Measuring topic network centrality for identifying technology and technological development in online communities

Zaoli Yang[1], Weijian Zhang[2], Fei Yuan[1], Nazrul Islam[3,*]

[1]College of Economics and Management, Beijing University of Technology, Beijing 100124, China;
[2]Tianjin 600 light-year intelligent technology Co. LTD, Tianjin 3001433, China
[3,*]University of Exeter Business School, Exeter, UK

Abstract: Online communities are a rapidly growing knowledge repository that provides scholarly research, technical discussion, and social interactivity. This abundance of online information increases the difficulty of keeping up with new developments difficult for researchers and practitioners. Thus, we introduced a novel method that analyses both knowledge and social sentiment within the online community to discover the topical coverage of emerging technology and trace technological trends. The method utilizes the Weibull distribution and Shannon entropy to measure and link social sentiment with technological topics. Based on question-and-answer and social sentiment data from Zhihu, which is an online question and answer (Q&A) community with high-profile entrepreneurs and public intellectuals, we built an undirected weighting network and measured the centrality of nodes for technology identification. An empirical study on artificial intelligence technology trends supported by expert knowledge-based evaluation and cognition provides sufficient evidence of the method's ability to identify technology. We found that the social sentiment of hot technological topics presents a long-tailed distribution statistical pattern. High similarity between the topic popularity and emerging technology development trends appears in the online community. Finally, we discuss the findings in various professional fields that are widely applied to discover and track hot technological topics.

Keywords： Technology identification; Network centrality; Online communities; Sentiment analysis; Weibull distribution

1 Introduction

Online media, such as communities, forums, microblogs, and social networking platforms, contain a wide range of topics that are discussed by the general public and professionals (Rotolo et al., 2015; Li, 2019). Such online information about hot events or technological topics popularized via online social networks helps people perceive the associated development trends (Li, 2019). Professional online communities, such as "Zhihu", have been attracting attention as new mobile technologies mature, and the number of topics related to hot and emerging technologies discussed in online communities is growing exponentially. Most internet users may learn about hot or attractive technological topics from the personalized recommendations of online communities. However, such topics are disorganized and inefficient due to a large amount of interacting information, and it is incredibly challenging to obtain "tailored" technological topics and development trends. Thus, a significant research direction in the field of technology management is to construct an

effective method of mining relevant information on technological topics in online communities and accurately identifying and forecasting technology development trends (Rotolo et al., 2015; Hong and Han, 2002; Breitzman and Thomas, 2015).

Moreover, the development stages of hot technology could be identified and tracked to provide guidance for policymakers to formulate economic and management policies and enterprises to adopt R&D strategies for relevant technologies (Yoon and Park, 2007). Academia has investigated methods of identifying and tracking the popularity trend of technologies. Traditional methods include subjective experiments, which are still the mainstream approach and include subjective scoring based on expert evaluation (Alberto et al., 2020；Simon et al., 2020) and evaluating processes impacted by subjective factors (Wang et al., 2015; Sakti et al., 2017). Subjective methods require expert evaluation results or decision-maker opinions and require experts to review many documents. These methods have drawbacks. Creating conditions for long-term, large-scale identification and tracking is difficult, and the prediction results are influenced by the experts' tendencies, which makes the results unreasonable. Therefore, subjective methods based on expert evaluation have high-efficiency advantages but are weak in accuracy, consistency, and stability and unqualified as systematic methods.

Another approach includes objective methods, such as the evaluation index method based on patent citation data, which builds a patent citation network and identifies the trend of technology development by analysing the characteristics of network structure (Liu and Wang, 2010; Du et al., 2020；Shubbak, 2019; Berg, 2019; Lee et al., 2018；Kyebambe et al., 2017; Mejia, Kajikawa, 2020; Li, 2020). The objective methods recognize the possible existence of new technologies in the future by analysing the structural characteristics of patent citation networks. Thus, the findings correspond to observable reality with rationality and provide a theoretical basis for further studies (Kyebambe et al., 2017; Mariani et al., 2019; Li et al., 2019). Authors have directly analysed the patent development trend through statistical data upon patent (Guo et al., 2018; Evangelista et al., 2020). However, this approach ignores the phenomenon that new and emergent technology may be breakthrough technology, which only slightly depends on previous patent outputs, thus leading to objectivity "failure" when used to monitor emerging technologies.

