Chapter 35
Attention-Based Multi-fusion Method for Citation Prediction

Juefei Wang, Fuquan Zhang, Yinan Li and Donglei Liu

Abstract As the most common research style in the process of academic exchange, the paper plays a vital role in the specific links of knowledge communication, academic cooperation, and scientific research education. However, in the traditional field of bibliometrics, how to quantitatively evaluate the influence of a paper generally depends on the number of citation as a reference standard. The number of citation is an important indicator for evaluating papers, and it has serious problems of lag. Therefore, based on the relevant meta-information generated in the publication process of the literature, the prediction of the future influence of the literature can make up for the above defects. In order to accurately predict the future citations of the paper, this paper constructs the Attention Convolution Neural Network model and combines the bibliometrics and alternative metrology-related features to enrich the input vector. Experiments on data sets collected from WOS and ResearchGate show that the model has improved accuracy compared to traditional prediction models.

Keywords Attention convolution neural network · Bibliometrics · Altmetrics

J. Wang · Y. Li · D. Liu (✉)
Computer School, Beijing Information Science and Technology University, Beijing 100101, China
e-mail: liudonglei@bit.edu.cn

J. Wang
e-mail: pstaion2@163.com

Y. Li
e-mail: 2120171031@bit.edu.cn

F. Zhang
Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China
e-mail: 8528750@qq.com

J.-S. Pan et al. (eds.), *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, Smart Innovation, Systems and Technologies 156,
https://doi.org/10.1007/978-981-13-9714-1_35

35.1 Introduction

Throughout the academic exchange process, literature is the most common scientific research style which play as a vital role. First of all, in the entire academic research, it serves as the main tool for academic information dissemination, comprehensively, truly, and systematically describes the scientific research results. Besides, it plays an important role in promoting academic exchanges, achievement of results and development of science and technology as an important basis for exploring academic issues and conducting academic research. However, as the level of research continues to increase, the number of papers published each year increases exponentially. On the other hand, relevant institutions also need a relatively stable and generalized evaluation method to evaluate the impact level of the paper.

Since Garfield [1] proposed the citation indexes in 1955, which can be used to evaluate the influence of a single paper, it is more and more common to evaluate papers by citations. The paper has a cumulative process of citations. This process lasts for half a year or more. This makes these traditional indicators based on citations of papers lag behind, and some newer papers cannot be evaluated or lack fairness. Therefore, the demand for the early prediction of the cited frequency is generated, and the prediction of the citation has become an important research topic for the evaluation of academic papers.

Usually, the citations of the paper is predicted by traditional bibliometric indicators, such as journal impact factor, h-index, and citations in previous years. In recent years, with the development of the Internet age, academic activities have gradually moved from offline to online, resulting in new evaluation indicators such as praise, collection, and comments, and thus Altimetric came into being. Compared with traditional bibliometrics, Altimetric has the advantages of diversity of evaluation indicators, high real-time performance, and comprehensive evaluation results. On the other hand, Altimetric also has various problems, such as the reliability and the coverage of data. In general, the emergence of Altimetric enriches the means of paper evaluation.

In this paper, by combining the bibliometric indicators and the alternative metrology indicators, the CNN+ Attention model is constructed to predict the citation frequency of the paper. The results are improved compared with the traditional methods.

35.2 Related Work

The research on the citation frequency of papers mainly focuses on two aspects: future selection [2–6] and prediction methods [7–11].

For the selection of indicators, the study by Ming yang Wang et al. [12] shows that the level of the first author and the quality of the article play a key role in the future reference of the paper. Vanclay [13] believes that journal impact factors, article length and publication type, and journal self-citing will affect the paper citation.

Jamali et al. [14] studied the relationship between title types and downloads and citations. In Altimetric, Costas et al. [15] confirmed the positive correlation between Altimetric and citation through the study of Altimetric scores in publications from social sciences, humanities, medicine, and life sciences, but relatively weak. The results of Thelwall et al. [16] show that the ranking based on the academic social website ResearchGate statistics has a strong relationship with the traditional paper ranking.

