Analyzing Citation-Distance Networks for Evaluating Publication Impact

Drahomira Herrmannova, Petr Knoth, Robert Patton
Oak Ridge National Laboratory, The Open University, Oak Ridge National Laboratory
Oak Ridge, TN, USA; Milton Keynes, UK; Oak Ridge, TN, USA
herrmannovad@ornl.gov, petr.knoth@open.ac.uk, pattonrm@ornl.gov

Abstract

Studying citation patterns of scholarly articles has been of interest to many researchers from various disciplines. While the relationship of citations and scientific impact has been widely studied in the literature, in this paper we develop the idea of analyzing the semantic distance of scholarly articles in a citation network (citation-distance network) to uncover patterns that reflect scientific impact. More specifically, we compare two types of publications in terms of their citation-distance patterns, seminal publications and literature reviews, and focus on their referencing patterns as well as on publications which cite them. We show that seminal publications are associated with a larger semantic distance, measured using the content of the articles, between their references and the citing publications, while literature reviews tend to cite publications from a wider range of topics. Our motivation is to understand and utilize this information to create new research evaluation metrics which would better reflect scientific impact.

Keywords: citation networks, publication impact, semantic similarity

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan[1].

1. Introduction

With the enormous and ever-growing number of research articles being published every year (Jinha, 2010; Laakso and Björk, 2012), researchers more and more often resort to using research evaluation metrics such as journal and publication citation indexes, as a proxy to quality and importance. Research metrics have been applied in various scenarios, from search and recommendation (Carevic and Schaer, 2014; Belter, 2016), to grant and tenure awards (Meho, 2007). Due to many drawbacks and limitations of the purely citation-based methods (Seglen, 1992; Seglen, 1997; Priem et al., 2010), recent years have seen the emergence of many new approaches and alternatives to the traditional *bibliometrics*, most notably metrics often referred to collectively as *altmetrics* (Priem et al., 2010) and *webometrics* (Almind and Ingwersen, 1997), which rely on data collected from the Web, such as download counts and social and news media mentions, and *semantometrics* (Knoth and Herrmannova, 2014), which measure how far each discovery takes us by utilizing publication full texts.

While citation networks have been widely studied in the literature, a number of works have recently developed the idea of studying citation patterns in combination with content similarity. In this paper we further explore this area and study how these citation-distance patterns reflect scientific impact. Our motivation is to explore whether certain patterns could be utilized in research evaluation to create new research evaluation metrics which reflect scientific impact more accurately than the widely used citation counts. We argue that measuring just the number of interactions in the scholarly communication network does not provide enough information for a sufficient understanding of the contributions a publication had, and posit that publication manuscript, in addition to the number of interactions, is needed to asses the value of a publication. For example, literature review publications are known to be highly cited, yet their main aim is to educate rather than influence a research area the way seminal publications do.

To this end we study the differences between citation-distance patterns of seminal publications and literature reviews. In this sense, these two types of papers represent extreme cases, as literature reviews, by definition, do not provide new ideas, while seminal publications greatly influence later developments. We believe research evaluation metrics, especially those focused on research impact, should be able to distinguish between these publication types. Our motivation is to understand and utilize information about the citation-distance patterns of these publications to create new research evaluation metrics which would better reflect scientific impact.

2. Related Work

A number of researchers have recently explored the idea of studying citation patterns in terms of content similarity, and utilizing these patterns for various tasks related to research evaluation. (Gerrish and Blei, 2010) have used a dynamic topic model to model thematic changes of content of documents, which was then used to create a Document Influence Model for measuring the importance of individual documents within a collection. (Yan et al., 2012) have used similarity between a publication and its references, which was calculated using Kullback-Leibler divergence of the content, to assess novelty. (Knoth and Herrmannova, 2014) have used semantic distance between publications

[1] `http://energy.gov/downloads/doe-public-access-plan`

which have cited a given publication and the publications cited by the publication to assess research contribution. In their case, distance was calculated using cosine similarity between $tf{-}idf$ term-document vectors. (Whalen et al., 2015) have used distance between a publication and publications that cite it, which was also calculated as cosine similarity, to predict future citations. In this paper we further explore the idea of studying citation patterns in terms of content similarity. To do this, we use the recently released TrueImpactDataset (Herrmannova et al., 2017) which contains publications of two types, seminal publications and literature reviews, and compare the citation patters of these two types of publications in terms of content distance.

