PAPER

Citation Count Prediction Based on Neural Hawkes Model

Lisha LIU^{†,††}, Student Member, Dongjin YU^{†a)}, Dongjing WANG[†], and Fumiyo FUKUMOTO^{††b)}, Nonmembers

SUMMARY With the rapid development of scientific research, the number of publications, such as scientific papers and patents, has grown rapidly. It becomes increasingly important to identify those with high quality and great impact from such a large volume of publications. Citation count is one of the well-known indicators of the future impact of the publications. However, how to interpret a large number of uncertain factors of publications as relevant features and utilize them to capture the impact of publications over time is still a challenging problem. This paper presents an approach that effectively leverages a variety of factors with a neural-based citation prediction model. Specifically, the proposed model is based on the Neural Hawkes Process (NHP) with the continuous-time Long Short-Term Memory (cLSTM), which can capture the aging effect and the phenomenon of sleeping beauty more effectively from publication covariates as well as citation counts. The experimental results on two datasets show that the proposed approach outperforms the state-of-the-art baselines. In addition, the contribution of covariates to performance improvement is also verified.

key words: citation count, neural hawkes process, covariates, cLSTM

1. Introduction

With the rapid development of scientific research, a large number of publications, such as scientific papers and patents, are published every year in the world. However, it is difficult for even experts in the corresponding field to identify the quality and future impact of so many publications [1], [2]. In other words, it remains a challenging research question on how to find high quality publications from a large volume of publications efficiently.

On the one hand, peer review data can be utilized to measure the quality of academic publications [3]–[5]. It includes effective clues to predict the impact of the paper. However, the data through a long period is difficult to collect, while more recently, the open review system has been available and made it possible to obtain it [6]. Another difficulty to make use of peer review data for citation count prediction is how to effectively extract evidence from the textual review data which supports the quality of the paper. Different reviewers which focus on different aspects of the same paper make it more problematic.

On the other hand, citation count can also be used to

measure the publication quality. Many models, for example, regression model [7], have been proposed for predicting the future citation of a paper. Recently, citation prediction based on deep learning techniques has been intensively studied, which include Bidirectional Long Short-Term Memory (Bi-LSTM) and recurrent neural network [8]. However, existing models mainly explored citation count behavior, without consideration of different publication covariates simultaneously.

Figure 1 shows that the number of citations of different papers and patents at various times is diverse. Moreover, due to lots of influencing factors (time, place, subject, author, etc.) to precisely predict the future citation frequency of the publication still remains a challenging problem. Motivated by the previous works, we propose a novel method for effectively leveraging a variety of factors with a neural-based citation prediction model. Our model is based on the Neural Hawkes Process (NHP) with the continuous-time Long

Fig. 1 The Figure on the left shows the citation count of the randomly selected 40 papers from the Semantic Scholar search service (between 1975 and 2015); The figure on the right gives the citation status of 30 patents randomly selected from the National Bureau of Economic Research (NBER) dataset after publication. The x-axis indicates the year since the paper / patent was published and the y-axis shows the number of cited papers / patents per year.

Manuscript received March 9, 2020.

Manuscript revised June 19, 2020.

Manuscript publicized August 3, 2020.

[†]The authors are with the College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China.

^{††}The authors are with the Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, Faculty of Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.

a) E-mail: yudj@hdu.edu.cn (Corresponding author)

b) E-mail: fukumoto@yamanashi.ac.jp (Corresponding author)

DOI: 10.1587/transinf.2020EDP7051

Short-Term Memory (cLSTM). NHP is a model that captures streams of discrete events in continuous time. It constructs a neutrally self-modulating multivariate point process that evolves the intensities of multiple event types by utilizing a novel continuous-time LSTM. It enables to capture the aging effect and the phenomenon of sleeping beauty more effectively from publication covariates as well as citation counts. Firstly, we use the number of citations from the history of publications to establish the citation prediction model based on the Neural Hawkes Process. Then, we combine the publication's various covariates with continuous-time LSTM (cLSTM) to build the Neural Hawkes-Process based Citation Prediction Model (NHPCM). Finally, we use relevant hyperparameters to optimize our model and make it more robust.

The main contributions of this work can be summarized as follows: (1) We propose a method for citation count prediction which effectively leverages a variety of covariates such as the author's h-index and venue-wise with a neural-based citation prediction model. (2) We introduce the NHP which is coupled with the cLSTM to capture the aging effect and the phenomenon of sleeping beauty more effectively from publication covariates as well as citation counts. (3) Experimental results on two real datasets show that our method outperforms the state-of-the-art baselines. In addition, the contribution of covariates to performance improvement is also verified.

The remainder of this paper is structured as follows. Section 2 introduces related works, focusing on the achievements of citation prediction in recent years. After Sect. 3 gives the detail of the proposed model and related algorithms Sect. 4 demonstrates the experimental results as well as the analysis. In the last section, we summarize the paper and outline future works.

2. Related Work

In this section, we briefly review some related works, including the studies on the influence factors of citation frequency and the techniques of citation prediction, and discuss the significance of our approach.

Influence factors of citation frequency. The citation of publications is a complex act that is influenced by many factors. Many researchers have attempted to predict the future citation of a paper by utilizing citation count [9]. For academic papers, Redner et al. mentioned the basic information about the relative popularity of citation distribution for scientific publications and provided a complete measure of popularity rather than the average or the total number of citations [10]. Kuo et al. proposed a method for journal validity measurement based on reliability-based citation impact factor (R-impact), which is widely used in short-term (two-year) to measure the life cycle of published articles and the marketability of published articles [11]. Ke et al. emphasized a complex feature of citation dynamics ("sleeping beauty") that provides empirical evidence against the use of short-term citation indicators to quantify scientific

impact [12].

