Information Processing and Management 58 (2021) 102594

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Detecting research topic trends by author-defined keyword frequency

Wei Lu [a,b], Shengzhi Huang [a,b], Jinqing Yang [a,b], Yi Bu [c], Qikai Cheng [a,b], Yong Huang [a,b,*]

[a] School of Information Management, Wuhan University, Wuhan, Hubei, China
[b] Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan, Hubei, China
[c] Department of Information Management, Peking University, Beijing, China

ARTICLE INFO

Keywords:
Scientometrics
Bibliometrics
Deep learning
Word frequency prediction

ABSTRACT

Detecting research trends helps researchers and decision makers to promptly identify and analyze research topics. However, due to citation and publication delay, previous studies on trend analysis are more likely to identify *ex-post* trends. In this study, we employ author-defined keywords to represent topics and propose a simple, effective, and *ex-ante* approach, called author-defined keyword frequency prediction (AKFP), to detect research trends. More specifically, the proposed AKFP relies on the long short-term memory (LSTM) neural network. Four categories of features are proposed as input variables: Temporal feature, Persistence, Community size, and Community development potential. To verify the effectiveness and feasibility of the AKFP, we also proposed a simple but effective method to build a balanced and sufficient data set and conducted extensive comparative experiments, based on data extracted from the ACM Digital Library. Our empirical result confirms the feasibility of word frequency prediction by forecasting precision. Specifically, the short- and medium-term word frequency prediction achieved excellent performance, and the long-term word frequency prediction obtained acceptable prediction accuracy. In addition, we found that these proposed features have a significant but inconsistent impact on the AKFP. Specifically, the temporal feature is always an unignorable factor. The persistence has a strong correlation with the community size, and both are more important in the short- and medium-term prediction. In contrast, the community development potential is particularly significant in the long-term prediction.

1. Introduction

Continuous growth of scientific publications makes it more challenging for decision makers and researchers to follow frontiers and trends in a timely and accurate manner (Huang & Zhao, 2019; Katsurai & Ono, 2019; Peset et al., 2020). Detecting research topic trends in advance and continuously tracking them plays a vital role in research and development (R&D), which not only provides support for policy-making and funding allocations, but also enables researchers to gain a deeper understanding of the evolution of disciplines (Behrouzi, Zahra Shafaeipour, Hajsadeghi & Kavousi, 2020; Chang, Huang & Lin, 2015; Duvvuru, Radhakrishnan, More,

* Corresponding author.
E-mail addresses: weilu@whu.edu.cn (W. Lu), ShengzhiHuang@whu.edu.cn (S. Huang), Jinq_yang@163.com (J. Yang), buyi@pku.edu.cn (Y. Bu), chengqikai0806@163.com (Q. Cheng), yonghuang1991@whu.edu.cn (Y. Huang).

https://doi.org/10.1016/j.ipm.2021.102594
Received 31 December 2020; Received in revised form 10 March 2021; Accepted 12 March 2021
Available online 26 March 2021

Kamarthi & Sultornsanee, 2013; Jia, Wang & Szymanski, 2017; Li, Ding, Feng, Wang & Ho, 2009; Santa Soriano, Álvarez & Valdés, 2018; Wang, 2018; Xu, Hao, An, Yang & Wang, 2019; Zeng, Shen, Zhou, Fan & Havlin, 2019).

The purpose of trend analysis is twofold: one is to map the intellectual structure of the discipline, which helps researchers to understand the cognitive structure and dynamics of research trends (Duvvuru et al., 2013; Li et al., 2009), and the other is to discover new topics, especially emerging topics in science (Wang, Cheng & Lu, 2014). Many studies regard keywords as the core element of expressing topics (Asghari, Sierra-Sosa & Elmaghraby, 2020; Chang et al., 2015; Duvvuru et al., 2013; Huang & Zhao, 2019; Katsurai & Ono, 2019; Khasseh, Akbar, Afshin, Moghaddam & Sharif, 2017; Liu, Hu & Wang, 2012; Peset et al., 2020; Trevisani & Tuzzi, 2018). In particular, the author-defined keyword (AK), which is a type of keyword hand-picked by the writer, contains topics that the author considers to be the most relevant to their research (Huang & Zhao, 2019; Lu et al., 2020; Zhao, Mao & Lu, 2018). Word frequency has long been utilized as the primary indicator of a topic's vitality, and high-frequency keywords are deemed to reveal the 'hot' topics (Huang & Zhao, 2019; Khasseh et al., 2017; Zhao et al., 2018). In essence, the temporal evolution of word frequency mirrors the historical development of the corresponding topics (Trevisani & Tuzzi, 2018).

The most commonly used techniques of trend analysis are citation-based analysis and keyword-based analysis (Wang et al., 2014). Citation-based methods including direct citation, bibliographical coupling and co-citation can enhance the understanding of the structure and behavior of the discipline from a collection of articles (Chang et al., 2015; Chen & Redner, 2010; Duvvuru et al., 2013; Mccain, 2014). However, due to the time lag between the publication and the citation, and because not all citations from an article are created equally (Zhu, Turney, Lemire & Vellino, 2015), the previous research has found it difficult to analyze research trends in time (Lee, Kwon, Kim & Daeil, 2018; Xu et al., 2019). Unlike citation analysis, keyword-based methods focus on AKs and/or keywords generated by articles, which can be analyzed immediately after the publication of the article. Keyword-based trend analysis of research topics may be either popularity-based or network-based (Choi, Yi & Lee, 2011). Network-based methods, such as keyword co-occurrence network analysis or keyword-citation-keyword network, have been proven to be effective in identifying the research trends and detecting hotspots in research (An & Wu, 2011; Chang et al., 2015; Cheng, Wang, Lu, Huang & Bu, 2020; Choi et al., 2011; Dehdarirad, Villarroya & Barrios, 2014; Duvvuru et al., 2013; Liu et al., 2012). However, it generally ignores the inherent life cycle of keywords and is sometimes restricted to simple descriptions of the network (Huang & Zhao, 2019). Popularity-based methods focus on analyzing the keyword frequency, which is recently regarded as a primary metric in signaling research trends (Huang & Zhao, 2019; Li et al., 2009; Peset et al., 2020; Trevisani & Tuzzi, 2018). However, to the best of our knowledge, word frequency has been more often analyzed in the retrospective analysis of research trends, and the quantitative analysis of the evolution of keyword frequency in the future is still a blank spot.

In this study, we utilized AKs to represent the research topics and aimed to detect research trends ahead of time by predicting AK frequency. Therefore, we proposed the author-defined keyword frequency prediction task (AKFP), which is essentially a regression task for fitting the life cycle of keywords in the specific field. More specifically, to approximate any keyword count trajectories and determine unified parameters to keep a close relationship between keywords in a specific field, the AKFP is fulfilled based on the long short-term memory (LSTM) neural network. Four categories of features (Temporal feature, Persistence, Community size and Community influence potential) for measuring the novelty, the current popularity, the human resources, and the potential development power of the AK, respectively, are proposed as the input variables of the AKFP. The keyword frequency in the following years, which is a proxy for measuring the popularity of topics, was used as output variables. To verify the effectiveness and feasibility of the AKFP, we also proposed an effective method to build balanced and sufficient data set based on articles extracted from the ACM Digital Library and conducted extensive comparative experiments. In addition, a cross-validation method and the "leave-one-out model" were employed to reveal the importance ranking of these features employed in this study.

The current study has the following theoretical and practical implications. Different from previous studies, we focused on predictive analysis of research trends rather than retrospective analysis of research trends. We proposed four categories of features which has a significant but inconsistent impact on future keyword frequency. The importance ranking of these features helps researchers to gain a deeper understanding of dynamics of research trends and provides guidance for trend detection. The feasibility of word frequency prediction in short-, medium- and long-term has been verified by the AKFP based on the LSTM neural network. Therefore, the AKFP can be used not only to detect trends of new topics and identify emerging topics, but also to reveal obsolete topics and outdated technology in advance, which can provide support for policy-making and/ or an early warning for decision makers to avoid unnecessary economic losses. In addition, the proposed method of training set construction can also be utilized in other prediction tasks encountering uneven data distribution such as citation count prediction for building a balanced and sufficient training set. At the end of this paper, we also offer the practical guidelines and potential application scenarios of the AKFP.

The rest of this paper is organized as follows. Section 2 reviews recent studies on keyword-based trend research and machine learning approaches for word frequency and citation count prediction. Section 3 presents the objectives of the current study. Section 4 clarifies the definition and implementation method of the AKFP, and proposes the features employed in this study. Section 5 entails the preparation of data, experimental setup and analysis of empirical results. Section 6 discusses the contributions and limitations of this research.

2. Related work

2.1. Keyword-based trend analysis of research topics

The research topics can be regarded as a group of coherent research problems, concepts, and methods related to the discipline of interest to researchers (Braam, Moed & Van Raan, 1991). Utilizing keywords to represent topics and the core ideas of articles are

proven to be effective and feasible, and trend analysis based on keywords also achieves a good performance (Cheng et al., 2020). Keyword-based trend analysis of research topics may be either popularity-based or network-based (Choi et al., 2011).