3. Methodology

Our greatest interest lies in understanding how can we take into account publication content to improve automated research evaluation. We investigate citation networks in terms of content distance and study how the uncovered patterns can be used for identifying highly influential publications. Specifically, we investigate the relations in a citation network studied in (Yan et al., 2012), (Knoth and Herrmannova, 2014) and (Whalen et al., 2015), which are depicted in Figure 1.

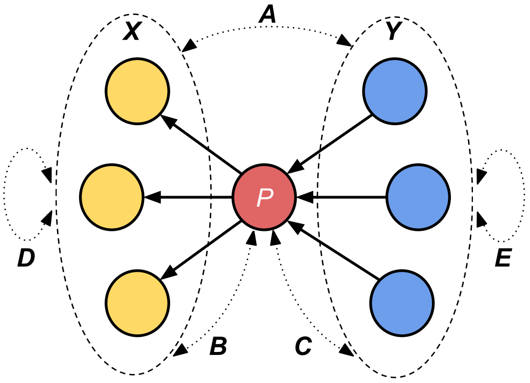

Figure 1: Neighborhood of a single publication P and relations between publications in the neighborhood which we investigate.

In the figure, node P represents the publication of interest, the yellow nodes (set X) represent publications cited by P (its references), and the blue nodes (set Y) represent publications that cited P. The relations between these publications, which are studied in this paper, are labeled A, B and C in the figure. In addition to the relations mentioned above, we also study the relations between the cited papers (regardless of whether there is a direct citation edge between them, these relations are labeled D), and between the citing papers (E).

The idea of studying the citing and cited publications is based on the process of how research builds on the existing knowledge in order to create new knowledge. In citation networks, the nodes, which are scholarly publications, are connected by citation relations which represent knowledge flows between the publications. While the cited papers are representative of the state-of-the-art in the domain of the publication in question (the publication itself contains only a fraction of the knowledge on which it is built, while the cited publications represent this knowledge more completely), the citing publications represent areas of application of the publication in question. Incorporating content into the analysis of citation behavior enables us to distinguish articles which draw on or influence diverse research areas from those which work within one specific discipline. Furthermore, the assumption, and a hypothesis made by both (Knoth and Herrmannova, 2014) and (Whalen et al., 2015), is that useful innovation will propagate in the form of new knowledge to the citing publications leading to a higher distance between the cited and citing publications.

To measure the distance we use the cosine similarity measure on $tf{-}idf$ term-document vectors created from the publications' abstracts. We then calculate the distance of two publications as $dist(p_1, p_2) = 1 - sim(p_1, p_2)$, where $sim(p_1, p_2)$ is the cosine similarity between the $tf{-}idf$ term vectors. Each set of relations A-E described above is represented as a set of distances (for example a set of distances between a publication and each of its references). We define a set of metrics applied on the distributions induced by the distances. An example of what characteristics we aim to distinguish is whether literature review publications typically cite a wider range of topics than seminal publications and whether seminal publications tend to work within a narrower area. The metrics we use to describe the distance distributions are: (1) minimum, (2) maximum, (3) range (difference between maximum and minimum), (4) sum of the distances, (5) mean distance, (6) standard deviation, (7) variance of the distances, (8) 25th percentile, (9) 50th percentile (median), (10) 75th percentile, (11) skewness, which is a measure of the asymmetry of the distribution, negative skew means the left tail is longer, positive skew means the right tail is longer, and (12) kurtosis, which is a measure of whether the data are heavy- or light-tailed, higher value means sharper peak. Because we describe each of the 5 distance distributions with 12 metrics, we have 60 features (features F1-F60) describing a publication's neighborhood.

