Entity-Citation-Driven Academic Impact Measurement in Scientific Papers[1]*

Xinyan Gao[1,2], Chong Chen[1,*] Wenxi Li[1], and Yongxin He[1]

[1] *Beijing Normal University, No.19 Xnjiekouwai Street, Haidian District, Beijing, China*
[2] *Peking University, No.5 Yiheyuan Road, Haidian District, Beijing, China*

Abstract

Citation is a manifestation of the academic impact of scientific papers. Among diverse citation motivations, the most crucial one is that the research elements of a paper, e.g., the proposed problem, methods, models, etc., inspire the research of peers. As the tags of research elements are known as knowledge entities, citations that refer to certain knowledge entities of a cited paper are called entity citations. Both the position and the strength of an entity citation indicate the impact that a certain research element has. In this study, the academic impact of a cited paper is measured by the entity citations. A measurement approach is proposed with the technique of knowledge entity recognition and entity citation detection. The impact of a paper can be more precise and more interpretable with the proposed approach. The findings of this study can enhance the impact evaluation of both papers and knowledge entities, as well as improve the ranking quality in knowledge retrieval applications.

Keywords

impact evaluation, academic impact, knowledge entity, entity citation, citation context

1. Introduction

When evaluating the impact of a scientific paper, it is necessary to understand what has inspired peers' studies besides simply counting the citation number. Among different citation motivations, the research elements that describe problems, methods, models, and so on are the most crucial factors that drive citations, as they outline the important parts of solving problems. In scientific papers, these elements are called knowledge entities [1]. In this study, the citation that refers to certain knowledge entities of a cited paper is called entity citation. It helps to explain the reason for the impact of a scientific paper and thus plays an important role in the study of knowledge transmission [2]. We detect the entities in the context of citing papers and propose the entity-citation-driven measurement to evaluate the impact of scientific papers.

In previous studies, entity types in the domain of chemistry, biology, and medicine such as genes [2][3] have been concerned. In this paper, we focus on research problems and methods of machine learning since the studies of this field output rich theories and methodologies that have been foundations for many disciplines such as artificial intelligence, neurobiology, automation, etc. The research problems and methods in this field are important entity types that is likely referred to by other studies [4].

As research builds upon previous work, these entities can be explicitly or implicitly mentioned in the context of citing papers. Thus, the semantic meaning of citation context may associates with certain knowledge entity. From the perspective of influence, the importance of a knowledge entity is related to the frequency and the position its citation appears in citing papers, and also to the

---

influence of the citing papers themselves. Consequently, the more important entities a paper contains, the more impact it has.

This study aims to discover entity citations in the citing papers and evaluate the impact of cited papers. The contribution lies in that it not only helps to explain the exact reasons for the paper's impact but also improves the academic impact measurement at the granularity of knowledge entities. Besides, the study that identifies the important knowledge entities also helps to discover the core knowledge of the research domain and improves the rank of knowledge retrieval.

2. Related work

Academic influence evaluation for publications, journals, or other academic products has been a hot topic in the field of information science for a long time. The metrics include the citation number of papers, the impact factor of journals, and the number of likes, comments, and retweets on social media. However, the accounting of citation is mainly based on coarse-grained objects, such as papers. While some studies have conducted citation analysis at the lexical level, they often treat keywords as highlight terms without leveraging specialized knowledge entities [5]. Moreover, the evaluation that elucidates motivations [6][7] and sentiment [8] focuses more on the citing side, instead of the inspiring knowledge of the cited side.

Nakov et al. consider the sentences surrounding citations as an important tool for the semantic interpretation of cited papers. They define the text span of citation sentences and illustrate that a set of citation sentences expresses the same concepts in different ways [9]. It implies the possibility of entity citation study, i.e., obtaining knowledge entities of the cited paper from citation contexts by semantic analysis.

A series of studies contribute to revealing the academic value of cited papers through micro-level analysis using citation content. Thelwall et al. argue that being cited by a highly valuable paper indicates a significant influence of the cited paper, integrating the citation frequency of referenced works into the evaluation system for paper importance [10]. Sombatsompop et al. propose the citation position impact factor, which refers to the ratio of the number of times a citation appears in different positions in the cited paper to the total number of cited papers, as a way to evaluate the quality of papers [11]. Yang et al. make use of weights corresponding to different citation functions of citation context and multiply the weights with the values of citation strength, sentiment, etc., which together constitute the influence evaluation of papers [12]. We consider the citation number of a citing paper, the citation strength and the positions of entity citations as impact indicators of the knowledge entities.