A keyword network vividly depicts the relationships of keywords and the centrality of keywords, which maps the knowledge structure in a series of articles (Hu & Zhang, 2015; Huang & Zhao, 2019; Katsurai & Ono, 2019). Hence, keyword networks have been widely employed to provide insight into the topic evolution in a field. For example, Choi et al. (2011) constructed a keyword network to analyze how keywords are associated with each other and revealed the knowledge evolution in the MIS field. Liu et al. (2012) utilized keywords to present the research topics and employed co-word analysis to highlight the research advances of the digital library (DL) field in China. Duvvuru et al. (2013) argued that keyword networks formed from keyword frequency of use are an effective tool for comprehending research trends, and analyzed the difference between structured keyword networks and unstructured keyword networks based on keywords from two prominent business management journals from the USA and India. Their results indicated that structured keyword networks are better than unstructured keyword systems to reveal research trends and highlight the emerging areas. Dehdarirad et al. (2014) utilized co-word analysis and hierarchical cluster analysis to cluster keywords based on 652 articles and reviews extracted from WOS, by which they identified the evolution and current status of the literature on gender differences in science. An and Wu (2011), Hu and Zhang (2015) and Khasseh et al. (2017) also employed co-word analysis and cluster analysis to analyze the research patterns and evolutionary trends of stem cell field during the period of 2001–2010, Recommendation System in China during the period of 2004 −2013 and iMetrics during the period of 1978 to 2014, respectively. Chang et al. (2015) combined keyword, bibliographic coupling, and co-citation analysis to analyze the research trend in library and information science (LIS). They revealed that “Bibliometrics” became predominant and “information seeking (IS) and information retrieval (IR)” showed a decreasing trend between 1995 and 2014. However, although the keyword network helps researchers to yield fruitful results, it generally ignores the inherent life cycle of keywords. Some network analyses were restricted to simple descriptions of the network, and quantitative studies were seldomly carried out to analyze the trends (Huang & Zhao, 2019).

When discussing the popularity-based method, Trevisani and Tuzzi (2018) proposed the "life cycle" of words by clustering words that have similar normalized keyword count trajectories, and traced a possible evolution of statistics. However, their research was a retrospective study and was unable to detect trends in the future. Peset et al. (2020) argued that the appearance and disappearance of keywords provided insight into some relevant aspects of the evolution of the LIS area. They quantified the probabilities of the new author keywords surviving for 10 years as a function of the impact of the journals. The purpose of their research was similar to ours, but they focus more on the probabilities of the survival of keywords, rather than on the specific keyword count in the future. Li et al. (2009) provide insights into the trend of stem cell research based on exponential fitting of the trend of publication outputs during 1991–2006, distribution of source title, author keyword, and keyword plus analysis. They predicted that the number of publications related to the term stem cell in 2011 would double that of 2006, and revealed that “embryonic stem cell” and “mesenchymal stem cell” are the main direction of stem cell research in the 21st century. Although their research successfully predicted the number of publications in the specific field, they lacked quantitative prediction for the development of topics in the field. Huang and Zhao (2019) proposed a novel indicator called PAFit to measure keyword popularity, which achieved an outstanding prediction performance on the growth of word frequency and word degree. Their study revealed that the popularity of ecological topics obeys the “rich get richer” and “fit get richer” mechanism. Their research was most similar to ours, as it predicted keyword frequency in the following three years by simple linear regression. However, simple linear regression does not meet the assumption that the life cycle is plotted as an S-shaped curve, and the prediction of keyword frequency over a long time span needs to be further explored.

Zhao et al. (2018) examined the relationships among word frequency and network-based metrics on co-word networks. They found that the strong correlations between word frequency and network-based metrics, which confirm frequency as a simple but effective method to detect research trends in a field. Hence, the purpose of this study is to explore comprehensively the feasibility of word frequency prediction and propose a quantitative method to detect research trends by forecasting word frequency. The proposed AKFP aims to achieve a real sense of trend prediction rather than identifying no more than current hotspots.

2.2. Machine learning approaches for forecasting word and citation frequency

Machine learning approaches have attracted much attention and achieved fruitful results in the area of Scientometrics. These common machine learning approaches have been widely used in prediction tasks such as word frequency prediction, word network links predictions. Huang and Zhao (2019) used their proposed popularity metrics from the past 27 years to predict the growth of word frequency and word degree in the following three years, based on a simple linear regression model. Their goodness of fit achieved a good performance in short-term prediction. However, the prediction of keyword frequency in a long time span is still a blank spot and needs to be further explored. Behrouzi et al. (2020) utilized five different supervised machine learning algorithms and three different topology-based prediction methods for link prediction to reveal the future structure of the keyword networks, and provided insight into the future trends of the computer science field. Their study focused more explicitly on the growth of the number of links of the whole keyword network rather than on the temporal evolution of a single keyword’s frequency. In addition, with network topology-based metrics and their temporal evolutionary information as input variables, Choudhury and Uddin (2016) also employed supervised learning approaches for link prediction in co-word networks. Both citation count prediction and word frequency prediction are fitting regression tasks and a high degree of similarity exists between them. Machine learning approaches have also achieved an excellent performance on impact prediction (e.g. citation count prediction). Abramo, D’Angelo and Felici (2019), Geng et al., 2018, Chakraborty, Kumar, Goyal, Ganguly and Mukherjee (2014) and Yan, Huang, Tang, Zhang and Li (2012), Robson and Mousques (2014) utilized linear regression model (LR), eXtreme Gradient Boosting (XGBoost), random forest (RF), support vector machine (SVM) and k-nearest neighbor (KNN) to predict citation counts, respectively, and achieved ideal results. In this study, four common

machine learning approaches (i.e. LR, KNN, XGBoost, and RF) were tried to achieve better performance on the AKFP.

The neural network was proposed by Rumelhart, Hinton and Williams (1986), and is currently one of the most popular machine learning algorithms in prediction tasks. Neural networks do not require a strict assumption of data distribution and possess a large number of adjustable parameters, so they have sufficient capacity to model complicated tasks and their performance is generally better than that of the common machine learning algorithms (Guo et al., 2020). At present the application of neural network algorithms in word frequency prediction is relatively scarce, but is more commonly used in impact prediction (e.g. citation count prediction). Lee et al. (2018) used a multi-layer feedforward neural network to predict patent citation count. Unlike the outcomes of previous studies, which are more likely to present current key technologies, their study can identify highly cited patents in the early stage of patent publication. Ruan, Zhu, Li and Cheng (2020) also used a four-layer feedforward neural network to predict citation counts, and their fitting results are better than those of common machine learning algorithms. However, compared with feedforward neural networks, which are unable to effectively fit sequence nature, the recurrent neural network (Elman, 1990) and long short-term memory neural network (Hochreiter & Schmidhuber, 1997) process time series data in their inherent order, so the input sequence is considered (Alkhodair, Ding, Fung & Liu, 2020; Ketkar & Santana, 2017). Abrishami and Aliakbary (2019) proposed a sequence-to-sequence method for predicting long-term citations of a paper based on the short-term citation counts, and their prediction accuracy outperforms state-of-the-art methods. In this study, to effectively grasp sequence nature and achieve better performance, the AKFP was also implemented based on the LSTM neural network.

2.3. Research objective

Keyword-based trend analysis of research topics has been proven to be effective in mapping the intellectual structure of the discipline, detecting hotspots and discovering new topics. However, the existing literatures generally suffer from the following limitations. First, they sometimes were confined to the simple descriptions of word network structure (Huang & Zhao, 2019). Second, word frequency has been more often analyzed in the retrospective analysis of research trends rather than predictive analysis of research trends. Third, the inherent life cycle of keywords, as depicted by Trevisani and Tuzzi (2018), was generally ignored, and few researches quantitatively explored the feasibility of word frequency prediction.

Therefore, the main objective of this work is to propose a quantitative method (AKFP) for detecting research trends by forecasting keyword frequency, that addresses the above-mentioned shortfalls. More specifically, we should achieve the following two sub goals. First, we need to explore what factors affect the future word frequency. Therefore, this study proposed four categories of features (i.e. Temporal feature, Persistence, Community size, and Community development potential). These features are proved to have a significant but inconsistent impact on future word frequency. Second, we should verify the feasibility of the word frequency prediction. Therefore, we built a balanced and sufficient data set by the data extracted from the ACM Digital Library during the period of 1969 – 2018. We then fulfilled AKFP based on four common machine learning algorithms and LSTM neural network to verify the feasibility of word frequency prediction in short-, medium-, long-term. Our empirical results also showed that the AKFP achieves satisfactory effect in the computer science field.

3. Method

3.1. Problem definition

The author-defined keyword frequency prediction task (AKFP) aims to fit and then extrapolate the developing pattern of the topics based on the historical data of the author-defined keywords (AKs). More specifically, the AKFP takes the features of consecutive m years of the AK, $X(x_1,x_2,\cdots,x_m)$, as the input variable, and the AK frequency in the following n-th year, $Y(y_n)$, as the output variable to find the complex functional relationship $Y = f(X)$. Hence, the AKFP is fundamentally a fitting regression task. In the following paper, we call m and n the time window and the time span of the AKFP, respectively. Briefly, the goal of the AKFP is to quantitatively predict the frequency of use in any life-cycle stage of the AK and reveal the developing pattern of topics in a specific field.

3.2. Prediction model

As knowledge carriers of research topics and technologies, the keywords should obey the life-cycle theory. The technology evolution over time is generally plotted as an S-shaped curve to represent its life cycle (Ernst, 1997; Rezaeian, Montazeri & Loonen, 2017; Taylor & Taylor, 2012). However, due to the technology renaissance, a few technologies and their representative keywords, such as "deep learning", may experience multiple stages of ups and downs, which means many growth and diffusion processes consist of several sub-processes. In other words, the life-cycle curve may be more likely the composition of multiple S-shaped curves (Rezaeian et al., 2017). In addition, the frequency of some commonly used keywords continues to increase in accordance with the number of publications. Therefore, the common S-curve is not applicable to all keywords and ignores the connection between keywords in the same field. Unlike S-shaped curves, the neural network algorithm does not require strict assumption of data distribution and a fully connected feed-forward neural network can approximate any continuous function at any desired level of precision (Hornik, Stinchcombe & White, 1989). Consequently, the neural network algorithm is suitable for fitting the developing pattern of keywords and determining unified parameters to keep a close relationship between keywords in a specific field.