4. Data

To collect all data needed for studying the relations introduced in the previous section, we have used three data sources:

1. TrueImpactDataset[2] (Herrmannova et. al., 2017) (Herrmannova et al., 2017), which provides us with seminal publications and literature reviews (i.e. the P node in Figure 1),
2. Microsoft Academic (MA) API[3] (Sinha et al., 2015) which we use to collect metadata (authors, year, venue, DOI, etc.) of the citing and cited publications (blue and yellow nodes in Figure 1),
3. Mendeley API[4] which we use to collect abstracts (since MA does not contain abstracts).

[2] http://trueimpactdataset.semantometrics.org/

[3] http://aka.ms/academicgraph/

[4] http://dev.mendeley.com/

Table 1 shows the size of the dataset. After collecting all needed data the size of the dataset was reduced to 276 publications (i.e. publications with at least one reference or at least one citation) – 126 literature reviews and 150 seminal publications.

Publications in TrueImpactDataset	314
TrueImpactDataset publications in MA	298
Pubs with at least one citation in MA	269
Pubs with at least one reference in MA	215
At least one cit. and one ref. in MA	209
Total number of citing papers	154,056
Total number of references	13,599

Table 1: Dataset size. The table shows for how many of the TrueImpactDataset publications we managed to get the needed metadata and how many additional publications we collected.

5. Experiments and Results

We begin by comparing the properties of seminal publications and literature reviews. We investigate how these two types of papers are situated with regard to the extracted features. To understand which features might assist with the task we calculate an independent one-tailed t-test for each feature. The t-test is a measure commonly used to assess whether two sets of data are statistically different from each other. In other words, it helps to determine the features that can distinguish literature reviews from seminal papers. To test the significance, we set the significance threshold at 0.05. Out of the 60 features, 27 result in p-value higher than 0.05. In this case we accept the null hypothesis of equal means. As the t-test tells us the values of these features are not significantly different for the two sets of papers, we remove these features from further analysis. The removed features are crossed out in Table 2.

From Table 2 it is obvious that there is not a single type of feature which describes well all five distributions. Furthermore, as most of the features describing the distribution E (distances between papers citing a publication) were removed, it seems this distribution does not offer much information for this task.

Next, we create a histogram for each feature and by comparing these histograms for the two publication types we gain insight into norms and placement of seminal publications and literature reviews in terms of their citation patterns. Figure 2 shows histograms of the remaining features, with seminal publications and literature reviews distinguished by color. In all of the histograms literature reviews are represented with dashed lines with circle points, while seminal publications with full lines with square points. The numbers in the legend of each plot show how many publications were used to produce each histogram (the numbers differ in case our data was incomplete and we could not calculate the given feature for all publications). To preserve space we do not show here histograms of all of the remaining features F1-F60, but instead we select 15 features with interesting properties.

	A	**B**	**C**	**D**	**E**
	(C-R)	**(P-R)**	**(C-P)**	**(R-R)**	**(C-C)**
min	~~F1~~	**F13**	~~F25~~	**F37**	**F49**
max	~~F2~~	**F14**	~~F26~~	~~F38~~	~~F50~~
range	~~F3~~	**F15**	~~F27~~	**F39**	**F51**
sum	~~F4~~	**F16**	**F28**	**F40**	**F52**
mean	**F5**	**F17**	**F29**	**F41**	~~F53~~
std	**F6**	**F18**	**F30**	**F42**	~~F54~~
variance	**F7**	**F19**	**F31**	**F43**	~~F55~~
p25	**F8**	**F20**	**F32**	**F44**	~~F56~~
p50	~~F9~~	**F21**	~~F33~~	~~F45~~	~~F57~~
p75	~~F10~~	~~F22~~	~~F34~~	~~F46~~	~~F58~~
skewness	**F11**	**F23**	~~F35~~	**F47**	~~F59~~
kurtosis	~~F12~~	~~F24~~	**F36**	**F48**	~~F60~~

