Will your paper get promoted by a citation? A case study of citation promoter in computer science discipline

Luo, Feiheng; Sun, Aixin; Raamkumar, Aravind Sesagiri; Erdt, Mojisola; Theng, Yin-Leng

2018

Luo, F., Sun, A., Raamkumar, A. S., Erdt, M. & Theng, Y. (2018). Will your paper get promoted by a citation? A case study of citation promoter in computer science discipline. IEEE Transactions On Emerging Topics in Computing, 9(1), 238-245. https://dx.doi.org/10.1109/TETC.2018.2861321

https://hdl.handle.net/10356/148426

https://doi.org/10.1109/TETC.2018.2861321

Will your paper get promoted by a citation? A case study of citation promoter in computer science discipline

Feiheng Luo, Aixin Sun, Aravind Sesagiri Raamkumar, Mojisola Erdt, and Yin-Leng Theng

Abstract—Reseachers have investigated numerous factors influencing citation counts of cited papers. One factor investigated has been the number of gained citations, as this could increase the visibility of cited papers and subsequently induce further citations. In this paper, aiming to identify a particular kind of citation that could trigger a rapid growth in the citation counts of cited papers, a concept of a "citation promoter" was proposed. We defined citation promoters based on the annual citation rates of the cited papers and the co-citation counts received by the pair of cited and citing papers. The comparative results showed that papers would obtain a sharp rise in citation counts shortly after they were cited by citation promoters. Papers that received citation promoters at an early age also outperformed other papers in long-term citation counts. In addition, we developed a classification model for predicting whether a citing paper would be a citation promoter for its cited paper. Since it was a class imbalanced problem (4% positive instances), and there was a lack of content and author features in our dataset, our preliminary models achieved moderate performance with an F_1 score slightly higher than 0.5, while the F_1 score obtained by random guessing was 0.07.

Index Terms—citation promoter, citation analysis, co-citation, citation prediction, imbalanced binary classification.

1 Introduction

Evaluating the quality of scientific papers is a complex task. Various impact indicators have been proposed to facilitate the process of evaluation [1], [2], [3], [4], where citation count is the most widely used metric, though the usage of citation counts remains controversial [5], [6]. Hence, a number of studies have investigated the factors influencing citation counts. While some studies explored the factors associated with cited papers, such as authors and publication venues of cited papers, other studies focused on citations, i.e., citing papers [7], [8], [9], [10], [11], [12], [13], [14].

Citations could increase the visibility of cited papers and subsequently induce further citations [7]. With the accumulation of citations, cited papers have an advantage of obtaining more and more citations. In addition, early citations indicate that cited papers have received recognition or feedback from the research community in the initial stage after publication. Exploring the relationship between early citations and long-term citations remains a popular research topic in citation analysis, and early citations can also be used to make predictions of the future impact of cited papers [12], [15]. Furthermore, previous studies have also investigated the importance of citations. Citations can be considered as scholarly votes or credits assigned to cited papers [16]. It is suggested that citations should be given different weights, since they vary in many aspects, such as the publication venues of citing papers and the citation time interval [13], [17].

In our study, a new concept of a "citation promoter" was proposed, with an aim to identify a particular type of citation which has a greater importance to cited papers, in terms of leading the rapid growth of citation counts of the cited papers. Specifically, once a paper receives a particular citation, and its total citation count increases swiftly and considerably, then the citing paper can be considered as a citation promoter for the cited paper. The citation promoter could be helpful in promptly identifying potentially influential papers. It could also provide more information and improve the performance of predicting future citation counts.

As a newly proposed concept, citation promoter was evaluated in this study. First of all, we learned from historical citation data to observe the rapid growth of papers' citation counts. An explicit definition of a citation promoter was introduced based on the annual citation rate, i.e., the yearly average citation count of the cited paper, and the co-citation count of the pair of cited and citing papers. While the annual citation rate could indicate the growth of citation counts, the co-citation counts could be partly considered as the direct impact that the citing paper exerts on the cited paper. This could be because co-citations are less common than citation counts and they tend to represent shifts in research activity [18]. Next, we studied the citation promoters identified by our definition. The effectiveness of the definition was investigated through a series of comparative studies on the citation performance between papers with citation promoters and other papers. After understanding the impact of citation promoters, it was expected to explore the factors which could influence a citing paper to become a citation promoter. These factors have not been revealed by

- *F. Luo, A. Sesagiri Raamkumar, M. Erdt and Y. Theng are with the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. E-mail: leavenlfh@gmail.com; aravind002@ntu.edu.sg; mojisola.erdt@ntu.edu.sg TYLTheng@ntu.edu.sg*
- *A. Sun is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. E-mail: axsun@ntu.edu.sg*

Manuscript received April 19, 2005; revised August 26, 2015.

our definition, since the definition was generated from the detected growth of citation counts. However, in practice, the growth is unknown, and it is meaningful to identify citation promoters before the promoting effect is exposed. Hence, we developed preliminary models to predict if a citation would become a citation promoter for its cited paper. The models acquired moderate predictive performance due to the limitations of our dataset, and we will improve the models in future work.