Recurrent neural network (RNN) (Elman, 1990) and long short-term memory neural network (LSTM) (Hochreiter & Schmidhuber, 1997) effectively fit the sequence property of time series data. The LSTM, as an improvement of the RNN, avoids the phenomena of

gradient vanishing and gradient explosion (Alkhodair et al., 2020). Therefore, this study selects the LSTM as the prediction model. As shown in Fig. 1, the neural network framework employed in this study consists of input layer, hidden layer, and output layer. The whole neural network is composed of multi-layer computing units connected in turn, and the data flows through the network by matrix operations. The *i*-th feature is a time series data with length of *m*, $(x_{1i}, x_{2i}, \ldots, x_{mi})$. All proposed features are concatenated as the input variables of the model, $X(x_1, \ldots, x_j(x_{j1}, x_{j2}, \ldots, x_{jk}), \ldots, x_m)$. As shown in Eq. (1), in the input layer, a single-layer feedforward neural network (FNN) is employed to transform the initial low-dimensional feature representations into the high-dimensional feature space. The activation function σ converts the input into a nonlinear output, which enhances generalization ability of the model.

$$x_j = \sigma\left(W_{input} x_j + b_{input}\right) \tag{1}$$

Then, the multi-layer LSTM is utilized to deal with time series data as shown in Eqs. (2)-6.

$$i_j = \sigma_{sig}\left(W_i x_j + U_i h_{j-1} + V_i c_{j-1} + b_i\right) \tag{2}$$

$$f_j = \sigma_{sig}\left(W_f x_j + U_f h_{j-1} + V_f c_{j-1} + b_f\right) \tag{3}$$

$$c_j = f_j c_{j-1} + i_j tanh\left(W_c x_j + U_c h_{j-1} + b_c\right) \tag{4}$$

$$o_j = \sigma_{sig}\left(W_o x_j + U_o h_{j-1} + V_o c_j + b_o\right) \tag{5}$$

$$h_j = o_j tanh\left(c_j\right) \tag{6}$$

where σ_{sig} is the logistic sigmoid function, and *i*, *f*, and *o* are the input, forget, output gates, respectively. h_j,c_j represents the short-term memory and long-term memory of the LSTM in time *j*, respectively. Finally, the output of the hidden layer is taken as the input of the output layer, which still employs a single-layer FNN to transform the output into the dimension we need, as shown in Eq. (7).

$$\hat{y} = \sigma\left(W_{output} c_m + b_{output}\right) \tag{7}$$

In this study, prediction target (the AK frequency in the following *n*-th year) is a scalar, so the output layer exports a scalar. Finally, we adopted the mean square error as the loss function (Eq. (8)).

$$loss = \frac{1}{N}\sum_{l=1}^{N}(y_l - \hat{y}_l) \tag{8}$$

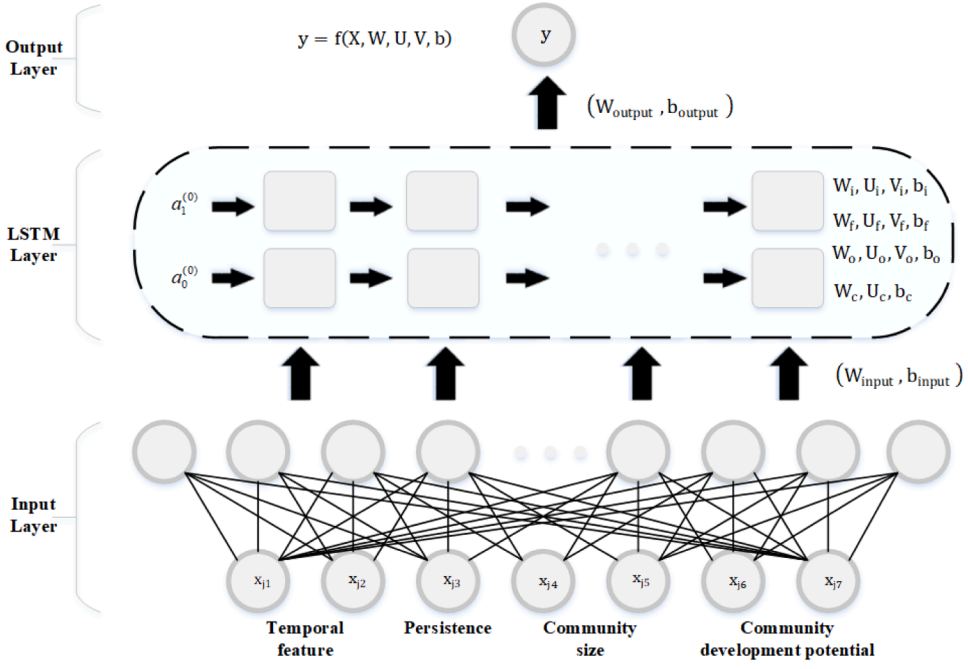

Fig. 1. Author-defined keyword frequency prediction model.

where y_i represents the actual keyword frequency, $\widehat{y_i}$ indicates the predicted keyword frequency, and N denotes the training sample size.

The operation of neural networks generally involves two steps: training and testing. In the training step, a back-propagation algorithm is utilized to modify the connection weights and bias until the algorithm converges. In the testing step, the performance of the trained neural network is evaluated on a test set. Once the model generalizes to an acceptable deviation level, it is ready to be applied to new data.

3.3. Keyword features

The following four categories of features are proposed in this study: (1) Temporal feature, (2) Persistence, (3) Community size, and (4) Community development potential. A total of seven sub-indicators are employed as input variables for the AKFP.

3.4. Temporal feature

Temporal feature includes two types of sub-indicators to represent the novelty of AKs. Before discussing the temporal feature, we introduce the concept of potential development year (PDY) proposed by Tu and Seng (2012). They defined the PDY as the period from the first year to the current year when a topic becomes a research topic that does not include any year with zero papers in the following years. They utilized the novelty index (NI) defined as the inverse of PDY to indicate whether a topic is novel. In this paper, we simplify the definition of PDY as the period from the first year when a keyword is published to the current year, and also use NI to indicate whether a keyword is novel. The formula for NI is as follows:

$$t_{relative} = \frac{1}{t - t_0 + 1} \tag{9}$$

In Eq. (9), t_0 indicates the time when the keyword was first selected as an AK, t means the current year. When an AK is first published, NI is normalized to 1. In its second year, the NI should be 1/2. For convenience, we rename NI as relative time ($t_{relative}$). In recent years, the rapid development of science and technology has led to the explosive growth of academic literatures, and the keyword frequency has also increased rapidly. This means the keyword frequency should be subject to the influence of current time. Here we define the current year, t, as the second sub-indicator, $t_{absolute}$, which denotes the absolute time of the AK. In the empirical experiments, $t_{absolute}$ is the difference relative to a fixed time point.

3.5. Persistence

The keyword frequency is a quantitative metric to measure the popularity of a keyword. In this research, the keyword frequency at time t is taken as one of the features, which is denoted as n_t. The topic detection and continuous tracking help to identify emerging topics (Suominen & Newman, 2017). Therefore, the time window, m, of the AKFP is set to three for persistent tracking, which means the keyword frequency of three consecutive years (n_t, n_{t+1}, n_{t+2}) is taken as input variables. The time series data provides more abundant information than that of a single year. For example, the time series data of keyword frequency contains the growth between two years ($n_{t+1} - n_t$), which also reflects the growth rate of keyword frequency. It is worth noting that we can set the keyword frequency of the first two years as 0 to predict the keyword frequency for the AK published for the first time. This trick allows us to make an evaluation and prediction for the new AK in time.

3.6. Community size

Community is key in the process of topic evolution. The names of people and/or organizations must be folded into the equation to determine community (Suominen & Newman, 2017). We argue that the size of the community also affects the growth of keyword frequency and regard the size of the community as a feature of AKs, called community size. The AK adopted by a large number of scholars and institutions essentially has sufficient human and material resources behind it, and its representative technologies may achieve technical breakthroughs rapidly. In this study, we employ the number of authors who select the AK at time t (a_t) and the number of institutions which select the AK at time t (i_t) as two sub-indicators to depict the size of the community. Similarly, the numbers of authors and institutions in three consecutive years (a_t, a_{t+1}, a_{t+2} and i_t, i_{t+1}, i_{t+2}) are used as input variables of the AKFP.

3.7. Community development potential

Community development potential is closely related to community size, and two sub-indicators are also constructed from the perspective of authors and institutions. Hu, Tai, Liu, Cai and Egghe (2020) utilized the cumulative number of papers published by authors as an author-based feature to identify highly cited articles. This study inherits the spirit of this work and measures the current development potential and contribution of the AK using the number of accumulative publications of authors and institutions. Specifically, we use pa_t and pi_t to denote the cumulative number of papers published by all authors who selected the AK in time t and the cumulative number of papers published by all institutions which adopted the AK in time t, respectively. These metrics, pa_t and pi_t, can be mathematically formulated as Eqs. (10)-(11):

$$pa_t = \sum_{\alpha \in A_t} p_t^{\alpha} \tag{10}$$

$$pi_t = \sum_{\beta \in B_t} p_t^{\beta} \tag{11}$$

where t denotes the absolute time ($t_{absolute}$), A_t indicates the set of authors who select this AK at time t, and p_t^{α} is defined as the cumulative number of papers published by the author, α, up to time t. pi_t has a similar meaning to pa_t, but it is formulated from institutions. B_t indicates the set of institutions which use the AK at time t, and $p_t^{\beta}$is defined as the cumulative number of papers published by the institution β up to time t. The larger the p_t^{α}, the stronger the academic ability of the author, α, and the AK adopted by α is more likely to be further studied and promoted. Similarly, institutions with a large number of publications are more capable of leading the direction of science and technology development. Hence, the larger the p_t^{β}, the broader the research prospects of the AK adopted by β.