A third type of methodology is patent/literature-based text mining, such as text clustering (Lee et al., 2020; Zhou et al., 2019; Martin and Hüseyin, 2019; Kim et al., 2020; Kwon and Park, 2018), co-occurrence analysis (Diego et al., 2019; Dotsika and Watkins, 2017; Jaewoo, Woonsun, 2014; Ravikumar et al., 2015), topic modelling (Chen et al., 2017; Erzurumlu and Pachamanova, 2020; Jeong et al., 2019; Chen et al., 2020; Qiu and Wang, 2020), machine learning (Lee et al., 2018; Kristjanpoller and Minutolo, 2018; Kim and Sohn, 2020; Xu et al., 2019; Noh and Lee, 2020) and bibliometric analysis (Merino, 1990; Zhao et al., 2020), which are performed to identify technological topics and forecast development trends. The text-mining method to identify emerging technologies provides insights beyond patent data evaluation, reveals the transparency of patent structures, and may enhance the accuracy of identifying emerging technologies.

These three methodologies represent different strategies for identifying technology trends and enrich the theoretical potential of technological forecasting. Nevertheless, these methods of investigating technology are limited to patents or academic literature and ignore the

external social and societal needs that will impact technology development in the future. Ena et al. (2016) pointed out that an increasing number of data sources address different phases of technology development. Some scholars analyse social media data to examine the topical change of emerging technologies and identify development trends of emerging technologies on Twitter (Li et al., 2019) and monitor new research topics or fields according to Facebook mentions and citations (Mohammadi et al., 2020) and other web news (Hong and Han, 2002; Li et al.; 2020). Li et al. (2019) and Mohammadi et al. (2020) pinpointed the advantages of mining user-generated content in social media for technological forecasting. First, the importance of a technology is determined and characterized by engineers and public users' shifting attention but not by the relative position of the technology in its scientific and technological system. Second, it is more timely and forward-looking to identify emerging technologies and trends by social media data than by patents or academic literature. Third, recent efforts (Li et al., 2019; Mohammadi et al., 2020; Li et al.; 2020) have expanded the technology trend monitoring methodology from the perspective of multisource data utilization. In particular, monitoring and identifying emerging technologies' popularity through public perception is conducive to discriminating different development stages of emerging technologies.

Nevertheless, the core techniques of current social media analysis methods mainly focus on text mining and topic modelling, and they are not differentiated from the traditional technology trend identification method based on text mining. The current study considers user-generated topics and social activity-derived data, such as Likes, Sharing, and Forwards. These minimally socially derived data reflect a topic's degree of popularity based on the importance of an emerging technology and social network communication characteristics. However, users with a central location of social networks may exaggerate the popularity of some topics that do not correspond to reality and a technical topic posted by a professional influencer may be forwarded by their followers; therefore, the technical topic's influence may exceed that of the related topic posted by many other ordinary professionals, which leads to the distortion of information, resulting in the inaccuracy of technology trend prediction.

In addition, mainstream internet data mining maps a network and determines the significance of a node in the network by calculating the centrality or intermediate level of nodes (Du et al., 2020; Li et al., 2019). Previous methods considered the networked node weight, which is not equivalent to random connections in a social network. Hence, determining methods of trimming the "social network attributes" from topic-based social sentiment data, measuring the weight of networked nodes, and discovering their real popularity represents a critical issue for monitoring technology trends using online data.

This paper aims to develop a topic network centrality algorithm for identifying hot technological topics based on online information measurement. The method's first step is to build a technology topic network organized by a question-and-answer tag group obtained from crawling from the Zhihu platform. The second step is to calculate the weight of connections in the topic network. Here, we apply the Weibull model to fit the random distribution characteristics of social sentiment data and then use Shannon entropy theory to measure the amount of information. Consequently, we take the average amount of information as the weight of connections. The third step is to measure the centrality of each node according to complex network theory. Finally, the popularity trend curve of

technological topics with trimmed "social network attributes" is constructed based on the topic network centrality.