The main contribution of our study is in proposing a novel concept of a citation promoter for a better understanding of the impact of citations, so as to discover influential papers promptly. A series of comparative experiments were conducted to evaluate the impact of citation promoters. In addition, preliminary efforts were exerted in building models of predicting citation promoters.

The rest of this paper is organized as follows. Section 2 reviews important literature related to this study. Section 3 introduces our definition of a citation promoter. The impact of citation promoters is presented in Section 4, and the predictive models are developed and evaluated in Section 5. Finally, in the conclusion in Section 6, research limitations and future work are discussed.

2 Related work

2.1 Citation Related Metrics

While citations are usually unweighted in the evaluation of research outputs, researchers have proposed to assign different weights to citations for reflecting the citation properties such as the prestige and time interval. Various methods have been developed to calculate the weights of citations. Eigenfactor score [19] is a popular alternative to the impact factor for evaluating scientific journals. Citations from highly ranked journals are weighted to contribute more to the Eigenfactor score. Yan and Ding [13] proposed a weighted citation score by considering the prestige of the citing journal and citation time interval, where the prestige of the citing journal was quantified by Article Influence score which is the average Eigenfactor score per article published in that journal. While the majority of papers with large citation counts were found to have high weighted citation scores, a portion of papers with relatively low citation counts were also detected to have high prestige in terms of the weighted scores. A series of PageRank based methods were also applied in the constructed citation network for weighted citations [17]. It was found that citation counts and the generated PageRank scores were positively correlated. At author level, the work of [20] employed weighted citations to distinguish between an authors popularity and prestige, where the popularity was measured by the total citation counts the author had gained, and the prestige was measured by the number of times an author had been cited by highly cited papers.

Apart from weighted citations, studies have also explored the properties of citations from other perspectives. A recent study [21] investigated a method of determining the prestige of citing papers by their affiliations. It was shown that papers that received citations from top universities obtained higher total citation counts as compared to other papers that never received those kinds of citations. The works of [14] and [22] were comparable. They both proposed a supervised classification approach to identify meaningful citations, where the importance of citations was modeled by several aspects, such as the location of the citation in the text. In addition, the work of [23] classified citations into different categories based on sentiment analysis.

2.2 Co-citation Analysis

The studies mentioned above indicate the merits of understanding the impact of different citations. In our study, it is expected to identify the citations that play a significant role in the growth of citation counts of the cited papers. Co-citation counts were leveraged in the definition of the proposed citation promoter. Author co-citation analysis (ACA) and document co-citation analysis are traditionally used to quantify the similarity between authors or documents, and to identify the nature of specialties in a scientific field [24]. In recent years, researchers started to take the citation content into consideration. The work of [25] showed that content-based author co-citation analysis not only achieved similar intellectual structure as the traditional methods, but also provided more details about the sub-disciplines. A formulation of contextual co-citation strength was proposed in [26]. [R4] proposed a hybrid method by combining author co-citation relations, citations publication time, citations publication venue, along with citations keywords to build co-citation matrices. Co-citation context and proximity have also been proposed to classify reasons for a citation. In one such study [R5], content and proximity-based data have been used to construct ACA. Based on these previous studies on co-citations, it could be a potential implementation to use the co-citation count as an indicator of the impact exerted on cited papers.

2.3 Prediction of Publication's Impact

Predicting the future impact of a scientific paper is a challenging research topic. Most related studies defined the task as a regression problem. While the works of [15] and [27] made a prediction of future citation counts, the work of [28] predicted the long-term citation impact using a quantile regression approach. Additionally, a recent study evaluated the impact of early citers, i.e., the authors of early citations, on the long-term citation counts [12]. By incorporating the properties of early citers, their regression model gained a significant performance improvement. In contrast, authors in [29] formulated the prediction of whether a published paper will increase the h-indexes of the authors as a binary classification problem. They collected a total of 26 features from six aspects, such as content and venue, and then fitted the data into popular classifiers including logistic regression, random forest and so on. In our study, the task of predicting whether the citing paper is a citation promoter for the cited paper was also formulated as a binary classification problem.