Meanwhile, some authors have attempted to take other factors into account. For example, Jian et al. analyzed the influence mechanism of several basic factors, such as the average review period, the average number of references and the annual distribution of references, on the trend of impact factors, especially among journals in different disciplines [13]. Martinez-Ruiz et al. focused on patents and proposed five factors related to patent value: patent value, technological usefulness of the invention, knowledge stock used by the company to create the technology, technological scope of the invention, and international scope of protection [14]. Liu et al. divided the quality of patents into two types of attributes, i.e., intrinsic (static) features and time-varying features to predict future citations [15]. The results demonstrated that the model integrating two types of attributes have contributed to improving the overall performance. While conceptually similar, our model differs from Liu's approach in that our model is based on cLSTM which can eliminate the independent and additional influence of historical citation events.

Citation count-based predicting approaches. There are at least two main paradigms for predicting the frequency of citations [1]. One is to establish a citation network model and use graph mining techniques to solve the problem of citation quantity prediction. Pobiedina et al. have established a feature, called, GER-score, which estimates the number of citations for future academic publications [16]. Butun et al. attempted to define the number of citations for predictive scientists as link prediction problems in directed, weighted, and time-referenced networks [17]. The results by using 12 undersampled datasets from the Aminer and HEP-Th dataset show that the method achieved good performance. Chuan et al. Proposed a mathematical concept of link prediction in a co-author network and a link prediction algorithm based on topic modeling to determine new interactions between authors that may occur in the future [18]. However, the method based on link prediction cannot perform dynamic reference counting at any point in the future as it is difficult to capture the various patterns in citation count which change across periods.

Other approaches focusing on the issue of citation prediction correlate time series with the number of citations or fit these time series with functions of certain classes. Some commonly used functions include regression model [7], the counting process [19], the point process or the specific Poisson Process [20], reinforces the Poisson Process (RPP) [21], self-excited Hawkes Process [22], RPP-self-excited Hawkes Process [23], and a long-term citation count prediction model with long short-term memory units [24]. Mei et al. proposed three approaches, i.e., Self-Modulating Multivariate Point Process, Hawkes Process with Inhibition, and Neural Hawkes Process [25]. They evaluated their methods by using synthetic and real-world media dataset and showed that the methods improved the prediction performance on the course of future events. Although our approach is based on the approaches of that of

Mei et al., the network structures are different. We incorporate multiple features into NHP with cLSTM and learns these features as well as citation counts simultaneously.

3. Model and Algorithm

3.1 Problem Definition

Let $\{C_i^d\}_{i=0}^T$ be the citation sequence of the publication d at the past timestamps $0, \dots, T$. The goal of this model is to predict the number of future citations of the publication d , i.e., $C_i^d, i \geq T+1$. In this paper, we use lowercase letters for model parameters, bold lowercase letters for vector, and no-bold lowercase letters for scalar (such as α). Table 1 shows a list of notations used in this paper.

3.2 The Proposed Model

The intensity function that predicts the number of citations of a single publication d is used within the time interval $[0, \infty)$ at the beginning of its published year, and the timestamp sequence is denoted by $\{t_i^d\}_{i=0}^\infty$, where $0 = t_0^d \leq t_1^d, \dots, t_i^d, \leq \dots \leq t_\infty^d$. For the number of citations per year of a single publication, we use event type to describe it. In other words, for a certain publication d , there is a citation sequence (k_i, t_i) , where each k_i ($k \in 0, 1, 2, \dots, K$) is an event type (number of citations) and $0 < t_1 < t_2 < \dots < t_i < T$ is the time when it was cited.

Next, we introduce three models, namely Hawkes-Process based Citation Prediction Model (HPCM), Inhibition Hawkes-Process based Citation Prediction Model (IHPCM) and Neural Hawkes-Process based Citation Prediction Model (NHPCM). Overall speaking, these models are based on (temporal) point processes and can model the data sequential generated in many real scenarios, such as earthquake location, equipment failure inspection, and customer purchase behavior.

HPCM: Hawkes-Process based Citation Prediction Model. The most basic event-time model, the Hawkes process, also known as the **self-excited point process** [26], has many applications in science and engineering. Hawkes process considers that the excitation of past events is positive, additive and decaying over time. For example, for scientific papers, it is generally believed, the number of citations per year is higher in the first few years, and then decreases year by year. However, for the paper with “sleeping beauty”, this is not true.

Table 1 Notations used in this paper

Notation	Description
d	the publication
$\mathbf{q}_d(t)$	the sequence of covariate about publication d at time t
k_i	the sequence of citation count
α, β, δ	coefficients that affect the number of citations
$\mathbf{w}, \mathbf{v}$	the vector coefficients corresponding to citations and covariates
$\lambda_k(t)$	the intensity function

In this paper, the basic Hawkes process is introduced into the space-time sequence relation of citation frequency count of publications. In the time period $[t, t+dt]$, $\lambda_k(t)$ gives the intensity of the number of citations in the publication (that is, the event of type), known as the **intensity function**, as indicated by Eq. (1):

$$\lambda_k(t) = \beta \mathbf{q}_d(t) + \sum_{h:t_h < t} \alpha_{k_h, k} e^{-\delta(t-t_h)} \quad (1)$$

where $\mathbf{q}_d(t) = (q_{1d}, q_{2d}, \dots, q_{nd})$ is the row covariable of publication d , β is the column coefficient relative to the covariable, and $\alpha_{k_h, k}$ means the relationship between the historical citations of the publication (k_h) and the current citations (k), which is always positive in Hawkes Process. Besides, $e^{-\delta(t-t_h)}$ is the decay kernel over time, and δ is the time coefficient of historical reference number excitation.