However, the accumulative number of publications published by the author and institution increases over time. For example, the cumulative number of papers published by an author in 2000 is at least equal to that before 2000. That is to say, pa_t and pi_t tend to give the recent AKs a higher score, which leads to inequality in time. In order to eliminate the time factor, we adopted the Z-score method to standardize the pa_t and pi_t of all the AKs within one year, which gives the comparison of pa_t and pi_t in different years a relatively fair starting point.

4. Experiments

4.1. Data

The ACM Digital Library is the world's most comprehensive database in computer science field. We collected literatures in the ACM Digital Library from 1969 to 2018 comprising 201,394 articles. After data pre-processing, there are in total 231,384 AKs, 265,371 authors and 7605 institutions. The abbreviation database was automatically built based on regular expression match. The database comprises 3247 key value pairs, in which the key is the abbreviation of the AK and the value is the full name of the AK. The distribution of publications, keywords, authors, and institutions in the ACM data set is shown in Fig. 2 (a-d).

Table 1

Before implementing the AKFP, the AK need to be standardized because it can be expressed differently for the same meaning. For

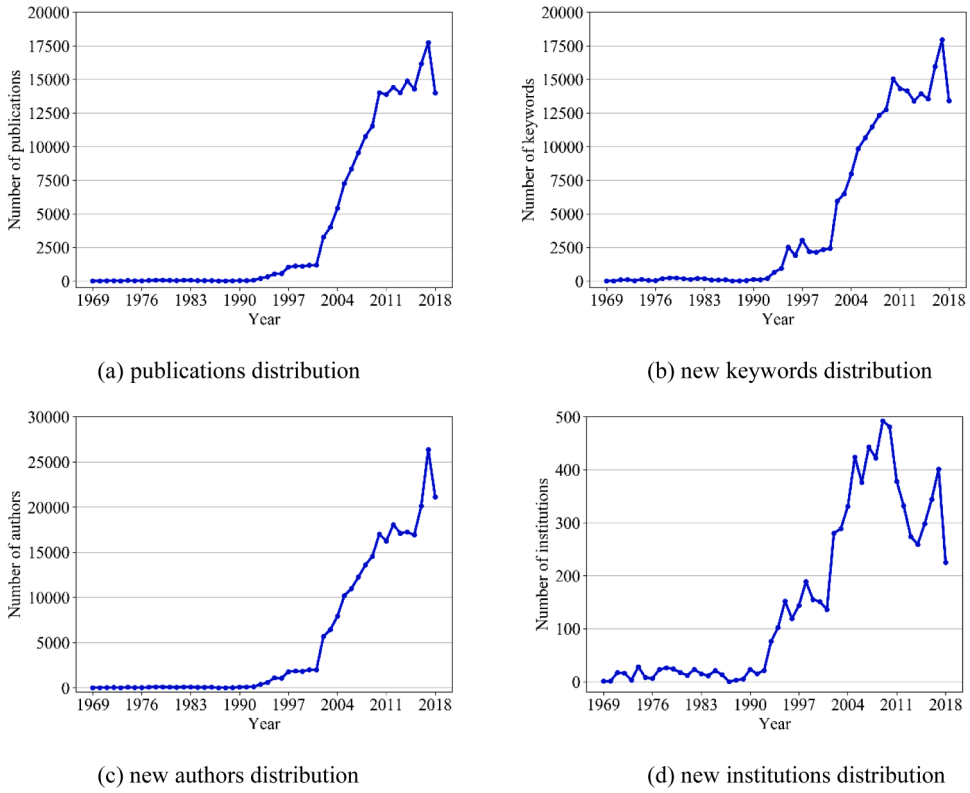

Fig. 2. Statistical distribution of publications, keywords, authors, and institutions.

example, the singular/plural form of the AK may be abused (e.g. social network/social networks); there are hyphens and other symbols in the AK (e.g. e-commerce/e commerce); there are both the original word and abbreviated form in the AK list (e.g. support vector machine/svm). Hence, considerable time and effort have to be invested in editing the AK. In order to alleviate the problem of irregular usage of the AK as far as possible, this study obeys the rules proposed by Choi and Hwang (2014) to refine the AK as per the following steps. First, as keywords are not case-sensitive, all the AKs were converted to lower case. After that, we standardized the AK into its singular form, and then removed the punctuation. Subsequently, the abbreviation database (as shown in Table 2) was built by regular expression match (e.g. Bayesian network (bn)). Finally, the abbreviated form of the AK was consolidated into the original one using this database.

This abbreviation database is not reported here in its entirety owing to lack of space, but part of it is shown as Table 2. There are one-to-many mapping relationships between the abbreviated and original forms, because some keywords have the same abbreviations. But this rare case only accounts for 16.97% (551) in the database, and only 0.24% in all keywords. For simplicity, this research removed this case from the data set. Finally, we got 231,384 keywords. For each keyword, we extracted the features from the metadata of the articles. The features were then used to build the training set and test set of the AKFP in the following experiments. It is worth noting that, due to the large number of literatures collected from the ACM Digital Library, it is difficult to unify synonyms in articles. We assume that synonyms obey similar keyword count trajectories, so this phenomenon should not greatly affect the performance of the AKFP.

5. Experiment setup

5.1. Training set

To verify the feasibility of short-, medium- and long-term prediction of keyword frequency, in this study, m was set as 3, and n was set as 2, 4 and 7, respectively. For simplicity, we denote these AKFPs with different n as the AKFP ($m + n$) (i.e. AKFP 5, AKFP 7, and AKFP 10).

The above preprocessed AKs are used to construct the training set for these AKFPs. However, there are typically a small number of keywords that are used frequently and a much larger number of keywords that are utilized infrequently (Choi & Hwang, 2014; Choi et al., 2011; Hu & Zhang, 2015; Zhao et al., 2018). To build a balanced and sufficient data set for these AKFPs, we roughly divided the keywords into four levels according to the cumulative word frequency, followed by randomly sampling in each level. We then utilized the sliding window method to build the data set. Finally, the data set was split into training set, verification set and test set at a ratio of 8:1:1. These steps are detailed in (1) and (2) below.

(1) Keyword selection: Since the statistics for the AKs published after 2014 cover less than five years, we got rid of them to avoid any boundary effect. There are 168,842 keywords left. According to the word frequency distribution, we simply divided these keywords into four intervals, as shown in Table 3 below. The cumulative frequency of 157,882 keywords is less than 10 and only 962 keywords are more than 99. To build a relatively balanced data set, we randomly selected 1000 keywords from the frequency range of 0–9 as well as 10–49 and retained 2191 keywords with a frequency of more than 50. Finally, a subset of 4191 keywords was obtained.

(2) Sliding window method: To make full use of the historical information of each keyword, a sliding window with the fixed step size in m is designed to build the data set, which is similar to Xu et al. (2019). Taking "machine learning" as an example, it was first published in 1989, so we generated the training pairs as shown in Table 4. The first three columns (x_1, x_2 and x_3) are the proposed indicators for three consecutive years, which can immediately be calculated from the ACM data set. The four to six columns (y_2, y_4 and y_7) denote the actual word frequency as output variables. Taking the first row of Table 4 as an example, $feature_{1989}$ is features of the first year, $feature_{1990}$ and $feature_{1991}$ are the features of the second and third years respectively, while $frequency_{1993}$, $frequency_{1995}$ and $frequency_{1998}$ indicate the actual word frequency in the following two, four and seven years, respectively. Then we slid the window forward one year in turn, to build the remaining rows. It should be noted that, since our collected ACM data set goes up to 2018, the samples that could not be obtained due to the boundary effect were denoted as "*"

Table 1
Summary of indicators employed in this research.

Category	Sub-indicator	Indicator definition
Temporal feature	$t_{relative}$	The inverse of PDY
	$t_{absolute}$	Current time (t)
Persistence	n_t	Keyword frequency in time t
Community size	a_t	Number of authors using this keyword in time t
	i_t	Number of institutions using this keyword in time t
Community development potential	pa_t	The cumulative number of papers published by authors who use this keyword in time t
	pi_t	The cumulative number of papers published by institutions which use this keyword in time t

Note: Z-score method is adopted to standardize pa_t and pi_t of AKs within one year, respectively.

Table 2
Abbreviation database.

Abbreviation	Author-defined keyword
Abm	agent-based modeling
Bn	Bayesian network
Cbr	case-based reasoning/constant bit rate
Dct	dual clutch transmission/discrete cosine transforms
Ecd	energy conservation district
…	…

Table 3
Keyword frequency distribution.

Threshold	0–9	10–49	50–99	100-	Total
Number of keywords	157,882	8768	1229	962	168,842

Table 4
"machine learning" for $m = 3$, $n = 2$, 4 and 7.

	x_2	x_3	y_2	y_4	y_7
$feature_{1989}$	$feature_{1990}$	$feature_{1991}$	$frequency_{1993}$	$frequency_{1995}$	$frequency_{1998}$
$feature_{1990}$	$feature_{1991}$	$feature_{1992}$	$frequency_{1994}$	$frequency_{1996}$	$frequency_{1999}$
…	…	…	…	…	…
$feature_{2013}$	$feature_{2014}$	$feature_{2015}$	$frequency_{2017}$	*	*
$feature_{2014}$	$feature_{2015}$	$feature_{2016}$	$frequency_{2018}$	*	*

(star). The sliding window method built a one-to-many relationship between the keyword and its samples. We then applied the method to 4191 keywords. Finally, three sample sets for AKFP 5, AKFP 7, and AKFP 10 were obtained, as shown in Table 5.