3 Definition of citation promoter

Citation promoter is a new concept to understand citations proposed in our study. Thus, as a new concept, citation promoter could be defined in multiple ways. The definition

in our paper is based on *annual citation rate* as one of the possible definitions of this concept. As an initial study, we define a citation promoter in a simple way, based on the variation of the annual citation rate of the cited paper, and the co-citation count of the cited paper, and the citing paper. Co-citation counts are important in identifying citation promoters because without this metric, the promotion effect of the citing paper cannot be validated. Our definitions are based on the finding that the co-citations largely dictate the similarity of papers. This has been proven in some recent studies [30], [31]. In the following definitions, we use the term 'age' to refer to the length of time that a paper has remained published. Consider a paper *P* has $C_p^{y_p}$ citations at age y_p ($y_p = 1, 2, 3 \cdots$), and it is cited by a citing paper *CP* at age $y_{(p,cp)}$. The annual citation rate is the yearly average citation count. For paper *P*, the annual citation rate is $R_p^{y_p} = C_p^{y_p}/y_p$. Co-citations of papers *P* and *CP* are the citations referring to both papers. We consider the citation performance of the cited paper within a three year time window after it has been cited to identify the citation promoter. A three year time window was considered to ensure there is sufficient time for citations to be picked up by papers so that meaningful comparisons could be performed. Also, it helps us in comparing the annual citation rate for each year in this window. Specifically, the citing paper *CP* is defined as a citation promoter for *P*, if two requirements are satisfied: 1) the annual citation rate of paper *P* increases in the three year window, i.e., $Max(R_p^{y_{(p,cp)}+k}) > R_p^{y_{(p,cp)}}$, where $Max(\cdot)$ denotes the maximum value, and $k \in [1, 2, 3]$; and 2) the accumulated co-citation count of papers *P* and *CP* within the three year window exceeds paper *P*'s past total citation count that it had gained in the year of being cited by paper *CP*, i.e., $CC_{(p,cp)}^3 > C_p^{y_{(p,cp)}}$.

From the definition, the citation promoter is determined according to the upcoming citations that the cited paper will receive. It is possible that a paper is cited by multiple citation promoters. However, the definition does not reveal the intrinsic properties of becoming a citation promoter. In addition, while the upcoming citations can be retrieved from an offline historical dataset in experiments, this kind of information is not available in practice. Hence, it is more important to develop a predictive model for detecting citation promoters. The model should be able to learn the intrinsic properties of citation promoters and their latent relationships with the cited papers, and then make predictions based on existing information before the promoting effect has been exposed. We will present how we build the predictive models after the impact of citation promoters, defined in this way, is evaluated.

4 Impact of Citation Promoter

According to the proposed definition in the last section, we extracted citation promoters from a Microsoft Academic Graph (MAG) [32] dataset released in February 2016, and studied their impact on cited papers. In this current study, we only considered the papers published between 2000 and 2005 in computer science journals and conferences. The publication time period was selected to ensure a ten year citation window for each paper. The computer science publication venues were retrieved from DBLP and mapped to the venue entries stored in the MAG dataset. Papers that had never been cited were removed. At the end, a total of 445,638 papers were retrieved, and there were 28,777 papers (6.5%) which had at least one citation promoter within the ten year citation window.

TABLE 1 Comparisons of ten year citation counts between papers with and without citation promoters. P — Papers with citation promoters; NP — Papers without citation promoters

Age	Paper count		Mean of ten year citation counts		Median of ten year citations	
	P	NP	P	NP	P	NP
1	16,039	199,804	51.1	18.5	32	10
2	5946	260,927	28.0	12.4	19	7
3	2646	287,643	19.7	9.3	14	6
4	1321	298,303	15.6	7.3	11	5
5	721	301,207	13.4	6.1	10	4

4.1 Comparisons of Long-term Citation Counts

Papers with citation promoters were compared to those that had never been promoted, in terms of ten year total citation counts, given that the two groups of papers had close citation counts at early ages. Specifically, papers were grouped according to age and the citation counts they had gained by that age. For papers with citation promoters, the age means when they received their first citation promoter.

In Table 1, we list various groups of papers in comparisons. At each age group, we considered papers which had the same range of early citation counts 1 to 10. From Table 1, it is observed that papers with citation promoters achieve better performance within the ten year window, where their mean and median values of ten year total citation counts are more than twice as many as the corresponding values of papers without citation promoters. These comparative results are also statistically significant ($p < 0.01$) under the Mann-Whitney tests. We can also see from Table 1 that a majority of papers were promoted in the first year after publication. The reason could be due to the definition of a citation promoter, which requires upcoming co-citation counts to be greater than the current citation counts. It is easier for papers to satisfy this requirement at an early age. Table 1 also suggests the effect of accumulative advantage, where papers could harvest more citations if they were cited earlier.