Table 2: The columns in the table represent the five distance distributions studied here, the rows represent the 12 metrics used to describe each of the distance distributions, and the cells represent individual features. The second row of the header provides an explanation for which distance each column represents, e.g. *C-R* means the distance between citing papers and references, *P-R* means the distance between the publication and its references, *C-P* means the distance between the citing papers and the publication in question, and so forth. The crossed-out features are those which we removed from further analysis.

In general, various metrics seem quite consistent across both groups. However, these metrics also reveal some important differences in citation patterns of seminal publications and literature reviews. First, one of our expectations and a hypothesis made by both (Knoth and Herrmannova, 2014) and (Whalen et al., 2015) is that useful innovation introduced by a publication will propagate in the form of new knowledge to the citing publications, leading to a higher distance between the publication and the citing publications (distance C) as well as between the references and citing publications (distance A). This is confirmed by higher average distances of both distributions in case of seminal publications (features F5 and F29). This is further supported by a lower standard deviation of the A and C distance distributions for seminal papers (features F6 and F30).

Secondly, the distribution of distances between a publication and its references seems consistent with our expectations. In the case of literature reviews, the minimal distance between the publication and its references is on average smaller than for seminal papers (F13). At the same time, the difference between the most similar and most dissimilar reference is higher for literature reviews (F15). Furthermore the sum of distances between the publication and its references is higher for literature reviews than for seminal papers (F16), which is likely because reference lists of literature reviews are typically long.

5.1. Citation Patterns and Publication Importance

In this section we explore the relation between the perceived impact of publications and the different metrics used to measure it. Although the above analysis of the separate