Due to the different time spans (n) of these AKFPs, the sample set of the AKFP with larger n contains fewer samples. These samples that cannot be constructed are shown as * in Table 4. Each training sample constitutes time-series data with a dimension (time step, number of indicators), in which the time step is 3 and the number of indicators is 7.

5.2. Evaluation

Four common machine learning approaches (i.e. LR, KNN, XGBoost, and RF) which have achieved good performance on word frequency prediction and/or citation count prediction are employed to fulfill the AKFP (Abramo et al., 2019; Behrouzi et al., 2020; Geng, Jing, Jin & Luo, 2018; Huang & Zhao, 2019; Robson & Mousques, 2014; Yan et al., 2012). These machine learning methods were implemented through the algorithm library encapsulated in scikit-learn (Swami & Jain, 2013). To ensure the satisfactory performance of the baseline models, the random search method (Bergstra & Bengio, 2012) was utilized to determine the parameter values. The random search algorithm firstly samples parameters according to their given distribution. Then, by traversing all the combinations of the chosen parameters and evaluating performance in a training set based on cross-validation method, the optimal parameter values are determined. Compared with the traditional grid search algorithm, the random search algorithm greatly improves the training speed while ensuring the algorithm's performance. A detailed description of the meanings and settings of the parameters of these baseline models is provided in Table 6.

Three popular criteria are utilized to evaluate the proposed method: First, mean square error (MSE), second, mean average error (MAE), and, third, the coefficient of determination (R^2). MSE measures the variation of the predicted values to the actual values and MAE measures the average of absolute errors between predicted and actual values. R^2 measures the overall relationship between predicted values and actual values. Thus, lower values of MSE and MAE and higher values of R^2 are desirable. The MSE, MAE and R^2 are defined as Eqs. (12)-(14).

Table 5
Sample size of AKFPs.

Task	Sample size
AKFP 5	59,748
AKFP 7	51,456
AKFP 10	39,850

Table 6
Parameters in the baselines.

Approach	Parameters	Description in scikit-learn	Default value	Search range	Setting value AK FP5	AK FP7	AKFP10
RF	n_estimators	The number of trees in the forest	10	(10, 200)	170	56	150
	max_depth	The maximum depth of the tree	None	(1, 20)	9	10	10
	min_sample_split	The minimum number of samples required to split an internal node	2	(1, 20)	2	3	12
	min_samples_leaf	The minimum number of samples required to be at a leaf node	1	(1, 20)	4	12	15
XGBoost	n_estimators	The number of estimators in the models	100	(10,200)	50	40	94
	min_child_weight	The minimum sum of the instance weights needed in a child	1	(1,20)	15	19	18
	max_depth	The maximum depth of the tree	3	(1,20)	7	8	3
KNN	n_neighbors	The default number of samples to use for neighbors' queries	5	(1,20)	16	19	18
	Weights	Weight function used in prediction	Uniform	Uniform or distance	uni	uni	uni
	leaf_size	Leaf size passed to BallTree or KDTree	30	(1,50)	40	7	27
LR	/	/	/	/	/		

$$MSE = \frac{1}{N}\sum_{l=1}^{N}(y_l - \widehat{y_l}) \tag{12}$$

$$MAE = \frac{1}{N}\sum_{l=1}^{N}|y_l - \widehat{y_l}| \tag{13}$$

$$R^2 = 1 - \frac{\sum_{l=1}^{N}(y_l - \widehat{y_l})^2}{\sum_{l=1}^{N}(y_l - \overline{y_l})} \tag{14}$$

In the above three equations, y_l indicates the actual value, $\widehat{y_l}$ indicates the predicted value, $\overline{y}$ represents the average of y_l, and N denotes the number of samples.

5.3. Parameters

In the training process of a neural network, the determination of hyper parameters is more like an art than a science, and an appropriate neural network architecture can dramatically improve the performance. The activation function, optimizer, and other parameters settings are shown in Table 7.

According to the MSE on the test set, we determined the optimal number of network layers and neurons in each layer. As shown in Fig 3, the x-axis denotes the number of neurons in each layer as a triple tuple and the double hidden layer is represented by brackets (2) in the legend. The y-axis represents MSE on the test set. When the number of neurons in each layer is small, the AKFP 5 and AKFP 7 achieve better performances in the double-layer LSTM, but, with the increase of neurons, the performance of the single-layer LSTM is similar to that of the double-layer LSTM. Hence, we determined the middle layer of the AKFP 5 and AKFP 7 as a single-layer LSTM, and the neurons in each layer were set to 256, 512, and 1, respectively. Owing to the difficulty of AKFP 10, the fitting result of the single-layer LSTM is always weaker than that of the double-layer LSTM. Therefore, we chose the double-layer LSTM and the neurons in each layer were set to 256, 512, and 1, respectively, in the AKFP 10.

The activation function transforms the input into a nonlinear output, which enhances nonlinear expressiveness of the neural network. In this research, "Rectified Linear Unit" (ReLU) was adopted as the activation function and its formula is as follows: $f(x) = max(0,x)$. This function is widely used in the field of deep learning and performs well in various neural network tasks (Glorot, Bordes &

Table 7
Parameters of the neural network model.

Parameters AKFP	AKFP 5	Values AKFP 7	AKFP 10
Number of units in each layer	256, 512, and 1	256, 512, and 1	256, 512 (2), and 1
LSTM layer	One layer	One layer	Two layers
Activation function	ReLU	ReLU	ReLU
Initial learning rate	1e-1	1e-1	1e-1
Optimizer	Adadelta	Adadelta	Adadelta
Batch size	64	64	64
Epochs	300	300	300

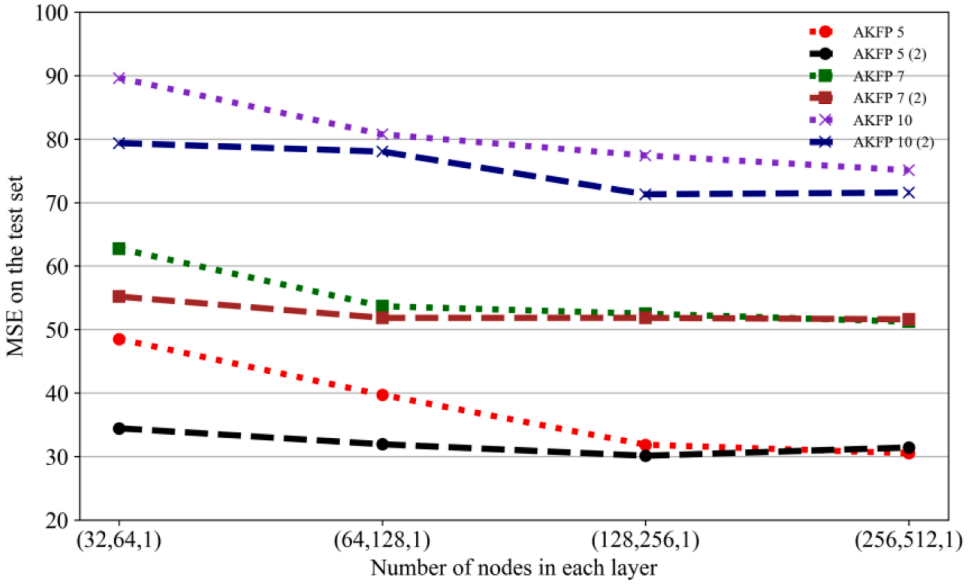

Fig. 3. MSE on test set under different neural network settings.

Bengio, 2011).

The learning rate is vital to the performance of the neural network algorithms. The high learning rate not only accelerates the training process and stops the neural network from dropping into the locally optimal solution, but also results in non-convergence of parameters. In contrast, the low learning rate helps the model fine tune around the optimal parameter values to achieve the optimal fitting results, but takes more time. To balance the training speed and performance of the neural network, an exponential decay learning rate is employed. Specifically, the initial value of the learning rate is set to 0.1, and decays exponentially with 0.9 in every 3000 batches. Because of the large initial learning rate, epochs are set to 300, so that the neural network will fine tune the parameters in the latter phases of training.

The choice of optimizer also play an important role in the performance of the neural network. Compared with the traditional gradient descent method which easily falls into the local optimal, the Adadelta optimization algorithm (Zeiler, 2012) is adopted. As an adaptive learning rate optimization algorithm, on the one hand, it allows each dimension of a parameter to have its own dynamic learning rate; on the other hand, it ensures that the units of the update match the units of the parameters. And it only uses first-order information and an approximation to the diagonal Hessian, which means it has high computing efficiency. It is worth mentioning that, although the Adadelta optimizer require no manual tuning of a learning rate, the exponential decay learning rate achieves better performance in our empirical experiments. Finally, we set the batch size to 64, and used a "tensorflow" framework to implement neural network training. The neural network settings are shown in Table 7.

5.4. Prediction results

During the training process, the parameters were updated along the negative gradient direction determined by the random gradient descent method, and the MSE on the training set gradually decreased with fluctuation. The MSE variations of different AKFPs are shown in Fig. 4, among which AKFP 5 had the fastest convergence speed and tended to converge after 200 epochs; AKFP 7 became stable after 240 epochs; and AKFP 10 converged slowly, with small amplitude oscillation at the end of training, but close to stability.