In Table 1, among the papers without citation promoters, there could be a number of papers that stopped earning citations, and thus the ten year citation performance of these paper groups could have been affected negatively. However, for the papers with citation promoters, the definition of a citation promoter has ensured an increment in citation counts. Hence, we made further comparisons in Table 2. For papers without citation promoters, we only considered those papers whose annual citation rate increased in the subsequent three years, similar to how we defined citation promoters. As seen in Table 2, papers with citation promoters also outperform the other group of papers, though the differences in ten year citation counts have narrowed.

TABLE 2 Comparisons of ten year citation counts between papers with and without citation promoters, considering papers with ACR increased. P — Papers with citation promoters; NP — Papers without citation promoters

Age	Paper count		Mean of ten year citation counts		Median of ten year citations	
	P	NP	P	NP	P	NP
1	16,039	63,766	51.1	34.9	32	22
2	5946	114,957	28.0	19.0	19	13
3	2646	130,866	19.7	13.5	14	9
4	1321	130,095	15.6	10.5	11	8
5	721	123,976	13.4	8.7	10	6

TABLE 3 Number of papers in each group for extracting annual citation trend. CCR — Citation count range; P — Papers with citation promoters; NP — Papers without citation promoters

Group	Early CCR	Year 10 CCR	No. of P	No. of NP
1	$[1, 20]$ at age 1	$[100, 200]$	1761	4781
2	$[1, 30]$ at age 2	$[100, 300]$	289	4964
3	$[1, 50]$ at age 3	$[100, 500]$	61	5238

4.2 Comparisons of Annual Citation Trends

In the previous comparative analyses, we considered papers with close early citation counts. In addition, we retrieved papers which had the same ranges of early and final citation counts, and then extracted their annual citation trend. The basic statistics of the retrieved papers are presented in Table 3, and their annual citation trends are illustrated in Figure 1. Figure 1 shows annual citation trends of the two groups of papers. The x-axis refers to years from year-1 to year-10. The y-axis refers to the average citation count. For example, in the left subfigure (group 1), the first dot in the line of papers with promoters denotes that the average citation count is slightly lower than 7.5 in the first year after publication. Other patterns can be observed in all of the three subfigures in Figure 1, where the trends for papers with citation promoters notably display a rise in average citation counts. These papers obtained higher citation counts on average over the ten years, even though they gained fewer citations at early ages.

5 Prediction of Citation Promoter

In this section, we firstly clarify the prediction problem before we develop prediction models. As known from our definition of a citation promoter, while a paper could have many citations, only a few can be identified as the promoters, or even none. Further, a citation promoter, as a citing paper, generally cites a number of papers in its reference list, but obviously it will not be promoting all the references. From the retrieved 28,777 papers with citation promoters, there were 29,520 unique citing papers identified as citation promoters. All of the referenced papers of the 29,520 citation promoters were extracted. We found that on average, each promoter cited 25 references, 2 of which were promoted. Consequently, the prediction of a citation promoter actually considered each pair of cited and citing papers. Similar to the work of [29], we defined the task as a binary classification problem, in which every instance was a pair of cited and citing papers. If the citing paper was a promoter for the cited paper, then this instance was labeled positive, otherwise it was labeled negative. If all of the retrieved 445,638 papers and their citations were considered as input, a great demand of computing resources would be required. Since most of the citation promoters were obtained in the first year after publication (as shown in Table 1), we only considered papers and their first year citations in our classification experiments. There was a total of 346,705 instances or citation pairs, 14,082 (4.1%) of which were positive instances.

5.1 Extracted Features

Features were extracted from the MAG dataset for each instance. We considered features related to citations, references, and publication venues. However, paper abstracts were not contained in the MAG dataset, and author information was not available for a large number of papers. Facing this problem, we extracted additional features from the citations received by cited and citing papers in the following year after the citing paper was published. At the end, a total of 32 features were extracted, as presented in Table 4.

5.2 Classification Models

As our task was formulated as a binary classification problem, we used a series of different classifiers, including XGBoost (XGB) [33] and random forest (RF). We also conducted experiments with support vector machines, logistic regression, Naïve Bayes models, and k-nearest neighbors models, but these models were less suitable for our retrieved dataset and their performances were poor compared to the other three models. Consequently, we only present the results of XGB and RF.