3. Research design

As shown in Figure 1, the evaluation of the academic impact of scientific papers consists of two phases. Firstly, detecting entity citation by comparing the meaning of citation contexts with the knowledge entities of the cited paper. Secondly, evaluating the impact of cited papers based on the weight of entity citation. The weight combines the importance of the citing paper, the citation strength and the position of the citation context.

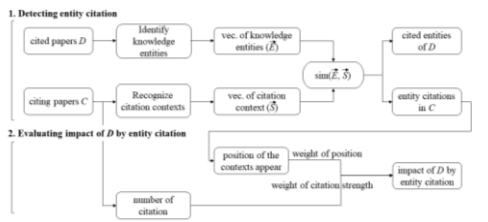**Figure 1:** Research design

Let D, C, and E respectively denote the cited papers, the papers that cite D, and the knowledge entities of D. The corpus D is composed of titles and abstracts of papers in certain fields. $C(d_i)$ is the

full text of papers that cites d_i, $d_i \in D$, $m=|C(d_i)|$. $e_k^{(i)}$ represents the kth entity identified from d_i. $s_l^{(j)}$ denotes the lth citation context in c_j, $c_j \in C(d_i)$. The knowledge entities are identified by a public toolkit based on the BiLSTM-CRF framework [13]. The embedding representations of $e_k^{(i)}$ and $s_l^{(j)}$ are weighted combination of SciBERT vector and Word2Vec vector. The latter is pretrained with a dataset of 76,274 machine learning papers built by [14] to complement the domain-specific semantic meaning to the former. In calculating the semantic similarity $(\vec{E},\vec{S})$, the entity with the highest similarity is selected for each citation context.

Set E includes the research problems E_p and methods E_m. The total number of E_p and E_m in d_i is r and r' respectively.

The academic impact of d_i, denoted as $I^{(i)}$, is defined as Formula 1. It considers the importance of the citing paper, the position weight of the entity citation appears, and the strength the entities are mentioned, which are respectively denoted as I_j, a_t and $f_{k,t}^{(i)}$. In this study, I_j is the citation number of c_j. a_t is assigned by Entropy Weight Method (EWM) to four different positions, i.e., Introduction & Background, Methods & Dataset, Experiment & Analysis, Conclusion. The weight of position is calculated in section 4.3. And $f_{k,t}^{(i)}$ means the number of citation contexts in position t of c_j that mention the entity $e_k^{(i)}$.

$$I^{(i)} = \sum_{j=1}^{m} I_j \cdot \sum_{t=1}^{4} a_t \sum_{k=1}^{r+r'} f_{k,t}^{(i)} \qquad (1)$$

4. Experiment and analysis

4.1. Dataset

One difficulty in data collection is maintaining the completeness of citation relations within a certain domain. We select a highly cited scholar in machine learning. All publications of the scholar are collected. Papers are screened out if the entities of problems and methods cannot be identified. The citing papers are crawled with the DOI obtained from DBLP if accessible. Finally, $|D|=418$ and $|C|=4730$. A total of 837 problem entities and 1593 method entities are identified from papers in D, and 10007 citation contexts are recognized semi-automatically by rules from C. Two domain experts annotated two subset of citation contexts, namely G_1 and G_2, as the gold standard. G_1 include all 259 citation contexts of the most frequent cited paper in D. G_2 consists of 100 citation contexts evenly sampled from four citation positions. The Kappa coefficient for annotation consistency is 0.852, with a significance p-value less than 0.001. The incoherent results are discussed and re-annotated by them. For example, a citation sentence is 'Generative Adversarial Networks: Goodfellow et al. proposed an adversarial learning model to train generative models, which showed promising performance in some computer vision tasks [41]–[45]'. In the papers it cites, the identified entities include 'deep hash', 'multi-task consistency-preserving adversarial hash' and 'cross-modal retrieval.' None of them explicitly appears in the citation sentence. According to the score of $(\vec{E},\vec{S})$, the matched entities should be 'cross-modal retrieval.' While, after discussed by the two domain experts, the corresponding entity is determine to be 'multi-task consistency-preserving adversarial hash.'