The main contributions of this paper are threefold: 1) to construct a technology topic network based on the technology topic and its social sentiment data; 2) to measure the weight of the random connections of nodes in a network based on the amount of information; and 3) to develop a topic network centrality algorithm based on online information measurement for technology identification and forecasting.

The rest of this paper is organized as follows. Section 2 provides an overview of the Zhihu platform and online data processing, including the basic structure of Zhihu and online data collection and processing within Zhihu. Section 3 introduces the methodology, including network construction of technological topics, centrality degree measurement, connection weight calculation, and the method of calculating the amount of information in observed data that combines information entropy with Weibull distribution. Section 4 discusses the results of an empirical study on the artificial intelligence case. Section 5 presents the conclusions of the paper as well as limitations and recommendations.

2 Data Platform and Processing

2.1 Data structure – Zhihu social platform

The Zhihu platform (https://www.zhihu.com/) is used to study the discovery and tracking of hot technological topics based on social networks. The following is a brief introduction to the essential characteristics of the Zhihu platform.

First, Zhihu is currently the largest cutting-edge technology exchange and social networking community in the Chinese Internet world, and information is organized in a tree-like hierarchical topic distribution structure. For example, the top topic of artificial intelligence is divided into subtopics, such as pattern recognition, algorithms, human-computer fighting, and smart cities, and there are corresponding subtopics under these subtopics mentioned above. For example, the subtopic of pattern recognition covers subtopics about detailed technology application directions, including character recognition, voiceprint recognition, image recognition, face recognition, and speech recognition. In this way, various technological topics in artificial intelligence can be covered, as shown in Fig. 1.

Fig. 1. Tree-like hierarchical topic distribution structure of artificial intelligence on the Zhihu platform

The tree-like topic structure shown in Fig. 1 is generated by the countless refinements resulting from the user's collective wisdom in the community. The topic tree accurately and comprehensively covers the content of each subtopic of artificial intelligence, and emerging and new technological topics will be quickly added to a specific position in the topic tree. The collective wisdom of spontaneous organization originates from many community users from all walks of life. The submission of new topics is quite timely, and the comprehensiveness and correctness of the topic tree are highly guaranteed after refinement by people with different professional backgrounds countless times. This Wikipedia-like phenomenon is called the power of collective wisdom.

Second, Zhihu has an effective intelligent recommendation system. When users pose questions on a topic, the questions will be pushed to relevant users to answer through two channels: the recommended users and the users with excellent answers under the related topic. Once the answer is completed, it will be pushed to the timeline of relevant users by the intelligent recommendation system. Social interactivities, such as "likes" and "comments", will occur as users browse. From the social interactivities, the intelligent recommendation system will select better answers and put them at the top for an efficient browsing experience. In other words, the better the answers to a question, the more quickly they will be browsed and combined via users to form a central hot topic in the community, thus indicating breakthrough or significant achievements of relevant technologies in a period.

Third, each question in the Zhihu community has multiple topic tags, which means that the

question is cross-correlated with various technologies. An example is shown in Fig. 2:

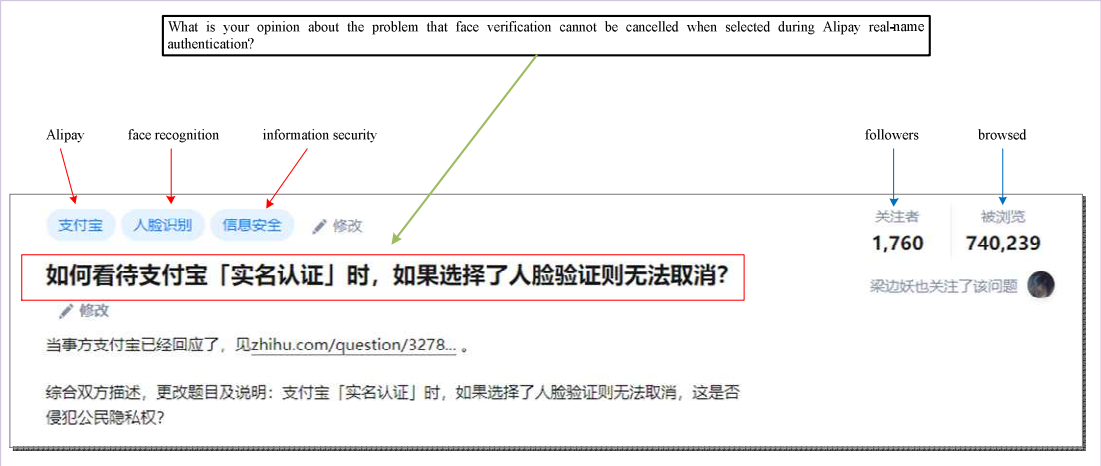

Fig. 2. Basic structure of the "Question" in Zhihu

In Fig. 2, the question "What is your opinion about the problem that face verification cannot be cancelled when selected during Alipay real-name authentication?" involves three topics, i.e., "Alipay", "face recognition", and "information security." "Alipay" is an enterprise, "face recognition" is a technology, and "information security" is an application. Thus, this problem connects many topics in different fields together, and through these connections, a wide variety of topics are organized into a network. Subsequently, hot topics in the network will be discussed and identified. In addition, Fig. 2 shows that the two tags "followers" and "browsed" are marked on the top right of each question to represent the "strong concerns" and "weak concerns" of community users about this topic.

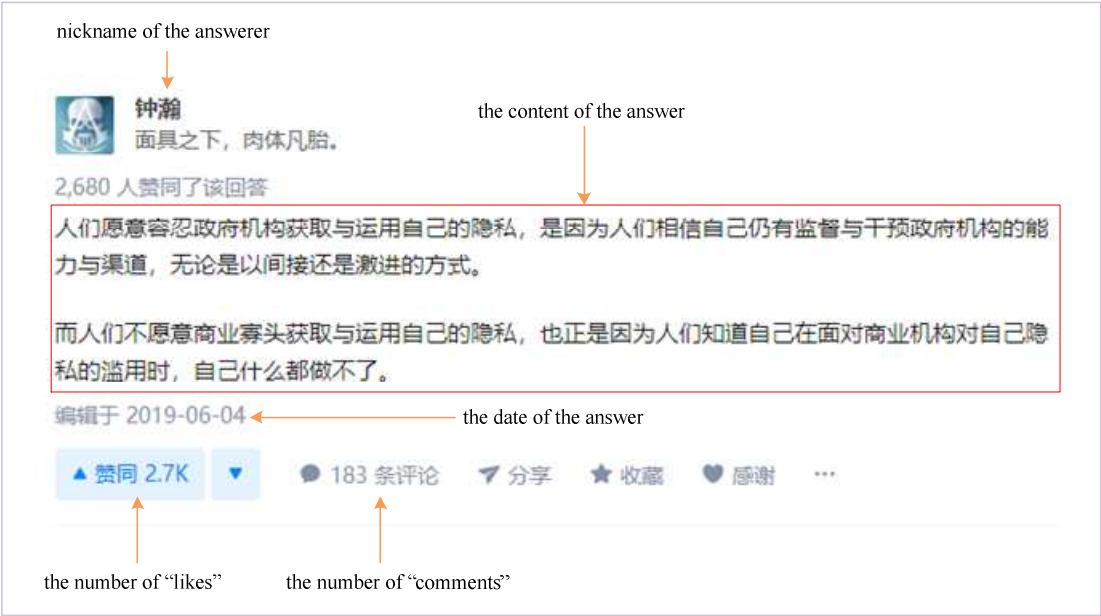

Fig. 3. Basic structure of the "Answer" in Zhihu

As shown in Fig. 3, each answer has many elements that show the popularity of the topic,

such as the number of "likes" and "comments" and the date of the answer. Moreover, the number of answers under each topic and each answer's length indicate how much attention the topic receives. Thus, the popularity of a topic or topic combination can be monitored according to these elements.

By considering these factors, we found that the Zhihu website is a high-quality and credible information source and provides enough information to monitor emerging technologies. Compared with patent information or published literature, Zhihu as a data source has the following three advantages.