IHPCM: Inhibition Hawkes-Process based Citation Prediction Model. The positive attribute conditions in HPCM may limit the exact representation of the number of predicted references. The ordinary Hawkes process will return to a steady state within a certain range of parameters. As shown in Eq. (1), $\alpha_{k_h, k}$ will always maintain the state of $\alpha_{k_h, k} \geq 0$. In order to eliminate the adverse effects of positive monotony in the future publications, we give another model, **IHPCM** (Inhibition Hawkes-Process based Citation Prediction Model), whose final intensity function $\lambda_k(t)$ can “adjust” the number of single publication citations. In other words, the intensity density function obtained by IHPCM can be either positive or negative.

To maintain the most basic decomposable properties of HPCM, we firstly keep the basic shape of the Hawkes process and then combine historical citation data with a **non-linear activation function** $f_k: R \rightarrow R_+$ to control the intensity of citation events in future publications. In this way, the positive constraint on $\alpha_{k_h, k}$ will be relaxed, allowing it to exceed a certain range, which is more in line with the publication citations in the real-world.

$$\lambda_k(t) = f_k(\tilde{\lambda}_k(t)) \quad (2)$$

$$\tilde{\lambda}_k(t) = \beta \mathbf{q}_d(t) + \sum_{h:t_h < t} \alpha_{k_h, k} e^{-\delta(t-t_h)} \quad (3)$$

In Eq. (2), we use the non-linear activation “Softplus” function for multi-class neural network output. Figure 2

Fig. 2 “Softplus” function and “ReLU” function.

illustrates “Softplus” and “ReLU” functions. We can see that “Softplus” can be regarded as “ReLU” smoothing as $m \rightarrow 0$, it is close to “ReLU”. “Softplus” can be regarded as “ReLU” smoothing (as shown in Fig. 2). We apply the scaled “Softplus” function $f(x) = m \log(1 + e^{\frac{x}{m}})$ to IHPCM. When $m \rightarrow 0$, it is close to “ReLU”. Therefore, Eq. (2) can be rewritten as $\lambda_k(t) = m_k \log(1 + e^{\frac{\lambda_k(t)}{m_k}})$. In addition, the intensity $\lambda_k(t)$ of the cited event may be excited (risen) or suppressed (declined) as the time passes. But in the end, it will be close to the basic publication covariates rate $\beta q_d(t)$.

NHPCM: Neural Hawkes-Process based Citation Prediction Model. To eliminate the independent and additional influence limits of historical reference numbers on $\lambda_k(t)$, NHPCM (Neural Hawkes-Process based Citation Prediction Model) uses a special recurrent neural network, i.e., continuous-time Long Short-Term Memory (cLSTM), combined with the Hawkes process. It can predict the probability of the citation frequency of a publication in the future by studying its citation frequency, time, and covariates.

In NHPCM, a continuous-time single-layer LSTM is used to predict the citation frequency and time of the publication. As mentioned earlier, the citation frequency event k of each publication d has a time-varying intensity $\tilde{\lambda}(t)$. When a new citation frequency event occurs, the dynamics of these changes are caused by the hidden state vector $\mathbf{h}(t)$. To be controlled, the corresponding $\mathbf{h}(t)$ is in turn associated with memory-cell $\mathbf{c}(t)$ in the continuous-time LSTM. The difference between the cLSTM and the normal LSTM in predicting the citation count of a publication is that in the interval after each citation event occurs, each memory-cell $\mathbf{c}(t)$ will decay exponentially toward the steady state-value $\bar{\mathbf{c}}(t)$ at rate δ . In addition, we adopt a linear expression as the covariate f_d related to publications, in which the covariate of the publication to optimize the intrinsic quality correlation coefficient v of the publication d . Finally, we take advantage of the idea of the Hawkes process, that is to say, the nonlinear activation function softplus is used to combine the cited frequency and covariate of the publication to obtain the intensity function $\lambda_k(t)$ of non-monotone fluctuation.

Here, Eq. (4) gives the definition of the intensity function of NHPCM at $t > 0$, where Eqs. (5) and (6) explain how the cLSTM’s hidden unit $\mathbf{h}(t)$ is obtained according to memory-cell $\mathbf{c}(t)$:

$$\lambda_k(t) = f_k(\mathbf{w}_k^T \mathbf{h}(t) + \mathbf{v}_k^T \mathbf{q}_k(t)) \quad (4)$$

$$\mathbf{h}(t) = \mathbf{o}_t \Theta(2\sigma(2\mathbf{c}(t)) - 1) \quad (5)$$

$$\mathbf{c}(t) = \bar{\mathbf{c}}_{i+1} + (\mathbf{c}_{i+1} - \bar{\mathbf{c}}_{i+1})e^{-\delta_{i+1}(t-t_i)} \quad (6)$$

As with the previous two models, $\lambda_k(t)$ here has some deterministic trend between the frequency of publication citations. The input vector $k_i \in \{0, 1, 2, \dots, K\}$ of the neural model is the number of citations (not including 0) per year since the release of the publication. Here, one-hot encoding is used for the vector k_i to enhance the non-linearity of the model. Besides, $\mathbf{h}(t)$ in the model has the strong correlation with the historical reference numbers. Similar to that in the previous LSTM model [27], $\mathbf{h}(t)$ can capture the sequence

of historical reference numbers $(k_1, k_2, \dots, k_i)$. However, in the NHPCM decay structure, it also reflects the time series $(t_1 - 0, t_2 - t_1, \dots, t_i - t_{i-1})$ in which the publication is referenced. Further, Eq. (6) shows that the memory-cell $\mathbf{c}(t)$ is determined by historical reference numbers. During the interval $(t_{i-1}, t_i]$, cLSTM updates the current memory-cell $\mathbf{c}(t)$ to a new initial state $\mathbf{c}_{(i+1)}$ based on the current hidden state $\mathbf{h}(t)$, and then reads (k_i, t_i) . In such way, $\lim_{t \rightarrow t_i} \mathbf{c}(t) = \mathbf{c}_{i+1}$. In addition, the linear input model’s input feature $\mathbf{q}_k(t)$ (i.e. covariates) is normalized to better reflect the comparability of different features.