We took 70% of the instances as training set and the remaining 30% of instances for validation. Since our task is imbalanced, where positive instances only accounted for 4.1%, we employed different methods such as data sampling to balance the two classes of instances before we fitted the training set data into classifiers. We left the validation set as imbalanced as it originally was. The methods that we used to balance data included random over-sampling, random under-sampling, Synthetic Minority Oversampling Technique (SMOTE), amongst others. We also allocated higher weights to positive instances in training the classifiers.

5.3 Prediction Results

The validation set contained 114,413 instances, of which 4647 were labeled with citation promoters. We quantitatively evaluated the predictability of the classifiers using precision, recall, and F_1 score. Since it was an imbalanced classification problem, an average evaluation score of both negative and positive prediction outcomes would be biased and less practical. We thus only focused on the predictive performance of the positive instances. Table 5 lists the best prediction results of different classifiers associated with

TABLE 4 Feature definition. Each instance is a pair of cited and citing papers. P — Cited paper; CP — Citing paper

	Feature	Description
Citation	CC-cur	P's citation count in the year of receiving CP
	CC-next	The number of citations P receives within the following year of receiving CP
	CC-diff*	The difference between CC-next and CC-cur
	CC-diff-rt*	The ratio between CC-diff and CC-cur
	CC-inc-rt*	The ratio between CC-next and CC-cur
	ACR	The yearly average citation count of P
	CP-cc	CP's citation count in the first year after publication
	Diff-P-CP	The difference of total citation counts between P and CP in the first year after CP's publication
	Diff-P-CP-rt*	The ratio between CP's citation count and P's citation in the first year after CP's publication
	Diff-CP-Pnext	The difference between CP-cc and CC-next
	Diff-CP-Pnext-rt	The ratio between CP-cc and CC-next
	CoCC*	The number of co-citations received in the first year after CP's publication
	CoCC-P-rt*	The ratio between CoCC and CC-next
	CoCC-CP-rt*	The ration between CoCC and CP-cc
Venue	V-P-if*	The impact factor of P's publication venue
	V-P-h	The h-index of P's publication venue
	V-CP-if*	The impact factor of CP's publication venue
	V-CP-h	The h-index of CP's publication venue
	V-if-diff*	The difference between V-P-if and V-CP-if
	V-if-diff-rt*	The ratio between V-if-diff and V-P-if
	V-h-diff	The difference between V-P-h and V-CP-h
	V-h-diff-rt	The ratio between V-h-diff and V-P-h
	V-same	A binary indicator of whether P and CP are published at the same venue
Reference	R-P-num	The number of references in P
	R-P-cc*	The total citation count of the references in P
	R-P-avgcc*	The average citation count of the references in P
	R-CP-num*	The number of references in CP
	R-CP-cc*	The total citation count of the references in CP
	R-CP-avgcc*	The average citation count of the references in CP
	R-co-num	The number of co-references in P and CP
	R-co-cc	The total citation count of the co-references in P and CP
	R-co-avgcc	The average citation count of the co-references in P and CP

Note: The asterisk (*) symbol indicates that the feature is an important feature for the results in Table 5.

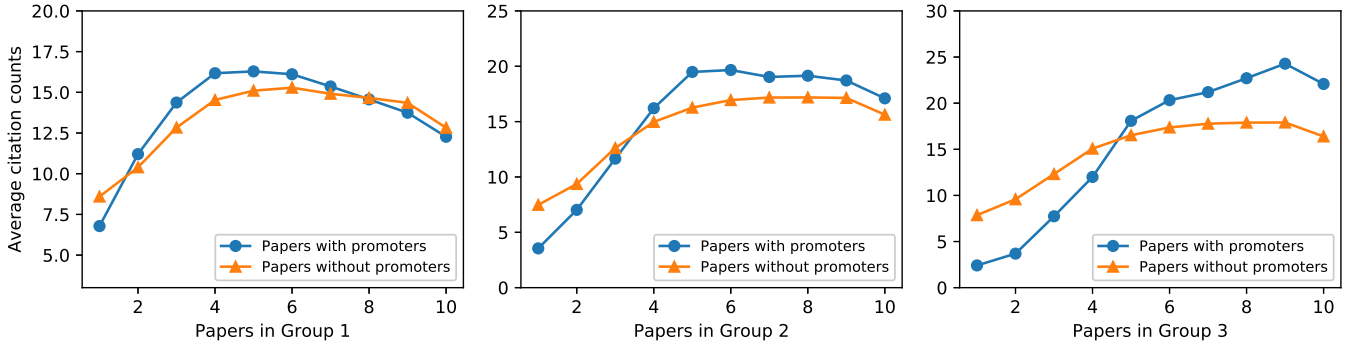


Fig. 1: Annual citation trends of papers with and without citation promoters, while the papers have close performance in terms of early and ten year citation counts.

different data balancing methods. The top 16 important predictive features are identified with the asterisk symbol

(*) in Table 4.