In terms of prediction methods, Yan et al. [17] compared the prediction effects of linear regression, k-Nearest Neighbor, support vector machine, and classification and regression tree (CART). The experimental results show that k-Nearest Neighbor with the worst prediction effect and CART predicts the best results with an accuracy rate of around 74%. Mistele et al. [18] use CNN to predict H-index. They show the results of a simple neural network that predicts the future citation counts of individual researchers based on various data from past researchers. Abrishami and Aliakbary [19] used the RNN neural network model, and their methods have better prediction accuracy than most current methods.

35.3 Methods

35.3.1 Feature Structure

Based on the summary of the feature of citations of papers, this paper constructs four kinds of original features from two aspects: bibliometrics and Altmetrics metrology. Detailed feature are listed in Table 35.1.

Domain Hotspots. Research hotspots often affect the citations of papers [20], and annual hotspots change with the year. In a given year, field research hotspots can be characterized by all the topics published in the current year. In this article, the topic of the article is constructed by the title, abstract, and keywords of the paper. In order to vectorize the paper text, we used doc2vec to process the title and abstract of each article, and use word2vec to process the keywords of each article.

Table 35.1 Detailed feature in the data set

Feature	Hotspots	Citation	Journal feature	ResearchGate feature

	Title;
	Abstract;
	Key word;
		3 years of citations;	Journal partition;
			3-year impact factor;
				Score_RG;
				ResearchItem_RG;
				ReadsNum_RG;
				CitationsNum_RG;
				RG_reads;
				Followers_RG;
				RelatedResearch_RG;

Citation by Year. The citation frequency of the previous years of the thesis will greatly affect the citation frequency of the paper [21]. In this paper, the citation of the first 3 years of the paper is selected as the relevant feature.

Journal Feature. Usually, the Journal published in the article will affect the citation frequency of the paper to a certain extent [13]. The citation of the paper, the higher the partition, the higher the impact factor, the papers published in journals with larger impact factors will be more likely to be cited. In this paper, the feature includes the journal partition level and the journal's 3-year impact factor.

Altmetrics Features. Altmetrics features cover a wide range of topics, including Twitter, Facebook, ResearchGate, Mendeley, etc. In this paper, we selected the data of the ResearchGate academic forum, and the feature of ResearchGate proved to be positively correlated with the paper.

35.3.2 Model Structure

The model used in this paper is shown in Fig. 35.1. The model in this paper first step is deals with the features of the document vectorization, and the feature input shape is (10, 200). Then, resampled to a (50, 40) shape, which facilitates merging with other features of the (5, 4) shape. A two-layer convolution of the features of (n, 50, 40) is made to have a shape of (5, 4). Then merged with authors, journals, and other numerical features to be the feature matrix with shape of (5, 5, 4), which through the attention layer. It allows the model to focus on key features. After three more convolution operations, the predicted output of (1, 1, 1) is finally obtained, which is the total reference number predicted for the next 3 years.