Compared to other models used to predict the citation count of publications, NHPCM has its special characteristics. Firstly, it tends to be non-monotonic and different event type k updates $\mathbf{h}(t)$ with different attenuation rates and turns to $\lambda_k(t)$. Secondly, we consider the various features of a single publication and the interaction between them, which enhances the credibility of the model. Thirdly, a reasonable activation function $f_k: R \rightarrow R_+$ makes the network’s approximation ability wider. Finally, this model is based on a combination of continuous-time LSTM and Hawkes processes, making the model logically rigorous.

3.3 Learning Algorithm

Through the log-likelihood function, we can obtain a generation point process model that fully fits the cited data of the publication, based on the observed citation sequence. Specifically, for the N observed sequence of publications $S = \{S_1, S_2, S_3, \dots, S_N\}$, each sequence is represented as $S_j = \{t_i^d\}_{i=0}^\infty$. Here, j is any index between 1 and N , $t_0 = 0$, and $[0, T]$ is called the time window of publications. According to Eqs. (1) ~ (6), the probability density function of the cited event at the time t_i can be expressed as:

$$P_k(t_i | t_1, t_2, t_3, \dots, t_{i-1}) = \lambda_k(t_i) e^{-\int_{t_{i-1}}^{t_i} \lambda_k(t) dt} \quad (7)$$

Therefore, the likelihood function of the entire model citation sequence can be calculated by Eq. (8), that is, the sum of the log strengths of the cited events occurring within the time when the publication is cited, minus the total intensity integral of the historical observation reference interval $[0, T]$:

$$\begin{aligned} L_k &= \log \prod_{i=0}^{N_k} P_k(t_i | t_1, t_2, t_3, \dots, t_{i-1}) \\ &= \sum_{i=0}^{N_k} \log \lambda_k(t_i) - \int_0^T \lambda(t) dt \end{aligned} \quad (8)$$

Here, we use the Monte Carlo algorithm to locally maximize L_k . Because we cannot guarantee all the prior information and covariates information related to publications in the essay environment, we calculate the approximate result on random sampling. Besides, as the sampling increases, the probability that the result is correct will gradually increase, and the result can be obtained by back-propagation. The Monte Carlo-based algorithm shown in

Algorithm 1. Integral Estimation (Monte Carlo)

```

1: Input: time interval  $[0, T]$ ; type event sequence  $(k_i, t_i)$ ; co-
   variates  $\mathbf{g}$  for each publication  $d$ 
2: Initialize: model parameters;
3:  $E \rightarrow 0$ ;  $\nabla E \rightarrow 0$ 
4: While  $D$  publications:
5:    $t \sim (0, T)$ 
6:   While  $k \rightarrow 1$  to  $K$  :
7:      $E += \lambda_k(t)$ 
8:      $\nabla E += \nabla \lambda_k(t)$ 
9:   End While;
10:   $E \rightarrow \frac{ET}{N}$ ;  $\nabla E \rightarrow \frac{\nabla ET}{N}$ 
11: End While;
12: Return  $(E, \nabla E)$ .

```

Fig. 3 Algorithm: integral estimation (Monte Carlo)

Fig. 3 averages multiple publication samples to reduce the variance of the model estimates.

The method of processing the integral in the likelihood function (Eq. (8)) is to give an unbiased estimate of the entire integral of the intensity estimate $\lambda_k(t)$ at the random time $t \sim (0, T)$, i.e., the expected value E . Since the gradient will vary with expectation, the unbiased estimate is ∇E .

4. Experiments and Discussion

We performed the experiments on two benchmark datasets to evaluate the properties of our models: Semantic Scholar dataset and National Bureau of Economic Research dataset.

4.1 Data Collection and Pre-Processing

4.1.1 Semantic Scholar Dataset

Data Collection. The Semantic Scholar[†] was founded in 2015 by a non-profit organization Allen, to create a scientific journal article search engine for AI support. As a search engine retrieves papers from several domains including machine learning, natural language processing, machine vision, computer science, earth science, and neuroscience, the collected data intuitively represents important elements of the paper (key figures, data analysis and graphical representation of citations) without extensive reading [28]. Compared to Google Scholar^{††} and PubMed^{†††}, Semantic Scholar aims to highlight the most influential and important papers and determine the connections between them. As of August 2019, the number of papers included in The Semantic Scholar has grown to more than 173 million. In addition, Lucaweih^{††††} had summarized the dataset from 1975 to 2015 for detailed classification. To facilitate the experiment, we downloaded the dataset of Lucaweih.

Pre-processing. The original dataset contains a total number of 1,193,650 articles published from 1975 to 2015. We selected 252,928 papers that were cited more than 3 times.

We counted the number of citations of these papers for each year and treated them as the event types. If the number of references in a certain year is 0, the count is automatically skipped.