TABLE 5 Prediction results generated by different models. R — Random guessing; XGB — XGBoost; RF — Random forest.

Classifier	Sampling	Class weight	Precision	Recall	F1
R	None	None	0.04	0.49	0.07
XGB	None	None	0.67	0.44	0.54
XGB	SMOTE	None	0.59	0.49	0.54
RF	SMOTE	None	0.56	0.48	0.52
RF	None	N:P=1:5	0.52	0.53	0.53

As seen in Table 5, while the dummy classifier with random guessing could achieve an F_1 score of 0.07 due to imbalanced data, both XGB and RF achieved much better performances with an F_1 score higher than 0.5, a 740% increase compared to that of random guessing. In our experiments, XGB and RF acquired close predictability. RF could be slightly improved if data balancing methods such as SMOTE was employed, but this was not applicable to XGB. Overall, both classifiers achieved moderate results.

5.4 Discussion of Prediction Results

When the citation promoter is applied to citation analysis, it is significant and practical to acquire a high accuracy of prediction if a citing paper would be a citation promoter for its cited paper. However, only a very small portion of citing papers could be identified as citation promoters, making the task an imbalanced classification problem, which is far more difficult than the balanced classification problem. The low F-score results from our experiments reflects the challenge in identifying such citation promoters. A lot of previous studies on improving classification performance for imbalanced data have proposed a sundry of sampling strategies. We have employed popular sampling methods such as SMOTE, but the achieved improvement on the predictability of the classifiers was limited.

A crucial factor affecting the predictive performance of the classifiers could be the lack of content and author related features in our current dataset. The importance of these features has been explored in research studies on citation analysis [7]. Especially, the study of [29] built a classification model to predict whether the published papers would contribute to the h-indexes of the authors within a given time period, in which the content related features played the most dominant role in contribution to the model's predictive performance. Therefore, we believe that a significant improvement of model predictability would be achieved if more features could be extracted and imported to our models.

6 Conclusion

In this paper, we proposed a new concept of citation promoter, which boosts the citation counts of the cited paper shortly after this citation is received. The citation promoter was defined based on the annual citation rate of the cited paper and the co-citation counts received by the pair of cited and citing papers. Through a series of comparative experiments, it was verified that the citation promoters identified by our proposed definition were indeed associated with the growth of citation counts of cited papers. Papers that received a citation promoter at an early age would outperform the other papers in terms of long-term citation counts within a ten year citation window. In addition, the annual citation trends of the papers with citation promoters displayed a sharp rise shortly after the citation promoters were received. However, as shown in previous studies [7], the increase of citation counts could be influenced by a number of factors. Therefore, more experiments demonstrating a directly promoting effect triggered by citation promoters are still needed in future.

Since our definition of a citation promoter relies on the upcoming citation performance of the cited paper, it would be more significant and practical to identify the citation promoter before the promoting effect is exposed. We thus considered it as a binary classification problem to predict whether a citing paper would be a citation promoter for its cited paper. The task was class imbalanced, in which the positive instances accounted for only 4%. Our proposed predictive models achieved moderate results, much better though than random guessing. The limitation of our models could be due to the lack of content and author features.

Given that the citation promoters could be identified promptly and correctly, they could be helpful in the prediction of future citation impact, similar to the work of [12], in which an improvement on predicting long-term citation count was achieved by incorporating the properties of early citers. Discovering sleeping beauties in science is another interesting topic [34]. Citation promoters could facilitate the process of identifying sleeping beauties. We could also employ citation promoters in the analysis of weighted citations. Overall, with the improvement of the prediction accuracy, citation promoters could be a promising concept in a wide range of applications.

We currently see some implications of our work to topics outside the identified scope of citation analysis studies, with some social media platforms as an exception. In the case of Twitter, the citation promoter concept could be extended to retweets. For instance, it is commonly observed that if a particular tweet receives a retweet from an important personality, the number of subsequent retweets increases substantially. Hence, the impact of a tweet could be ascertained by a retweet promoter concept. More importantly, this concept of citation promoters could be potentially used for evaluating the research impact of publications. For instance, if a publication has more citations which are identified as citation promoters, then the publication can be considered to be of good quality compared to other publications which have higher citation counts but less number of citation promoters. In fact, the authors of the citation promoter papers are ideal candidates for co-authorship for future papers because co-authorship is a strong indicator of better research activity and institutional ranking [35]. Secondly, previous studies have shown that just counting citations may not always be useful for identifying highly influential researchers [36]. Hence, more efficient metrics for evaluating citations are required. We foresee a citation promoter count for publications as one such useful metric.