In the optimal neural network model, the MSE of AKFP 5 on training set and test set is 28.034 and 30.513 respectively; the MSE of AKFP 7 on training set and test set is 54.527 and 51.230; and the MSE of AKFP 10 on training set and test set is 86.025 and 71.567. With the expansion of the time span (n), the difficulties of the AKFP increase step by step: on the one hand, the AKFP needs a deeper neural network; on the other hand, the MSE of the optimal model rises. In order to further analyze the results, we selected 12 keywords as a case study, according to the word frequency division in Table 3, as shown in Fig. 5 (a-l).

The sliding window method proposed above was utilized to predict keyword frequency. The black curve is the actual keyword frequency curve, and the red curve nearest to it denotes predicted results of AKFP 5, closely followed by the green curve, which indicates predicted results of AKFP 7. The blue curve predicted by AKFP 10 has the lowest accuracy. These results are consistent with the MSE variation on the test set; that is, the longer the time span is, the lower the accuracy of the corresponding AKFP. Therefore, the AKFP is a prediction task that pursues the tradeoff between timeliness and accuracy. For these high-frequency keywords shown in Fig. 5 (a-i), these prediction results are satisfactory. The predicted curves keep the track similar to the actual black curve, although occasionally there is a little lag phenomenon. For these keywords with lower frequency (j-l), these prediction curves are more oscillatory due to the randomness of actual word frequency, and this phenomenon is more significant with the decrease of keyword frequency.

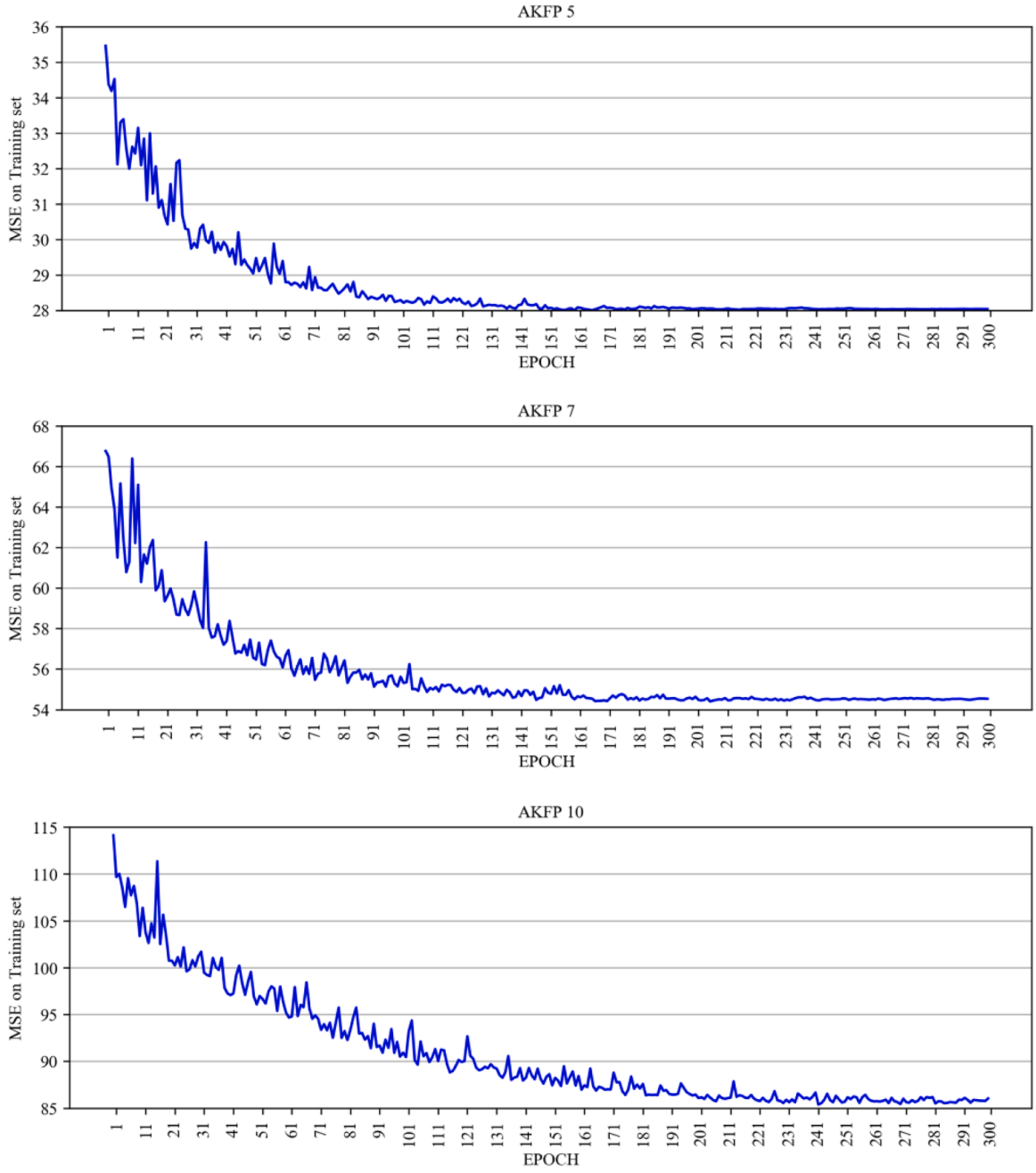

Fig. 4. MSE during the training process.

The overall performance of AKFPs on the test set is shown in Fig. 6. The horizontal axis represents the actual value, while the longitudinal axis indicates the predicted value. There are a small number of outliers which obviously deviate from the real word frequency, and this occurs more frequently in AKFP 7 and 10. These outliers are mostly below the diagonal, which indicates that the future keyword frequency is more likely to be underestimated by our model. This may be caused by the fact that the word frequency distribution obeys the power law distribution, which leads to fewer samples of high-frequency keywords in the training set. In addition, the Spearman correlation coefficients between actual and predicted word frequency are 0.907, 0.825 and 0.792 in AKFP 5, AKFP7, and AKFP10, respectively, which coincides with the changes of MSE. Overall, AKFP 5 and AKFP 7 achieved excellent performance, and AKFP 10 obtained acceptable prediction accuracy, all of which verifies the feasibility of keyword frequency prediction.

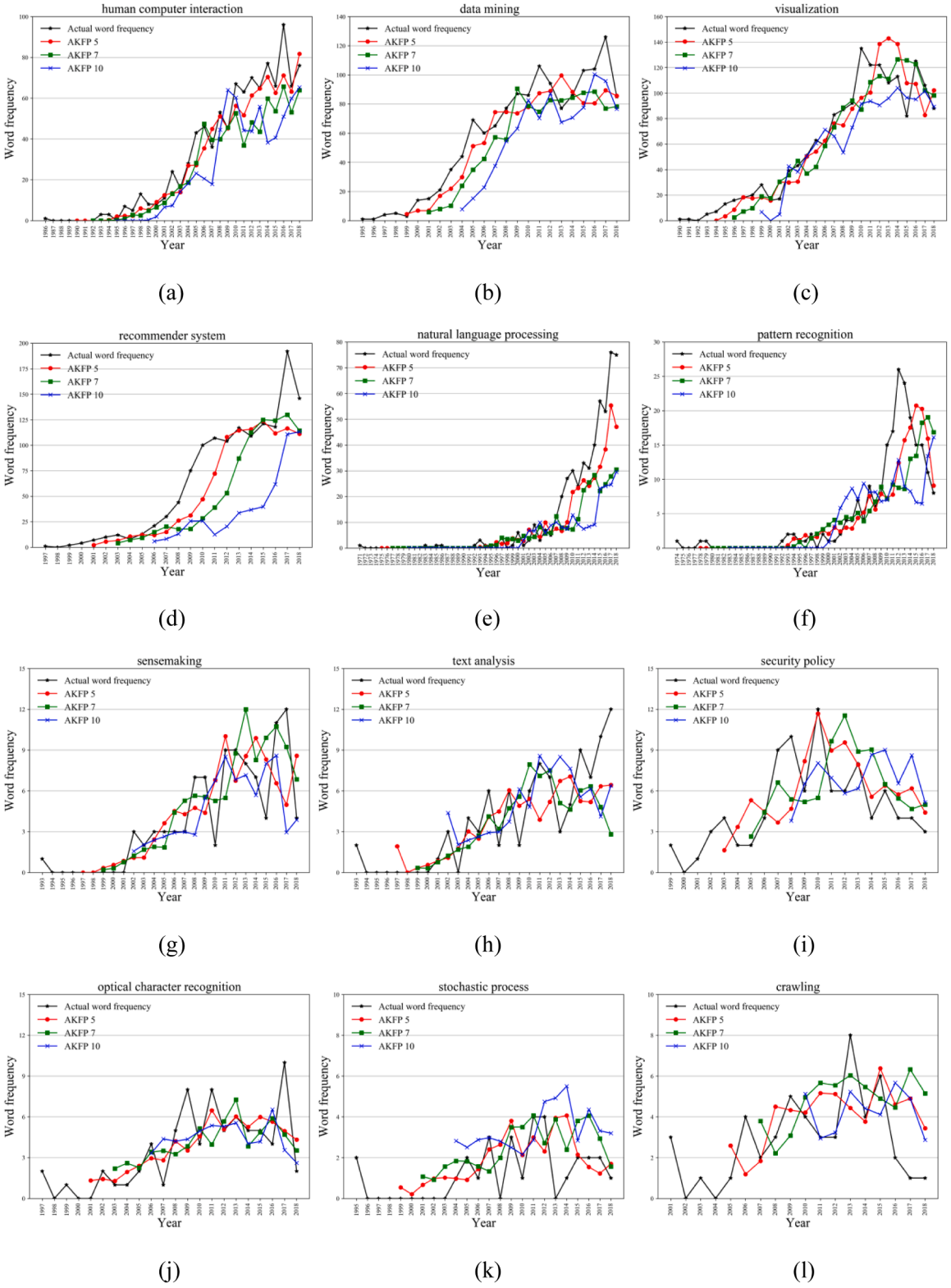

Fig. 5. Twelve cases of the AKFP.