Another remarkable data pre-processing is to extract the corresponding features (covariates) of the paper. The probability of a paper being cited is not random, and the author, the environment, the article itself and some other factors play crucial roles in citation prediction. For example, in [29], Yaminfirooz et al. indicate that journal IF, journal rank, journal subject quartile, and the first communication author's h-index, the number of documents produced, Scientific Journal Rankings (SJR), Source Normalized Impact per Paper (SNIP), and other variables are significantly positively correlated with the citation of the paper. In contrast, other variables, including the age of the paper, the type of paper, the number of references, the number of authors, and the type of index institutions and journals are not a good indicator to predict future citation count of the paper. However, this conclusion was based only on 200 papers searched from Scopus, which may be not comprehensive.

As shown in Table 2, we selected 63 covariates and classified them into three groups according to [30]. The three groups are Author features, Venue features, and Paper features. Each group consists of 46, 8, and 9 covariates, respectively. We fed the normalized data into our model for training and testing.

Figure 4 demonstrates the distribution of citations from the Semantic Scholar which gives the hint that the author's h-index, the site factor, and the number of authors of the paper, among the above-mentioned 63 covariates, may be the effective factors to estimate citation count in the future.

4.1.2 National Bureau of Economic Research (NBER) Dataset

Data Collection. Founded in 1920, the National Bureau of Economic Research (NBER)^{††††} is the leading non-profit economic research organization in the United States [31]. NBER provides downloadable patent data, which can be used to recall all cited invention patents (more than 3 million) in the US Patent and Trademark Office database from January 1963 to December 1999.

Pre-processing. The NBER patent data set includes all utility patents granted over 37 years (1963-1999), and the total number of patents is 3,258,856. We first screened patents that have been cited for a total of more than 5 years since the patent was issued. Then from the 537,857 patents obtained in the first step, we randomly selected 10,000 patents as our experimental data set. We further selected and normalized 6 patent features, as shown in Table 3. Finally, the number of past patent citations of a single patent and its features are used to predict the number of future patent citations.

[†]<https://www.semanticscholar.org/>

^{††}<https://scholar.google.com/>

^{†††}<https://www.ncbi.nlm.nih.gov/pubmed/>

^{††††}<https://github.com/Lucaweih/impact-prediction/>

^{†††††}<http://data.nber.org/patents/>

Table 2 Paper related covariates - author, venue and paper features

Author (Feature type)	
Covariates	Description
author_h-index_ {mean, min, max}	H-indices of authors
author_h-index_delta_ {mean, min, max}	Change in h-indices of authors in the last 2 years
author_citation_ {mean, min, max}	Cumulative citations for each author
author_citation_delta_ {mean, min, max}	Change in cumulative citation for each author in the last 2 years
author_key_citation_ {mean, min, max}	Cumulative key citation for each author
author_key_citation_delta_ {mean, min, max}	Change in cumulative key citation for each author in the last 2 years
author_mean_citation_ {mean, min, max}	Mean citation per paper for each author
author_mean_citation_delta_ {mean, min, max}	Change in mean citation per paper for each author in the last 2 years
author_paper_ {mean, min, max}	Number of papers published for each author
author_paper_delta_ {mean, min, max}	Number of papers published for each author in the last 2 years
author_unweighted_pagerank_ {mean, min, max}	PageRank of each author in the unweighted co-authorship network
author_weighted_pagerank_ {mean, min, max}	PageRank of each author in the weighted co-authorship network
author_mean_citation_rank_ {mean, min, max}	Rank of each author among all authors in terms of mean citation per paper
author_recent_num_coauthor_ {mean, min, max}	Number of coauthors each author had in the last 2 years
author_max_single_paper_citation_ {mean, min, max}	Maximum citation a single paper of each author has received
total_num_authors	Total number of authors for the paper
Venue (Feature Type)	
Covariates	Description
venue_h-index	H-index of the venue
venue_h-index_delta	Change in h-index of the venue in the last 2 years
venue_mean_citation	Mean citation per paper published in the venue
venue_mean_citation_delta	Change in mean citation per paper published in the venue in the last 2 years
venue_paper	Number of papers published in the venue
venue_paper_delta	Number of papers published in the venue in the last 2 years
venue_rank	Rank of the venue among all venues in terms of mean citation per paper
venue_max_single_paper_citation	Maximum number of citation any paper published in the venue has received
Paper (Feature Type)	
Covariates	Description
paper_age	Age of the paper in years (rounded up)
paper_citation	Cumulative citation count
paper_key_citation	Cumulative key citation count
paper_mean_citation_per_year	Average number of citation received per year
is_survey	Whether or not the paper is a survey
paper_citation_delta_{0,1}	Number of citation the paper received in the last year and the year before that
paper_key_citation_delta_{0,1}	Number of key citation the paper received in the last year and the year before that

Fig. 4 The visualization of the author's h-index, the site factor, and the number of authors of the paper, respectively, where the x-axis shows the corresponding covariate, and the y-axis indicates the citation count of the paper.**Table 3** Patent related covariates

Covariates	Description
asscode	Assignee Type
nclass	Main Patent Class (3 digits)
cat	Technological Category
subcat	Technological Sub-Category
creceive	Number of Citations Received
general	Measure of Generality

Table 4 The value of the log-likelihood of seq against the different number of hidden unit dimensions. The figures in bold indicate the best values of log-likelihood.

The # of hidden unit dimensions	Hawkes	16	32	64
Paper	-69.8045	-0.2689	-0.2526	-0.2512
Patent	-23.6921	-0.3449	-0.2978	-0.3184

The # of hidden unit dimensions	128	256	512	1,024
Paper	-0.2369	-0.2223	-0.2413	-0.2595
Patent	-0.3413	-0.3444	-0.3407	-0.3126

4.2 Experimental Setting and Evaluation Metrics

Experimental setting. Our experiments were run on RTX 2070 super, Intel Core i9 (3.6 GHz) machine. The experimental optimizer used Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (ADAM). The learning rates of SGD and ADAM were both initialized to 0.001, which achieved better convergence. In addition, the larger the number of hidden unit dimensions in cLSTM, the more details, and richer expression capabilities. However, it also brought about the disadvantages of overfitting and time-consuming. Moreover, different datasets will also affect the log-likelihood results. We thus tuned the number of hidden unit dimensions. Table 4 shows the log-likelihood of “seq” values by using the different dimensions of the hidden unit with cLSTM where “seq” means the combined prediction of “type” and “time”. As we can see from Table 4, the number

Table 5 The hyperparameters used in the experiment.