4.2. Parameters for detecting entity citation

In the experiment of detecting entity citation, the parameters for comparing the citation contexts with the knowledge entities are tuned for higher accuracy based on G_1. A total of 343 problem entities and 544 method entities are matched with all 10007 citation contexts. G_2 is used for evaluation the performance. The accuracy is 79%. Three aspects need to be highlighted. Firstly, the average accuracy is higher when the citation context only includes citation sentence than includes sentences before and after it. It means a larger context does not necessarily introduce desired semantic information about the cited knowledge entities. Secondly, the average accuracy increases from 43.6% to 51.4% when duplicating words from 1 to at most 10 before the citation notes for 5

times, as illustrated in Figure 2. It reminds us that effective information about the cited entities becomes denser in words preceding the citation notes. Thirdly, the accuracy increases further when combining the Word2Vec vectors with SciBERT vectors. The best performance is achieved by setting the weight 0.16 for SciBERT vectors and 0.84 for Word2Vec vectors, see Figure 2. It indicates richer domain knowledge can effectively compensate for the semantic representation since the word2vec model is trained in the domain-specific papers.
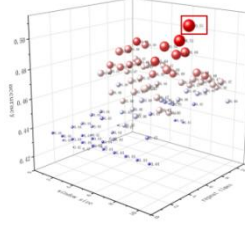

Figure 2: Tuning parameters for detecting entity citation

4.3. Paper impact evaluation driven by entity citation

Weight of citation position We first identify the position of each citation context in papers of C, then count the frequency of each position. Citation position weights are determined with EWM. The number and the normalized weights are shown in Table 1. What to be noted is although most citations appear in the first two positions, there are still 28% in the last two, which implies that the cited knowledge entities may provide experimental supports or theoretical foundations to the citing papers, and thus have a higher weight.

Table 1
Position weight of the citation context

Position of citation context	Number	Normalized weight
Introduction & Background	7145	0.16
Methods & Dataset	1220	0.20
Experiment & Analysis	1311	0.33
Conclusion	330	0.31

The weights derived in this study are largely consistent with those obtained through the expert scoring method in [15] and the questionnaires combined with AHP method as used in [12]. All of which emphasizes the importance of citations in the "Experiment & Analysis" and "Conclusion" sections. Additionally, Juyoung An et al. find that authors with the highest citation counts are often cited in the "Analysis" and "Conclusion" sections [16], further supporting the generalizability of this study's findings. Therefore, using the citation position weights derived through EWM for influence evaluation is reasonable.

Academic impact of papers In Figure 3, the total impact value of each paper is illustrated in descending order, also the impact contributed by the problem entities and the method entities of the cited paper D is shown. Generally, the papers with high impact only occupy a small proportion, most papers are not very influential. Notably, for highly impactful papers, contributions from both problem entities and method entities are substantial. The top six impact papers give evidence as shown in Table 2.

In a whole, the impact contributed by method entities is higher than problem entities. Considering the selected scholar is of high impact in the domain whose H-index ranked in the top 25 among the computer scientists of the world till Nov. 2023, according to Research.com, the results indicate the contribution of the scholar to the domain mainly lies in method innovation.
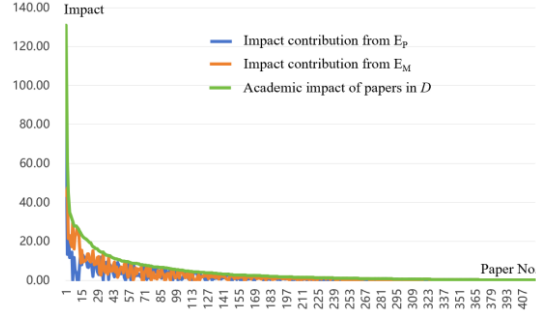
Figure 3: The academic impact of cited papers driven by entity citations

Table 2
Impact score of the top six papers

$I^{(i)}$	$I^{(E_p)}$	$I^{(E_m)}$
130.91	87.35	43.56
60.86	13.35	47.52
44.90	23.45	21.44
34.79	11.77	23.02
33.56	13.89	19.67
31.66	14.47	17.19

To further confirm the rationale of the academic impact measurement driven by entity citations, we calculate the correlation between the impact score proposed in this study and the citation number of papers, the classical impact metrics. The results in Table 3 indicate a high correlation of the two metrics; and in this research field, the impact contributed by method entities is notably higher than those of problem entities which is probably due to the substantial number of method entities.