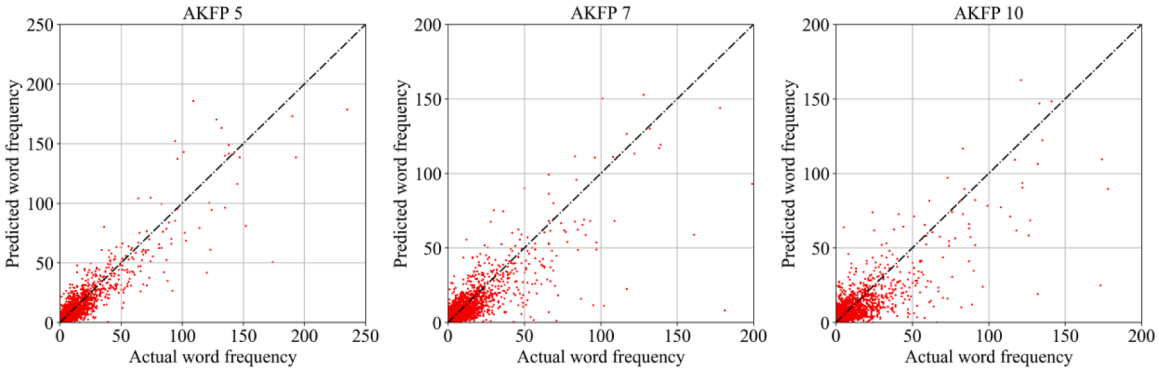

Fig. 6. Scatter of the predicted frequency and actual frequency.

5.5. Feature importance

To analyze the importance of the features employed in this study, we dropped one category of feature from input variables one by one and used the remaining features to train the neural network ("leave-one-out model"), after which the MSE on the test set was calculated. Specifically, we randomly divided the data 10 times at the ratio of 8:1:1, and 10 groups of training set, validation set and test set were obtained for AKFP 5, 7, and 10, respectively. Then, leave-one-out models and the full features model were trained on each training set. For example, AKFP 5 has 10 groups of results and each group contains a full model and four leave-one-out models. Subsequently, we calculated the difference of the MSE on the test set between the leave-one-out model and corresponding full features model, and used a paired *t*-test to test the significance level of these features on the prediction performance of the AKFP.

The experiment results are shown in Table 8 below. The third column of the table denotes the average difference of MSE on the test set. It shows that the MSE of the AKFP (5, 7, and 10) increases significantly if the temporal feature is dropped, which indicates that the temporal feature is an un-ignorable factor in the AKFP. Indeed, the explosive growth of scientific publications leads to more new keywords and more high-frequency keywords. Therefore, the characteristics of publications distribution make the temporal feature become a key factor to determine the topic trends. Although the persistence should have worked as the core feature, its information may be included by the community size, to a certain extent. In fact, there is a strong positive correlation between persistence and community size, because high-frequency keywords are generally adopted by more authors and institutions, while this is not the case for low-frequency keywords. Therefore, simply dropping one of persistence and community size may not have a significant impact on the performance of the model. However, we still identified that, with the increase in time span (n) of the AKFP, the results change from being significant for only one feature to not significant for both, which indicates that the influence of persistence and community size on future keyword frequency is gradually decreasing. In order to further analyze the effect of persistence and community size on the AKFP, the case of dropping both was trained. As expected, we found that persistence and community size significantly affect the MSE of AKFP 5, 7, and 10, but have the lowest impact (4.8506) on AKFP 10, which is consistent with the previous conclusion. Therefore, Persistence & Community size are the dominant factors in the short- and medium-term word frequency prediction. Finally, the effect of

Table 8
Difference between each 'leave-one-out' model and the full features model.

Time span	Dropped feature	Difference in test MSE
	Temporal feature	1.8384***
	Persistence	−0.2276
5	Community size	−0.8208*
	Community development potential	0.7175*
	Persistence & Community size	5.8060***
	Temporal feature	1.1853**
	Persistence	0.6374*
7	Community size	0.1541
	Community development potential	1.3966**
	Persistence & Community size	6.3894***
	Temporal feature	2.2322**
	Persistence	0.1967
10	Community size	−1.3125
	Community development potential	6.1537***
	Persistence & Community size	4.8506***

Notes:.
* indicates $p<0.05$,.
** indicates $p<0.01$,.
*** indicates $p< 0.001$.

community development influential on AKFP 5 is very weak (0.7175), but with the expansion in time span (n) of the AKFP, its impact gradually rises. Just as in AKFP 10, the effect of community development influential on MSE rises to 6.1537, which is about nine times as much as that of AKFP 5. Community development potential becomes the most important factor in long-term word frequency prediction. Hence, the scientific research capacity of authors and institutions will significantly affect the long-term development of the topics. For the sake of clarity, we ranked the importance ranking of these features in Table 10.

5.6. Comparisons

The fitting and prediction performances of the baseline models and the LSTM model are shown in Table 9. Compared with baselines, the LSTM model is slightly less effective on the training set than XGB and RF, but the fitting performance is significantly better than LR. The LSTM model exceeds all baselines on the test set. More specifically, in AKFP 5, the MSE of the LSTM on the test set is lower than 14.14% for XGB, 15.11% for RF, 24.95% for KNN, and 23.04% for LR. In AKFP 7, the MSE of the LSTM on the test set is reduced by 8.47% for XGB, 4.98% for RF, 12.41% for KNN, and 17.97% for LR. In AKFP 10, the MSE of the LSTM on the test set is decreased by 13.32% compared with XGB, 4.90% with RF, 17.58% with KNN, and 23.32% with LR. In addition, R^2 of the LSTM model achieves the best effect on the test set in AKFP 5, 7, and 10, which are 0.822, 0.679, and 0.624, respectively, illustrating that the LSTM neural network fits the developing process of keywords well and has better generalization capabilities than baseline models.

6. Discussion

In this study, the AKFP is defined as a regression problem. We built the AKFP based on the LSTM neural network and employed Temporal feature, Persistence, Community size, and Community development potential as input variables, and future keyword frequency as the output variable. In the following paper, we introduced the theoretical and practical implications, practical guidelines for the AKFP and limitations in this research.

6.1. Theoretical implications

The current paper has the following theoretical implications. First, we proposed the AKFP as a quantitative method to detect research trends in the future, which has been proved to be feasible by forecasting precision on short-, medium- and long-term prediction. More specifically, we illuminated that the AKFP is a prediction task that pursues the tradeoff between timeliness and accuracy. The current experimental results showed that short- and medium-term word frequency prediction achieved excellent performance, and long-term word frequency prediction obtained acceptable prediction accuracy. Second, the factors affecting the future word frequency have been explored. We proposed four categories of features (i.e. Temporal feature, Persistence, Community size, and Community development potential). These features are proved to have a significant but inconsistent impact on future word frequency. More specifically, we found that the temporal feature is always an unignorable factor in the AKFP. The persistence and community size, which measure the recent popularity of the topics, are the main influencing factors in short- and medium-term prediction. With the increase in time span of the AKFP, their influence gradually decreases. In contrast, community development potential is vital in long-term prediction, which becomes more and more significant with the expansion in time span of the AKFP. This also means the research ability of authors and institutions has a long-term impact on the development of the research topics. To clearly reveal the importance ranking of these features, we ranked them in different time spans of the AKFP, as shown in Table 10. Third, our experimental results also show that the neural network algorithms have better generalization ability than some common machine learning algorithms, which is consistent with previous studies (Ruan et al., 2020). Finally, in this research, we utilized the simple but effective keyword selection strategy and sliding window method to build a balanced and sufficient training set, which made full use of the data and achieved the purpose of data expansion. We argue that this technique can also be utilized in other prediction tasks encountering uneven data distribution such as citation count prediction.

6.2. Practical implications

The proposed AKFP is a simple, effective, and *ex-ante* approach to track research trends by forecasting keyword frequency. Different from previous studies, the AKFP focus on predictive analysis of research trends rather than retrospective analysis of research trends. Therefore, the AKFP can be used as a tool to assist in research such as emerging topic detection and hotspot identification, which provides support for policy-making and grant allocation. For example, through the statistical analysis of the keyword frequency, researchers can accurately grasp the future frontiers and hotspots from the micro perspective based on a single AK or from the macro perspective based on a class of AKs. In addition, the AKFP can be used to reveal obsolete topics and outdated technology in advance, which can provide an early warning for decision makers to avoid unnecessary economic losses. Moreover, the AKFP can provide help for some research methods to a certain extent. For example, the AKFP can be seen as an upstream task to enrich features based on word frequency, such as keyword growth employed by Uddin and Khan (2016).

6.3. Practical guidelines

Practically, researchers may employ the AKFP to fit the developing pattern of topics in a field. After that, the trained AKFP might be adopted to predict the word frequency and determine the AKs that need to be retained for analysis, according to the number and/or

Table 9
Fitting and prediction performances for different models.