Hyperparameters	Value
Train Ratio	1.0
Dimension of cLSTM	256 (paper) / 32 (patent)
Track Period	1,000
Max Epoch	50
Size Batch	10
Optimizer	ADAM/SGD
Learning Rate	0.001

of dimensions of the hidden unit is not necessary as large as possible because on both of the datasets, the best values of log-likelihood were obtained (-0.2223 / -0.2978) when we used 256 / 32 dimensions. Therefore, for the Semantic Scholar dataset, we used 256, and for the National Bureau of Economic Research dataset, we used 32 as the number of hidden unit dimensions. Besides, some hyperparameters used in our model are shown in Table 5.

Evaluation metrics. To evaluate the proposed approach and compare it with other approaches, we use three general criteria, namely, log-likelihood ratio, accuracy (ACC), and coefficient of determination (R^2). Log-likelihood ratio denotes the one with the most possibility of this result as a true estimate, as shown in Eq. (8). ACC denotes the ratio of the correct number of predictions divided by the total number of correct values on the test data. R^2 is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

where n is the number of predictions, y_i and $\bar{y}$ refer to each actual number of citations and the mean of the actual number of citations, respectively. f_i indicates the prediction value of the paper i .

4.3 Results and Discussion

In order to investigate if and how our model outperforms other models based on Neural Hawkes Process, we compare our model NHPCM with the following four models: (i) Neural Hawkes Process (HP), (ii) HP combined with a non-linear activation function (DHP), (iii) DHP combined with cLSTM (NHP), and (iv) DHP with covariates (IHPCM). The first three models are presented by Mei et al. [25]. In Table 6, we compare the log-likelihood of all models on two datasets, which is a good indicator frequently used to evaluate how the covariates influence the model.

As can be seen clearly from Table 6, our model NHPCM achieves the best performance on each dataset for all three results, i.e., *seq*, *type* and *time*, because it learns the past citation sequence and publication covariates simultaneously. Table 6 also indicates that on each dataset, the non-linear function is effective because the results obtained by DHP show a 0.6046 ~ 63.9201 improvement over the model HP on both datasets. Besides, our model achieves the most effective results for predicting the time when a certain number of citations is reached after the paper has been

Table 6 Log-likelihood of five models. The first three are the results that do not contain the publication covariates, while the latter two make use of the publication covariates. “Type” shows the prediction of the type sequence event citation each year since the publication of the publication, “time” shows the time prediction of the number of citations after publication, and “seq” shows the combined prediction of “type” and “time”. The figures in bold indicate the best values of log-likelihood.

(a) The Semantic Scholar dataset			
Results	seq	type	time
HP	-69.8045	-2.4990	-67.3055
DHP	-5.8844	-1.8944	-3.9900
NHP	-2.8230	-1.6546	-1.1685
IHPCM	-0.3360	-0.1749	-0.1611
NHPCM	-0.2223	-0.1405	-0.0818

(b) The National Bureau of Economic Research dataset			
Results	seq	type	time
HP	-23.6921	-3.5866	-20.1055
DHP	-5.2100	-1.3303	-3.8796
NHP	-3.7050	-1.1190	-2.5860
IHPCM	-0.3240	-0.0996	-0.2244
NHPCM	-0.2978	-0.1247	-0.1731

Fig. 5 The impact of different feature classes (author features, venue features, and paper features) on the number of times a paper is cited.

published or the patent has been granted. Similarly, cLSTM works well as the results by NHP show a 0.2113 ~ 3.0614 improvement over DHP. We note that the results obtained by our model NHPCM show a 0.9943 ~ 3.4072 improvement over the model NHP which does not integrate publication covariates. This demonstrates that publication covariates are very effective for citation count prediction.

To explore the impact of different types of covariates on the number of citations in publications, we compare three types of covariates (author, venue, and paper) of the paper from the Semantic Scholar dataset. As illustrated in Fig. 5, when we use only the paper features, the values of log-likelihood of being cited achieves the highest among the results of *seq* and *time*, compared with those when only using author features or venue features. In fact, the paper features, such as cumulative citation count and the average number of citations received per year, play a very important role for identifying the quality of the paper.

Further, we compare our model with five state-of-art baseline models on the same datasets. These models are:

1. Autoregressive-Integrated-Moving-Average model

(ARIMA): It models the next step in the sequence as a linear function of observations and residual errors at previous time steps, which is the most common model in statistical models used for time series prediction.

2. **Exponential Smoothing (ES):** Exponential smoothing is proposed by Robert G. Brown, it is actually a special weighted moving average model, which is characterized by giving different weights to the past observations.
3. **Long Short-Term Memory (LSTM):** Long Short-Term Memory is an artificial recursive neural network (RNN) architecture for deep learning [32], [33]. It is suitable for processing and predicting important events with very long intervals and delays in time series.
4. **Bi-directional LSTM (BLSTM):** Bi-directional LSTM is an extension of traditional LSTM and it trains two LSTMs on the input sequence instead of one LSTM. The first on the input sequence stays the same, an inverse copy of the second input sequence.
5. **Hawkes Process (HP):** The Hawkes Process is defined as a mathematical model of a self-excited process. It is a counting process that simulates some types of arrival sequence over time. Liu et al. designed special Hawkes Process models with time-varying features and a time-delayed effect kernel to predict self or non-self citations types [15].