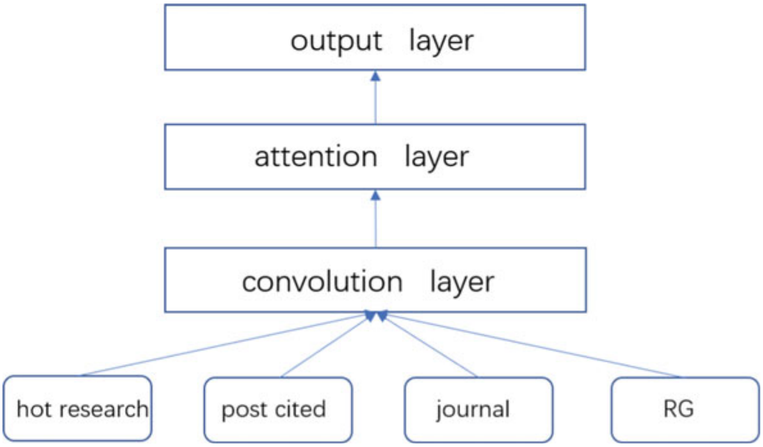

Fig. 35.1 ACNN module

35.4 Experiments

35.4.1 Dataset

The Literature data sets in this paper is collection form the published zone on internet. The journals and literature features mainly come from the Web of Science database and the public search system such as Google Scholar. The Literature Altmetrics feature come from ResearchGate database. Now, there is a collection of 34 literature data of all SCI journals in the transportation field for 10 years, with a total data for 40,000, ResearchGate data of 24,000, and the list in Table 35.2.

35.4.2 Training and Evaluation

The training uses a combination of gradient descent and backpropagation. Using the Huber loss as the loss function, which combines the advantages of the two MSE and MAE loss functions. When the loss is at $[0 - \delta, 0 + \delta]$ in the interval, the loss result is close to MSE, and close to MAE in the interval of $[-\infty, \delta]U[\delta, +\infty]$, which makes it possible to have better training effect when dealing with more abnormal values. In the original data set, we will randomly select 70% of the tag data to train the model, and the rest 30% as the test set.

The purpose of this experiment is to predict the development status in the next 3 years based on the development status of the first 3 years after the publication of the paper. The specific index of the model prediction is the total citation of the literature in the next 3 years. In order to determine if the prediction is accurate or not, we use the coefficient of determination to evaluate.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{1}$$

SSR: regression sum of squares, SST: total sum of squares, SSE: error sum of squares.

35.5 Results

In this section, we present the results of the evaluation by comparing different prediction models. We chose LR (Logistic regression), CART (Classification and Regression Trees), and the model ACNN shown in Fig. 35.1 for prediction. The results are shown in Table 35.3. It can be seen from the results that the ACNN model has a significant improvement in accuracy than the traditional LR and CART.

Table 35.2 34 journal on transportation field

Num	Journal name	Partition
1	Computer-Aided Civil and Infrastructure Engineering	1
2	Vehicular Communications	1
3	Transportation Research Part B-Methodological	2
4	Transportation Research Part C-Emerging Technologies	2
5	Transportation Science	2
6	IEEE Vehicular Technology Magazine	2
7	IEEE Transactions on Intelligent Transportation Systems	2
8	Transportmetrica B-Transport Dynamics	2
9	IEEE Transactions on Vehicular Technology	2
10	Networks and Spatial Economics	3
11	Transportation Research Part E-Logistics and Transportation Review	3
12	Transportation Research Part A-Policy and Practice	3
13	Transportation	3
14	Transportation Research Part D-Transport and Environment	3
15	IEEE Intelligent Transportation Systems Magazine	3
16	International Journal of Engine Research	4
17	Transportmetrica A-Transport Science	4
18	Journal of Advanced Transportation	4
19	Journal of Intelligent Transportation Systems	4
20	Proceedings of The Institution of Mechanical Engineers Part F-Journal of Rail and Rapid Transit	4
21	Proceedings of The Institution of Mechanical Engineers Part D-Journal of Automobile Engineering	4
22	IET Intelligent Transport Systems	4
23	International Journal of Automotive Technology	4
24	Journal of Transportation Engineering	4
25	European Transport Research Review	4
26	Transport	4
27	International Journal of Vehicle Design	4
28	Transportation Planning and Technology	4
29	Transportation Research Record	4
30	Transportation Letters-The International Journal of Transportation Research	4
31	Promet-Traffic & Transportation	4
32	Proceedings of The Institution of Civil Engineers-Transport	4
33	International Journal of Heavy Vehicle Systems	4
34	ITE Journal-Institute of Transportation Engineers	4

Table 35.3 Results from different models

Modules	R^2
LR	0.726
CART	0.761
ACNN	0.845

Table 35.4 Feature elimination experiment result

Feature	+	–
ALL	0.845	
- hot research	0.322	0.833
- post cited	0.736	0.502
- journal	0.368	0.821
- RG	0.240	0.822

In addition, we performed elimination experiments to further analyze the impact of features on the results. The detailed results are shown in Table 35.4, where the "+" indicates that the feature is used alone, and the "−" indicates that the feature is removed from all features. Through the results, we can see that the number of citations in the past years has the greatest impact on the subsequent prediction results, and the addition of research hotspots and alternative metrology features can effectively improve the prediction accuracy of the model.