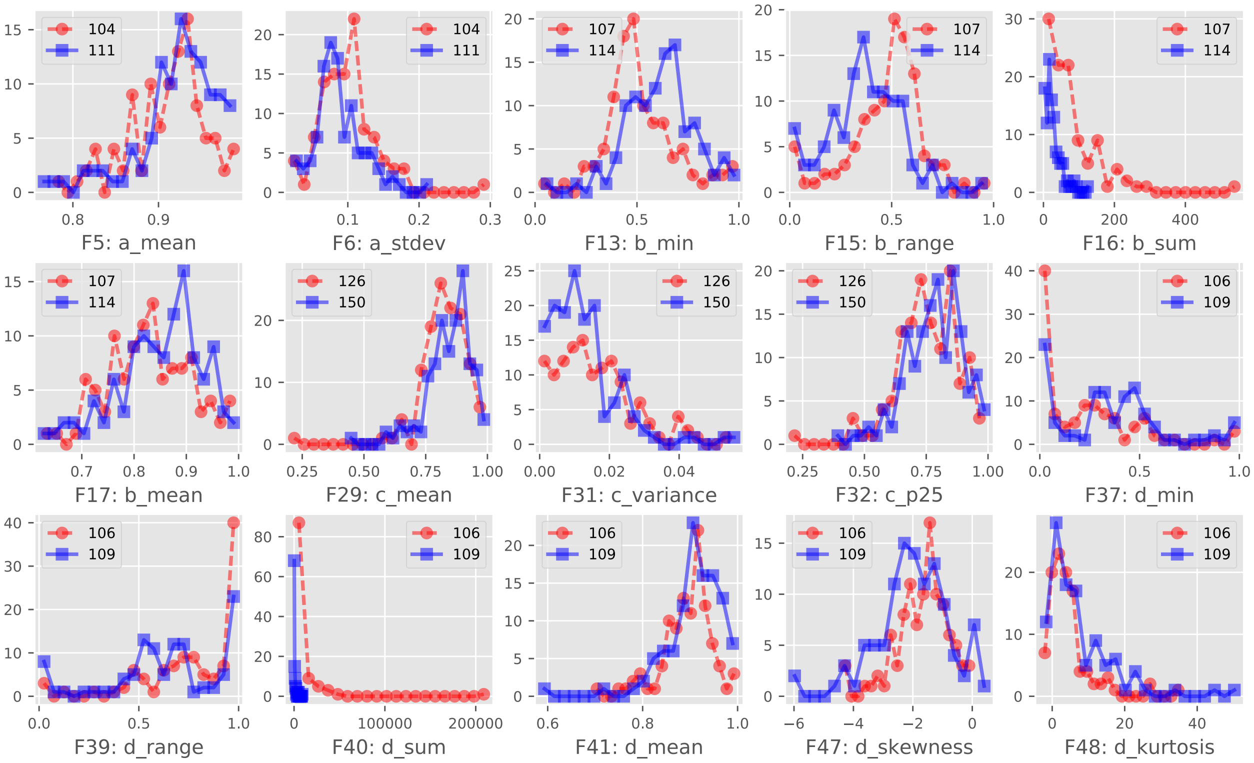

Figure 2: Histograms of selected features describing distance distributions *A*-*E* from Figure 1. In the figures, literature reviews are represented with dashed red lines with circle points and seminal publications with full blue lines with square points. The numbers in the legend of each subplot show how many publications were used to plot each line (the numbers differ across subplots in case we didn't have all data needed to calculate a given feature). Each subplot then shows how literature reviews and seminal publications are positioned with respect to that feature.

features revealed distinct differences between the citation behavior of seminal and literature reviews, we are interested in analyzing whether the revealed patterns help in distinguishing important seminal publications from literature reviews better than current research evaluation methods. To be able to compare features in terms of accuracy we approach this question as a classification task.

After testing different classifiers (specifically SVM, logistic regression, decision tree and Naïve Bayes), we have selected Naïve Bayes (NB) classifier as a classifier which works well on our dataset and our task across different setups. In our classification experiment we use a leave-one-out cross-validation setup, that is we repeatedly train on all but one publication and then test the performance of the model on the publication we left out of the training. The performance is evaluated using accuracy, considering seminal papers as the positive class. We compare the results against a baseline which always predicts the most frequent label (a seminal publication).

To understand the contribution of each feature, we trained and tested the NB classifier using a single feature at a time and calculated accuracy using each feature. This is again done in a leave-one-out cross-validation setup. For comparison we also train a classifier using citation counts as a single feature (specifically citation counts of each seminal and literature review publication obtained from MA). We are interested in analyzing whether some content-based features distinguish between these two types of publications better than citation counts. Table 3 shows results for top 10 features according to classification accuracy, as well as for citation counts, which were the 22. best feature out of the 27 features used in this experiment. All 27 features achieve better performance than the baseline.

#	Feature	Accuracy
1	F16: B sum	0.6897
2	F15: B range	0.6502
3	F13: B min	0.6453
4	F40: D sum	0.6355
5	F37: D min	0.6059
6	F31: C variance	0.6010
7	F39: D range	0.5911
8	F47: D skewness	0.5911
9	F32: C p25	0.5911
10	F48: D kurtosis	0.5813
22	Citations	0.5616
	Baseline	0.5025

Table 3: Classification performance when using individual features. The features are listed in descending order of accuracy.

The classification accuracy using the best performing feature F16 is $\sim$69%, while our baseline classifier achieves the accuracy of $\sim$50%. This means that by using F16 alone, it is possible to achieve 11% improvement over the widely used citation counts on this task. While in this study we only trained the classifier using a single feature at a time, in the future we will evaluate the performance of classifiers trained using a variety of well performing feature combinations.