Considering the accuracy (ACC) and coefficient of determination (R^2) are frequently used to evaluate models of citation count predication, we summarize these results in Fig. 6 and Table 7, respectively.

As of NHPCM, the accuracy values obtained from the Semantic Scholar dataset and the National Bureau of Economic Research dataset are 0.7081 and 0.8386 respectively, while the values of R^2 are 0.7793 and 0.6402 respectively. We can find from Fig. 6 that the prediction accuracy of our model ranks No. 1, followed by the HP model which is most similar to ours. In addition, the accuracy of our model drops more slowly compared with all other models. It is also noteworthy that all models perform better on the NBER dataset than on the Semantic Scholar dataset, as the accuracies drop much slower with the increasing number of years on the NBER dataset compared with those on the Semantic Scholar dataset.

As we know R^2 shows how well a regression line fits. From this point of view, Table 7 indicates that LSTM and BLSTM do not work well, especially when we use a large volume of the data, i.e., the Semantic Scholar dataset. By comparison, we find HP and our model preserve the good performance on both datasets, and our model achieves the best results.

In the current experiment, we combined the covariates of the paper and the patent to predict the number of future citations. However, the volume of covariate data is large which may include some interference items affecting overall performance. How the volume of covariate data affects the performance is a rich space for further exploration. More-

Fig. 6 The citation prediction accuracy (ACC) of five different models compared with our model. The x-axis gives the number of years since the publication and the y-axis indicates the accuracy of the prediction, reflecting the extent to which the prediction result is close to the true value. (a)-(e) illustrate the results based on the Semantic Scholar dataset and (f)-(j) show the results based on the NBER dataset.

Table 7 The R^2 result of six different models. When $R^2 \rightarrow 1$, the model is better. The figures in bold indicate the best values of R^2 .

Model	Paper R^2	Patent R^2
ARIMA	0.4353	0.4000
ES	0.5444	0.3333
LSTM	0.1843	0.2450
BLSTM	0.2909	0.2664
HP	0.5699	0.5153
NHPCM	0.7793	0.6402

over, we currently use only two datasets for comparative experiments. We need to evaluate our method by using other datasets for quantitative evaluation.

5. Conclusion

We have presented an approach to the prediction of future citations based on the Neural Hawkes Process which is coupled with the continuous-time Long Short-Term Memory to leverage a variety of features as well as citation counts. The comparative results with the traditional prediction models and recent popular models in predicting future citations show that our model is competitive and robust against the different volumes of the dataset. We also demonstrated how feature integration contributes to performance improvement. However, there are still a number of interesting directions for future research. We need to obtain further advantages by investigating the features needed to improve the performance. The datasets we used in the experiment are a period of 25 and 41 years. To further evaluate the robustness of our model, we are going to examine the different span of the dataset, especially how our model affects the overall performance for a short period of time span. In addition, we will extend our model to handle textual features from publications to predict the impact on citation frequency, such as keywords of publications, abstracts of academic papers,

polarity information from the review texts by utilizing the sentiment analysis, and other important information.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No. 61472112, No. 61702144), and Key Science and Technology Project of Zhejiang Province of China (No. 2017C01010).