In future work, it would be interesting to extend our

study to other datasets where more features could be collected for developing a more powerful predictor of citation promoters. A more comprehensive investigation on the impact of citation promoters is also planned. While only the computer science discipline was considered in this current study, an extended exploration of citation promoters in other disciplines is also planned.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its Science of Research, Innovation and Enterprise programme (SRIE Award No. NRF2014-NRF-SRIE001-019).

REFERENCES

[1] M. Thelwall, "Interpreting correlations between citation counts and other indicators," *Scientometrics*, vol. 108, no. 1, pp. 337–347, 2016.

[2] 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*, vol. 102, no. 46, p. 16569, 2005.

[3] E. Garfield, "The history and meaning of the journal impact factor," *Jama*, vol. 295, no. 1, pp. 90–93, 2006.

[4] L. Leydesdorff, "How are new citation-based journal indicators adding to the bibliometric toolbox?" *Journal of the Association for Information Science and Technology*, vol. 60, no. 7, pp. 1327–1336, 2009.

[5] M. H. MacRoberts and B. R. MacRoberts, "Problems of citation analysis: A critical review," *Journal of the American Society for information Science*, vol. 40, no. 5, p. 342, 1989.

[6] P. O. Seglen, "Why the impact factor of journals should not be used for evaluating research." *BMJ: British Medical Journal*, vol. 314, no. 7079, p. 498, 1997.

[7] I. Tahamtan, A. S. Afshar, and K. Ahamdzadeh, "Factors affecting number of citations: a comprehensive review of the literature," *Scientometrics*, vol. 107, no. 3, pp. 1195–1225, 2016.

[8] M. Patterson and S. Harris, "The relationship between reviewers quality-scores and number of citations for papers published in the journal physics in medicine and biology from 2003–2005," *Scientometrics*, vol. 80, no. 2, pp. 343–349, 2009.

[9] L. Bornmann, H. Schier, W. Marx, and H.-D. Daniel, "What factors determine citation counts of publications in chemistry besides their quality?" *Journal of Informetrics*, vol. 6, no. 1, pp. 11–18, 2012.

[10] D. W. Aksnes, "Characteristics of highly cited papers," *Research evaluation*, vol. 12, no. 3, pp. 159–170, 2003.

[11] A. Gazni and M. Thelwall, "The long-term influence of collaboration on citation patterns," *Research Evaluation*, vol. 23, no. 3, pp. 261–271, 2014.

[12] M. Singh, A. Jaiswal, P. Shree, A. Pal, A. Mukherjee, and P. Goyal, "Understanding the impact of early citers on long-term scientific impact," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, June 2017, pp. 1–10.

[13] E. Yan and Y. Ding, "Weighted citation: An indicator of an article's prestige," *Journal of the Association for Information Science and Technology*, vol. 61, no. 8, pp. 1635–1643, 2010.

[14] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying meaningful citations." in *AAAI Workshop: Scholarly Big Data*, 2015.

[15] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1247–1252.

[16] P. M. Davis, "Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts?" *Journal of the Association for Information Science and Technology*, vol. 59, no. 13, pp. 2186–2188, 2008.

[17] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with googles pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[18] Y. Pasadeos, J. Phelps, and B.-H. Kim, "Disciplinary impact of advertising scholars: Temporal comparisons of influential authors, works and research networks," *Journal of Advertising*, vol. 27, no. 4, pp. 53–70, 1998.

[19] C. Bergstrom, "Eigenfactor: Measuring the value and prestige of scholarly journals," *College & Research Libraries News*, vol. 68, no. 5, pp. 314–316, 2007.

[20] Y. Ding and B. Cronin, "Popular and/or prestigious? measures of scholarly esteem," *Information processing & management*, vol. 47, no. 1, pp. 80–96, 2011.

[21] F. Luo, A. Sun, M. Erdt, A. Sesagiri Raamkumar, and Y.-L. Theng, "Exploring prestigious citations sourced from top universities in bibliometrics and altmetrics: a case study in the computer science discipline," *Scientometrics*. [Online]. Available: https://doi.org/10.1007/s11192-017-2571-z

[22] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.

[23] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006, pp. 103–110.