Table 3
Correlation between the academic impact with the citation frequency

	$I^{(i)}$	$I^{(E_p)}$	$I^{(E_m)}$
Citation Frequency	0.87	0.69	0.87

5. Conclusion

Scholars cite previously published papers when the research elements of these papers enlighten their studies. They usually state the research elements with concise expression and, in most cases mention original knowledge entities of the cited papers. Such a citation driven by an entity indicates the influence of the inspiring knowledge. We propose an evaluation approach to academic impact with consideration of the entity citation. The contribution lies in that it not only helps to explain the exact reasons for the impact of a cited paper but also improves the academic impact measurement at the granularity of knowledge entities. Similar to traditional metrics like citation counts, our method cannot predict the impact of papers that have not been cited. However, it can reveal the specific reasons for papers' impact. Besides, the study that identifies the important knowledge entities also benefits to discovering the core knowledge of the research domain, and improving the rank of knowledge retrieval.

In the future, publications may be presented at a finer granularity of knowledge units. The method proposed in our study can be directly applied to the evaluation of research outputs, researchers, knowledge discovery, and information services.

Acknowledgements

Scientific Communication (IMSC) in the Era of Big Data" at 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Kimi in order to: translate a small portion of text into English and cheek spelling. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. Proceedings of the 11th International Workshop on Semant-ic Evaluation (SemEval-2017). (2017), 546–555.

[2] Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. Entitymetrics: measuring the impact of entities. PLoS One. 2013 Aug 29;8(8): e71416.

[3] Yesol Park, Gyujin Son, and Mina Rho. Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances[J].Applied Sciences,2024,14(20):9302-9302.

[4] Zhuoran Luo, Wei Lu, Jiangen He, et al. Combination of research questions and methods:a new measurement of scientific novelty[J]. Journal of Informetrics, 2022, 16(2):101282.

[5] Heng Huang, Donghua Zhu, and Xuefeng Wang. Evaluating scientific impact of publications: combining citation polarity and purpose[J]. Scientometrics, 2021, 127:5257-5281.

[6] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the Evolution of a Scientific Field through Citation Frames[J]. Transactions of the Association for Computational Linguistics, 2018(6):391-406.

[7] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, et al. Structural Scaffolds for Citation Intent Classification in Scientific Publications[J]. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1 (Long and Short Papers):3586-3596.

[8] Souvick Ghosh, Dipankar Das, and Tanmoy Chakraborty. Determining Sentiment in Citation Text and Analyzing Its Impact on the Proposed Ranking Index. [C]. In Proceeding of Computational Linguistics and Intelligent Text Processing. Springer, 2018:292-306.

[9] Preslav Nakov, Ariel Schwartz, and Marti A. Hearst. Citances: Citation Sentences for Semantic Analysis of Bioscience Text. In Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics.

[10] Mike Thelwall. "Should Citations be Counted Separately from Each Originating Section." J. Informetrics 13 (2019): 658-678.

[11] Narongrit Sombatsompop, Apisit Kositchaiyong, Teerasak Markpin, et al. Scientific evaluations of citation quality of international research articles in the SCI database: Thailand case study[J]. Scientometrics,2006,66(3):521-535.

[12] Siluo Yang, Ying Nie. Research on Evaluation Model of Papers' Influence Combined with Full-text Analysis[J]. Journal of Modern Information, 2022, 42(3):133-146.

[13] Heng Zhang, Chengzhi Zhang, Yingyi Wang. Revealing the Technology Development of Natural Language Processing: A Scientific Entity-Centric Perspective [J]. Information Processing and Management, 2024, 61(1):103574.

[14] Chong Chen, XingChen Ji, Denghui shang, and Yaxuan Lan. Identifying the Roles of Method Entities in Scientific papers[C]. Proceedings of 18th International Society for Knowledge Organization Conference (lSKO 2024). Advances in Knowledge Organization, Volume 20, 75-88.Ergon, Baden-Baden.

[15] Siniša Maričić, Spaventi J, Leo Pavičić, et al. Citation context versus the frequency counts of citation histories.[J]. Journal of the American Society for Information Science, 1998, 49(6):530-540.

[16] Juyoung An, Namhee Kim, Min-Yen Kan, et al. Exploring characteristics of highly cited authors according to citation location and content[J]. Journal of the Association for Information Science and Technology., 2017, 68(17):1975-1988.