References

- [1] S. Yuan, J. Tang, Y. Zhang, Y. Wang, and T. Xiao, "Modeling and predicting citation count via recurrent neural network with long short-term memory," arXiv preprint arXiv:1811.02129, 2018.
- [2] L. Leydesdorff, P. Wouters, and L. Bornmann, "Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—state-of-the-art report," *Scientometrics*, vol.109, no.3, pp.2129–2150, 2016.
- [3] S. Li, W.X. Zhao, E.J. Yin, and J.R. Wen, "A neural citation count prediction model based on peer review text," *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.4916–4926, 2019.
- [4] R. Smith, "Peer review: A flawed process at the heart of science and journals," *Journal of the Royal Society of Medicine*, vol.99, no.4, pp.178–182, 2006.
- [5] V.D. Kosteas, "Predicting long-run citation counts for articles in top economics journals," *Scientometrics*, vol.115, no.3, pp.1395–1412, 2018.
- [6] S.B. Bruns and D.I. Stern, "Research assessment using early citation information," *Scientometrics*, vol.108, no.2, pp.917–935, 2016.
- [7] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," *Proc. 20th ACM international conference on Information and knowledge management*, pp.1247–1252, 2011.
- [8] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol.45, no.11, pp.2673–2681, 1997.
- [9] M. Gutiérrez-Salcedo, M.Á. Martínez, J.A. Moral-Munoz, E. Herrera-Viedma, and M.J. Cobo, "Some bibliometric procedures for analyzing and evaluating research fields," *Applied Intelligence*, vol.48, no.5, pp.1275–1287, 2018.
- [10] S. Redner, "How popular is your paper? An empirical study of the citation distribution," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol.4, no.2, pp.131–134, 1998.
- [11] W. Kuo and J. Rupe, "R-Impact: Reliability-Based Citation Impact Factor," *IEEE Trans. Rel.*, vol.56, no.3, pp.366–367, 2007.
- [12] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying sleeping beauties in science," *Proc. National Academy of Sciences*, vol.112, no.24, pp.7426–7431, 2015.
- [13] Z. Jian, N. Cai, and T. Zong-Yuan, "Analysis of effects to journal impact factors via citation networks generated by distributed parallel model," *IEEE Access*, vol.7, no.1, pp.19775–19781, 2019.
- [14] M.-R. Alba and T. Aluja-Banet, "Toward the definition of a structural equation model of patent value: PLS path modelling with formative constructs," *REVSTAT Statistical Journal*, vol.7, no.3, pp.265–290, 2009.
- [15] X. Liu, J. Yan, S. Xiao, X. Wang, H. Zha, and S.M. Chu, "On predictive patent valuation: forecasting patent citations and their types," *Proc. Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*, pp.1438–1444, 2017.
- [16] N. Pobiedina and R. Ichise, "Citation count prediction as a link prediction problem," *Applied Intelligence*, vol.44, no.2, pp.252–268, 2016.
- [17] E. Bütün and M. Kaya, "Predicting Citation Count of Scientists as a Link Prediction Problem," *IEEE Trans. Cybern.*, pp.1–12, 2019.
- [18] P.M. Chuan, L.H. Son, M. Ali, T.D. Khang, L.T. Huong, and N. Dey, "Link prediction in co-authorship networks based on hybrid content similarity metric," *Applied Intelligence*, vol.48, no.8, pp.2470–2486, 2018.
- [19] D.Q. Vu, A.U. Asuncion, D.R. Hunter, and P. Smyth, "Dynamic ego-centric models for citation networks," *International Conference on Machine Learning*, pp.857–864, 2011.
- [20] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol.342, no.6154, pp.127–132, 2013.
- [21] H.W. Shen, D. Wang, A.-L. Barabási, and C. Song, "Modeling and predicting popularity dynamics via reinforced Poisson processes," *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp.291–297, 2014.
- [22] P. Bao, "Modeling and Predicting Popularity Dynamics via an Influence-based Self-Excited Hawkes Process," *ACM International Conference on Information and Knowledge Management*, pp.1897–1900, 2016.
- [23] S. Xiao, J. Yan, C. Li, and B. Jin, "On modeling and predicting individual paper citation count over time," *Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp.2676–2682, 2016.
- [24] S. Xiao, J. Yan, M. Farajtabar, L. Song, X.K. Yang, and H.Y. Zha, "Joint modeling of event sequence and time series with attentional twin recurrent neural networks," arXiv preprint arXiv:1703.08524, 2017.
- [25] H. Mei and J.M. Eisner, "The neural hawkes process: A neurally self-modulating multivariate point process," *Advances in Neural Information Processing Systems*, pp.6754–6764, 2017.
- [26] P. Embrechts, T. Liniger, and L. Lin, "Multivariate Hawkes processes: An application to financial data," *Journal of Applied Probability*, vol.48, no.A, pp.367–378, 2011.
- [27] S. Martin, S. Ralf, and N. Hermann, "LSTM neural networks for language modeling," *Thirteenth Annual Conference of the International Speech Communication Association*, pp.194–197, 2012.
- [28] S. Fricke, "Semantic scholar," *Journal of the Medical Library Association: JMLA*, vol.106, no.1, pp.145–147, 2018.
- [29] M. Yaminifirooz and F.R. Ardali, "Identifying the Factors Affecting Papers' Citability in the Field of Medicine: an Evidence-based Approach Using 200 Highly and Lowly-cited Papers," *Acta Inform Med*, vol.26, no.1, pp.10–14, 2018.
- [30] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee, "Towards a Stratified Learning Approach to Predict Future Citation Counts," *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp.351–360, 2014.
- [31] B.H. Hall, A.B. Jaffe, and M. Trajtenberg, "The NBER patent citation data file: Lessons, insights and methodological tools," *National Bureau of Economic Research (No. w8498)*, 2001.
- [32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol.9, no.8, pp.1735–1780, 1997.
- [33] N. Tax, I. Verenich, M. La Rosa, and M. Dumas, "Predictive Business Process Monitoring with LSTM Neural Networks," *International Conference on Advanced Information Systems Engineering*, Springer, Cham, pp.477–492, 2017.

Lisha Liu received a Bachelor's degree in Computer Science and Technology from School of Business, Anhui University of Technology (now called Maanshan College) in 2018. She is now studying computer science at Hangzhou Dianzi University in China and Yamanashi University in Japan for dual master degrees. Her current research interests are big data, machine learning, and natural language processing. Her e-mail is liulisha250@gmail.com.

Dongjin Yu is a professor at Hangzhou Dianzi University, China. His current research focuses on big data, service computing, and intelligent software engineering. He is the director of Institute of Big Data (IBD) and Institute of Computer Software (ICS) of Hangzhou Dianzi University. He is a member of ACM and IEEE, and a senior member of China Computer Federation (CCF). He is also a member of Technical Committee of Software Engineering CCF (TCSE CCF) and a member of Technical Committee of Service Computing CCF (TCSC CCF).

Dongjing Wang is currently a lecturer with Hangzhou Dianzi University, China. He received the B.S. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 2012 and 2018, respectively. He was co-trained at the University of Technology Sydney for one year. His current research interests include recommender systems, machine learning, and business process management. His e-mail is dongjing.wang@hdu.edu.cn.

Fumiyo Fukumoto received BSc in the Faculty of Sciences, Gakushuin University, 1986. Fumiyo Fukumoto received MSc in the centre for Computational Linguistics of UMIST, UK in 1993, and a Ph.D. degree in the Department of Information Science, University of Tokyo, Japan in 1997. She was a Research Assistant, University of Yamanashi until 1999, Associate Professor until 2010 and appointed as a Professor of Interdisciplinary Graduate School of Medicine and Engineering, Univ. of Yamanashi.

Her current research interest is Computational Linguistics in particular lexical semantics. She is a member of the Association for Computational Linguistics, Association for Natural Language Processing, and the Information Processing Society of Japan.