[24] C. Chen, F. Ibekwe-SanJuan, and J. Hou, "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis," *Journal of the Association for Information Science and Technology*, vol. 61, no. 7, pp. 1386–1409, 2010.

[25] Y. K. Jeong, M. Song, and Y. Ding, "Content-based author cocitation analysis," *Journal of Informetrics*, vol. 8, no. 1, pp. 197–211, 2014.

[26] A. Callahan, S. Hockema, and G. Eysenbach, "Contextual cocitation: Augmenting cocitation analysis and its applications," *Journal of the Association for Information Science and Technology*, vol. 61, no. 6, pp. 1130–1143, 2010.

[27] T. Yu, G. Yu, P.-Y. Li, and L. Wang, "Citation impact prediction for scientific papers using stepwise regression analysis," *Scientometrics*, vol. 101, no. 2, pp. 1233–1252, 2014.

[28] C. Stegehuis, N. Litvak, and L. Waltman, "Predicting the long-term citation impact of recent publications," *Journal of informetrics*, vol. 9, no. 3, pp. 642–657, 2015.

[29] Y. Dong, R. A. Johnson, and N. V. Chawla, "Will this paper increase your h-index?: Scientific impact prediction," in *Proceedings of the eighth ACM international conference on web search and data mining*. ACM, 2015, pp. 149–158.

[30] G. Colavizza, K. W. Boyack, N. J. van Eck, and L. Waltman, "The closer the better: Similarity of publication pairs at different cocitation levels," *Journal of the Association for Information Science and Technology*, vol. 69, no. 4, pp. 600–609, 2018.

[31] C. Giovanni, K. W. Boyack, N. J. Van Eck, and L. Waltman, "Exploring the similarity of articles co-cited at different levels," in *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics*, 2017, pp. 549–560.

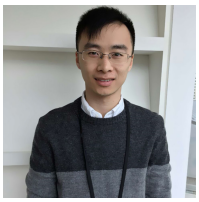
[32] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*. ACM, 2015, pp. 243–246.

[33] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

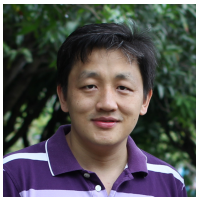
[34] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying sleeping beauties in science," *Proceedings of the National Academy of Sciences*, vol. 112, no. 24, pp. 7426–7431, 2015.

[35] J. D. Linton, R. Tierney, and S. T. Walsh, "Publish or perish: how are research and reputation related?" *Serials Review*, vol. 37, no. 4, pp. 244–257, 2011.

[36] L. Waltman, N. J. van Eck, and P. Wouters, "Counting publications and citations: Is more always better?" *Journal of Informetrics*, vol. 7, no. 3, pp. 635–641, 2013.



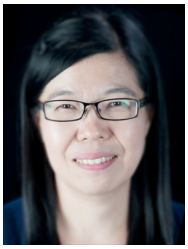
Feiheng Luo received Master degree from The University of Hong Kong in 2014. He has been a research engineer since 2015 at the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore. His research interests include informatics, bibliometrics, altmetrics, and data mining.



Aixin Sun received PhD degree from the School of Computer Engineering, Nanyang Technological University, Singapore, where he is an associate professor. His research interests include information retrieval, text mining, social computing, and multimedia. His papers appear in major international conferences like SIGIR, KDD, WSDM, ACM Multimedia, and journals including Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, and the Journal of the Association for Information Science and Technology.

Aravind Sesagiri Raamkumar received his PhD and Masters degrees from the Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, where he is currently a research fellow. His research interests include information filtering, research metrics, data mining, social media, and scholarly communication.

Mojisola Erdt is a Research Fellow on the Altmetrics project at the Centre for Healthy and Sustainable Cities (CHESS) in the Wee Kim Wee School of Communication and Information at NTU. She obtained her doctorate degree in Information Technology at the Multimedia Communications Lab at the Technische Universitaet Darmstadt, Germany in 2014. Her PhD thesis was on the topic of personalized recommendation of learning resources for resource-based learning.



Yin-Leng Theng is Professor and Director at the Centre of Healthy and Sustainable Cities (CHESS) at the Wee Kim Wee School of Communication and Information, and Research Director at the Research Strategy and Coordination Unit (Presidents Office), Nanyang Technological University (NTU, Singapore). Her expertise is in user-centred design, interaction design and usability engineering, and has been working recently on A*Star and NRF-funded projects on telehealth, patient care, elderly and technologies. She has participated in numerous research projects as PI, Co-PI and Collaborator with more than 200 publications published widely in international journals and conferences.