Time span	Models	Train			Test		
		MSE	MAE	R^2	MSE	MAE	R^2
5	LR	38.275	2.837	0.760	39.650	2.921	0.769
	KNN	/	/	/	40.655	2.830	0.763
	RF	25.544	2.409	0.840	35.945	2.665	0.790
	XGB	**22.812**	**2.382**	**0.857**	35.652	2.662	0.792
	LSTM	28.034	2.449	0.824	**30.513**	**2.629**	**0.822**
7	LR	69.505	3.791	0.612	62.455	3.712	0.609
	KNN	/	/	/	58.489	3.477	0.634
	RF	51.649	3.198	0.712	53.913	3.335	0.663
	XGB	**42.481**	**3.103**	**0.763**	55.976	3.367	0.650
	LSTM	54.527	3.214	0.694	**51.230**	**3.300**	**0.679**
10	LR	116.098	5.135	0.429	93.338	5.023	0.509
	KNN	/	/	/	86.828	4.650	0.543
	RF	87.452	**4.357**	0.570	75.253	4.415	0.604
	XGB	**83.217**	4.428	**0.591**	82.571	4.478	0.566
	LSTM	86.025	4.205	0.581	**71.567**	**4.270**	**0.624**

Table 10
Importance ranking of features.

Short-term prediction	1. Persistence & Community size are the dominant factors. 2. Temporal feature is the second most important factor. 3. Community development potential has minimal influence.
Medium-term prediction	1. Persistence & Community size are the dominant factors. 2. Temporal feature is the second most important factor. 3. Community development potential becomes more important.
Long-term prediction	1. Community development potential is the dominant factor. 2. Persistence & Community size are important factors. 3. Temporal feature is an un-ignorable factor.

ranking of future word frequency. The essence of the AKFP is to fit the life-cycle rules of keywords in a specific field, but these rules differ amongst disciplines. For example, in recent years, the number of publications in the computer science field has been significantly higher than that in traditional disciplines such as mathematics. Therefore, in the practical application of the AKFP, it needs to be adjusted systematically and updated in a timely manner according to technological contexts. Although the neural network algorithm was employed in this research, Table 10 shows that the common machine learning algorithms (RF, XGB) also achieved a good fitting performance on the AKFP. Therefore, for small- and medium-sized disciplines, it is vital for researchers to select algorithms appropriately according to the practical experience.

7. Limitations

There are at least four limitations in this study. Firstly, we utilized keywords to represent the research topics. However, research topics are regarded as a higher-level concept than keywords. Therefore, studies of combining co-word clustered analysis or topic model and the AKFP are worth exploring in the future. Secondly, due to the "black box" of the neural network, it is not possible to analyze the exact logical relationship between selected features accurately, which is a common drawback of neural networks. Thirdly, how to further improve the performance of AKFP 10 is a future goal. Some effective keyword features such as keyword popularity (Hu et al., 2020) and topological feature of the keyword network (Choi et al., 2011; Duvvuru et al., 2013) require further study. Finally, this paper focuses only on the AKs in the computer science field. The developing pattern of the AKs in other emerging fields needs to be explored.

CRediT authorship contribution statement

Wei Lu: Conceptualization, Methodology, Formal analysis, Supervision. **Shengzhi Huang:** Conceptualization, Methodology, Writing – original draft. **Jinqing Yang:** Data curtion, Formal analysis. **Yi Bu:** Investigation, Writing – original draft. **Qikai Cheng:** Data curtion, Writing – review & editing. **Yong Huang:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare no competing interests.

References

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*, 32–49.

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics, 13*, 485–499.

Alkhodair, S. A., Ding, S. H., Fung, B. C., & Liu, J. (2020). Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management, 57*, Article 102018.

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*, 133–144.

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. S. (2020). A topic modeling framework for spatio-temporal information management. *Information Processing & Management, 57*, Article 102340.

Behrouzi, S., Zahra Shafaeipour, S., Hajsadeghi, K., & Kavousi, K. (2020). Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics, 14*(4), Article 101079.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, 13*, 281–305.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for information science, 42*, 233–251.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM Joint Conference on Digital Libraries* (pp. 351–360). IEEE.

Chang, Y. W., Huang, M. H., & Lin, C. W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics, 105*(3), 2071–2087.

Chen, P., & Redner, Rns. (2010). Community Structure of the Physical Review Citation Network. *Journal of Informetrics, 4*, 278–290.

Cheng, Q., Wang, J., Lu, W., Huang, Y., & Bu, Y. (2020). Keyword-citation-keyword network: A new perspective of discipline knowledge structure analysis. *Scientometrics, 124*, 1923–1943.

Choi, J., & Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting & Social Change, 83*, 170–182.

Choi, J., Yi, S., & Lee, K. C. (2011). Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information & Management, 48*, 371–381.

Choudhury, N., & Uddin, S. (2016). Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics, 108*, 745–776.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2014). Research trends in gender differences in higher education and science: A co-word analysis. *Scientometrics, 101*, 273–290.

Duvvuru, A., Radhakrishnan, S., More, D., Kamarthi, S., & Sultornsanee, S. (2013). Analyzing Structural & Temporal Characteristics of Keyword System in Academic Research Articles. *Procedia Computer Science, 20*, 439–445.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14*(2), 179–211.

Ernst, H. (1997). The Use of Patent Data for Technological Forecasting: The Diffusion of CNC-Technology in the Machine Tool Industry. *Small business economics, 9*(4), 361–381.

Geng, Q., Jing, R., Jin, J., & Luo, Q. (2018). Citation Prediction and Influencing Factors Analysis on Academic Papers. *Library and Information Service, 62*(14), 29–40.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (pp. 315–323).

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., & Zamani, H., Wu, C., Croft, W. B., & (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*, Article 102067.

Hochreiter, S., & Schmidhuber, J. A. (1997). Long Short-Term Memory. *Neural computation, 9*(8), 1735–1780.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Hu, J., & Zhang, Y. (2015). Research patterns and trends of Recommendation System in China using co-word analysis. *Information Processing & Management, 51*, 329–339.

Hu, Y. H., Tai, C. T., Liu, K. E., Cai, C. F., & Egghe, L. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics, 14*(1), Article 101004.

Huang, T.-. Y., & Zhao, B. (2019). Measuring popularity of ecological topics in a temporal dynamical knowledge network. *PloS one, 14*, Article e0208370.

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour, 1*(4), 1–7.

Katsurai, M., & Ono, S. (2019). TrendNets: Mapping emerging research trends from dynamic co-word networks via sparse representation. *Scientometrics, 121*, 1583–1598.

Ketkar, N., & Santana, E. (2017). *Deep learning with python, 1*. Berkeley, CA: Apress.

Khasseh, Ali, Akbar, Chelak, Afshin, Mousavi, Moghaddam, Hadi, & Sharif, S. (2017). Intellectual structure of knowledge in iMetrics: A co-word analysis. *Information Processing & Management, 53*(3), 705–720.

Lee, Changyong, Kwon, Ohjin, Kim, Myeongjung, & Daeil. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting & Social Change, 127*, 291–303.

Li, L.-. L., Ding, G., Feng, N., Wang, M.-. H., & Ho, Y.-. S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics, 80*, 39–58.

Liu, G. Y., Hu, J. M., & Wang, H. L. (2012). A co-word analysis of digital library field in China. *Scientometrics, 91*, 203–217.

Lu, W., Liu, Z., Huang, Y., Bu, Y., Li, X., & Cheng, Q. (2020). How do authors select keywords? A preliminary study of author keyword selection behavior. *Journal of Informetrics, 14*, Article 101066.

Mccain, K. W. (2014). Assessing an author's influence using time series historiographic mapping: The oeuvre of conrad hal waddington (1905–1975). *Journal of the Association for Information Science and Technology, 59*(4), 510–525.

Peset, F., Garzón-Farinós, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J. L., et al. (2020). Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology, 71*, 462–473.

Rezaeian, M., Montazeri, H., & Loonen, R. C. G. M. (2017). Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation. *Technological Forecasting & Social Change, 118*, 270–280.

Robson, B. J., & Mousques, A. (2014). Predicting citation counts of environmental modelling papers. In *International Environmental Modelling and Software Society (IEMSs) 7th International Congress on Environmental Modelling and Software*.

Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics, 14*(3), Article 101039.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Santa Soriano, A., Álvarez, C. L., & Valdés, R. M. T. (2018). Bibliometric analysis to identify an emerging research area: Public Relations Intelligence—A challenge to strengthen technological observatories in the network society. *Scientometrics, 115*, 1591–1614.

Suominen, A., & Newman, N. C. (2017). Exploring the fundamental conceptual units of technical emergence. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1–5). IEEE.

Swami, A., & Jain, R. (2013). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Taylor, M., & Taylor, A. (2012). The technology life cycle: Conceptualization and managerial implications. *International Journal of Production Economics, 140*, 541–553.

Trevisani, M., & Tuzzi, A. (2018). Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based Systems, 146*, 129–141.

Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management, 48*, 303–325.

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics, 10*, 1166–1177.
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the association for information science and technology, 69*, 290–304.
Wang, X., Cheng, Q., & Lu, W. (2014). Analyzing evolution of research topics with NEViewer: A new method based on dynamic co-word networks. *Scientometrics, 101*, 1253–1271.
Xu, S., Hao, L., An, xin, Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*, Article 100983.
Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 51–60).
Zeiler, M.D. (.2012). Adadelta: An adaptive learning rate method. *arXiv preprint arXiv*:1212.5701.
Zeng, A., Shen, Z., Zhou, J., Fan, Y., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications, 10*(1), 1–11.
Zhao, W., Mao, J., & Lu, K. (2018). Ranking themes on co-word networks: Exploring the relationships among different metrics. *Information Processing & Management, 54*, 203–218.
Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science & Technology, 66*(2), 408–427.