It can be seen there are a number of features which work particularly well in distinguishing these two types of publications. These are particularly features describing the distance distributions B, C and D. The three best performing features are all related to the distance between a publication and its references. In particular, this experiment confirmed the features describing the distance between a publication and its references distinguish between the two types of papers (F13, F15, F16). Similarly, features describing the distribution of distances between a publication's references work well in this task, particularly features describing the "width" of the distribution and the shape of its peak (skew and kurtosis). One particularly interesting feature which outperforms simple citation counts by a significant margin is feature F31, which describes variance of the distance distribution C. This is interesting as it shows citations to literature reviews tend to come from broader mix of more and less distant citing publications.

6. Conclusion and Future Work

This paper studied the relationship between semantic distance of scholarly articles in a citation network and their impact. More specifically, following on the work of (Knoth and Herrmannova, 2014) and (Whalen et al., 2015) we investigated the novelty assumption, i.e. the idea that new useful ideas tend to propagate to the work of others, which in turn influences the semantic distance patterns in a citation network. To validate this assumption, we have used the new TrueImpactDataset (Herrmannova et al., 2017) to systematically evaluate a range of distance features characterizing the relationship between seminal and review publications, their references and citing publications. Our results show that there a number of features describing citation-distance patterns which significantly outperform widely used citation counts in distinguishing seminal publications from literature reviews on our dataset. This demonstrates content analysis might provide valuable information for research evaluation. While in this study we have focused on individual features, as future work, we are planning to evaluate the performance of a variety of well performing feature combinations. We also plan on experimenting with other semantic similarity measures, such as similarity computed on word2vec vectors, as well as investigating the effect of using abstracts compared to fulltext.

7. Bibliographical References

Almind, T. C. and Ingwersen, P. (1997). Informetric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics'. *Journal of Documentation*, 53(4):404–426.

Belter, C. W. (2016). Citation Analysis as a Literature Search Method for Systematic Reviews. *Journal of the Association for Information Science and Technology*, 67(11):2766–2777.

Carevic, Z. and Schaer, P. (2014). On the Connection Between Citation-based and Topical Relevance Ranking: Results of a Pretest using iSearch. In *BIR@ ECIR*, pages 37–44.

Gerrish, S. and Blei, D. M. (2010). A Language-Based Approach to Measuring Scholarly Impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 375–382.

Herrmannova, D., Patton, R. M., Knoth, P., and Stahl, C. G. (2017). Citations and Readership are Poor Indicators of Research Excellence: Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics. In *Proceedings of the 1st Workshop on Scholarly Web Mining*, pages 41–48. ACM.

Jinha, A. E. (2010). Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence. *Learned Publishing*, 23(3):258–263.

Knoth, P. and Herrmannova, D. (2014). Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, 20(11/12).

Laakso, M. and Björk, B.-C. (2012). Anatomy of Open Access Publishing: A Study of Longitudinal Development and Internal Structure. *BMC medicine*, 10(1):1. DOI: 10.1186/1741-7015-10-124.

Meho, L. I. (2007). The Rise and Rise of Citation Analysis. *Physics World*, 20(1):32.

Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010). Altmetrics: A Manifesto.

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, 43(9):628.

Seglen, P. O. (1997). Why the Impact Factor of Journals Should Not be Used for Evaluating Research. *BMJ: British Medical Journal*, 314(7079):498.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P., and Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246. ACM.

Whalen, R., Huang, Y., Sawant, A., Uzzi, B., and Contractor, N. (2015). Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science. *Quantifying and Analyzing Scholarly Communication on the Web (ASCW'15)*.

Yan, R., Huang, C., Tang, J., Zhang, Y., and Li, X. (2012). To Better Stand on the Shoulder of Giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 51–60. ACM.

8. Language Resource References

Herrmannova et. al. (2017). *TrueImpactDataset*. Distributed via `http://trueimpactdataset.semantometrics.org/`, 1.0, ISLRN 197-407-228-291-9.