35.6 Conclusion

In this paper, by constructing the model of attention convolution neural network, we predict the citation frequency of the paper. Compared with the traditional LR and CART prediction models, the accuracy rate is improved. Besides, we construct bibliometrics, alternatirices and construct feature vector which show the experimental results that adding research hotspots and the features of ResearchGate can be helped with prediction accuracy. In the next step, we will continue to study how to increase the Predictive accuracy by Improvement the model.

References

1. Garfield, E.: Citation indexes for science: a new dimension in documentation through association of ideas. Science **122**(3159), 108–111 (1955)
2. Joyce, C.W., Kelly, J.C., Sugrue, C.: A bibliometric analysis of the 100 most influential papers in burns. Burns **40**(1), 30–37 (2014)
3. Finardi, U.: Correlation between journal impact factor and citation performance: an experimental study. J. Inf. **7**(2), 357–370 (2013)

4. Abramo, G., Cicero, T., D'Angelo, C.A.: Are the authors of highly cited articles also the most productive ones? J. Inf. **8**(1), 89–97 (2014)
5. Gazni, A., Didegah, F.: *Investigating Different Types of Research Collaboration and Citation Impact: A Case Study of Harvard University's Publications*. Springer-Verlag, New York (2011)
6. Bornmann, L., Daniel, H.D.: Citation speed as a measure to predict the attention an article receives: an investigation of the validity of editorial decisions at Angewandte Chemie International Edition. J. Informetr. **4**(1), 83–88 (2010)
7. Wang, D., Song, C., Barabasi, A.L.: Quantifying long-term scientific impact. Science **342**(6154), 127–132 (2013)
8. Fu, L.D., Aliferis, C.F.: Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics **85**(1), 257–270 (2010)
9. Acuna, D.E., Allesina, S., Kording, K.P.: Future impact Predicting scientific success. Nature **489**(7415), 201 (2012)
10. Yu, T., et al.: Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics **101**(2), 1233–1252 (2014)
11. Ibáñez, A., Larrañaga, P., Bielza, C.: Predicting citation count of Bioinformatics papers within four years of publication. Bioinformatics **25**(24), 3303–3309 (2009)
12. Wang, M., Yu, G., Yu, D.: Mining typical features for highly cited papers [J]. Scientometrics **87**(3), 695–706 (2011)
13. Vanclay, J.K.: Factors affecting citation rates in environmental science [J]. J. Inf. **7**(2), 265–271 (2013)
14. Jamali, H.R., Nikzad, M.: Article title type and its relation with the number of downloads and citations [J]. Scientometrics **88**(2), 653–661 (2011)
15. Costas, R., Zahedi, Z., Wouters, P.: Do, "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective [J]. J. Assoc. Inf. Sci. Technol. **66**(10), 2003–2019 (2015)
16. Thelwall, M., Kousha, K.: ResearchGate vs. Google scholar: which finds more early citations? Scientometrics **112**(1), 1–7 (2017)
17. Yan, R.,Tang, J., Liu, X., et al.: Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1247–1252. ACM (2011)
18. Mistele, T., Price, T., Hossenfelder, S.: Predicting citation counts with a neural network [J]. arXiv preprint arXiv:1806.04641 (2018)
19. Abrishami, A., Aliakbary, S.: NNCP: a citation count prediction methodology based on deep neural network learning techniques [J]. arXiv preprint arXiv:1809.04365 (2018)
20. Sohrabi, B., Iraj, H.: The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. Scientometrics **110**(1), 243–251 (2017)
21. Yu, T., et al.: Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics **101**, 1233–1252 (2) (2014)