First, it is forward-looking. Patents and published literature represent outcome documents of emerging technologies, and as a result, they usually lag behind the development trend of these technologies. In contrast, a knowledge exchange community such as Zhihu aims to share or discuss popular expertise issues. After new technical requirements or new technical tools have been introduced and discussed on the social media platform, these new technologies attract science, technology, and innovation practitioners, which will lead to a more significant outcome and breakthrough of these technologies in the future.

Second, it is sensitive and self-organizing. Analysing outcome documents based on key technologies requires modelers to propose a model in advance and use it as a benchmark to obtain the corresponding prediction results. However, non-professionals cannot learn about emerging technologies in the embryonic stage, thus limiting the sensitivity of the mining algorithm. Every expert or professional may contribute and create content on the Zhihu platform, which is useful for a technology trend mining algorithm. Consequently, the emerging key technologies under a particular topic will always be found and added to the topic tree. Such technologies will be connected with other existing key technologies by topic tag groups to determine the relative position of these technologies in the whole technology network. In this case, the modelers can build the entire technology popularity trend index without knowing the potential hot technologies and their names in the future, and the model will automatically discover and focus on them from the technology network.

Third, it provides a perspective of the whole society. Professionals write both patent and paper texts. However, professionals' insights in a specific field are relatively narrow and cannot reflect attitudes from a wide range of societies corresponding to the development of key technologies. Zhihu is a comprehensive knowledge-based Q&A platform widely used by experts and professionals from various industries to share high-quality insights. Although answering technical questions requires a particular professional background, professional knowledge is not necessary for interactivity, such as "likes”, "comments”, "attention”, and "browsing." If a critical technology obtains many answers and a large number of page views, followers, comments, and likes, then this technology has received the collective attention of professionals and non-professionals and will more comprehensively reflect its impact on society in a wide range in the future.

2.2 Dataset preparation

This paper studies how to monitor the development trend of technology popularity based on the Zhihu platform by taking hot technology "artificial intelligence" as the sample. It is necessary to collect all questions and answers under the topic of "artificial intelligence" and related subtopics. In this paper, more than 700,000 answers were collected through a crawler technique, and they cover the following aspects.

First, subtopics of artificial intelligence in the topic tree and subtopics of subtopics are retrieved.

Second, all the questions under each topic are collected, including the questions' title, tag group, content, number of followers, page view number, and number of answers.

Third, all answers to each question, including the answerer's nickname, answer's content and time, and the number of likes and comments are retrieved.

The final dataset takes an answer as a basic data unit that consists various attributes. 1) The tag group. For example, in Fig. 2, the tag group of multiple answers is "Alipay, face recognition, information security." 2) Number of likes, which reflects the recognition degree of users for an answer. 3) Number of comments, which reflects the attention of users to the answer. 4) Text content of an answer. 5) Number of views, which reflects the current overall popularity of a question. 6) Number of followers. If readers are interested in a follow-up discussion of the question and want to obtain more relevant pushes, they may actively click on "Follow", which reflects the users' continued attention to the question. 7) Submission time of the answer. In addition, some auxiliary information are included, such as the ID, user name, signature, and title (excellent answerer) of the answerer.

3 Methodology

The method is used to measure the information amount of socially derived data related to network topics, the connection weight based on the amount of information, and the node centrality of the topic network. First, we introduce how to build a topic network based on the topic and its socially derived data and introduce the primary method to calculate the node centrality in the topic network. Second, due to the randomness of social data, the node connection weights in the topic network cannot directly be obtained by the absolute value of derived data; thus, the corresponding information amount and Shannon information entropy theory are used to convert the socially derived data into an information amount. The connection weight is obtained by calculating the average information amount. Third, to accurately measure the information amount, it is necessary to investigate the distribution characteristics of information propagated by nodes in the topic network so that we can introduce the distribution characteristics of the final propagation times of information (such as the "number of comments" and "number of thumbs-up") using the Weibull distribution. Afterwards, we integrate Shannon information entropy and the Weibull distribution model to introduce a complete calculation method for converting socially derived data into an amount of information. Finally, by combining the above process and merging the information measurement with a weighted topic network, we propose the framework of the topic network centrality algorithm based on information measurement for computing the network centrality and forecasting technology trends.

3.1 Network construction and centrality measurement of technological topics

After collecting all answers in the artificial intelligence field of the Zhihu platform through web crawler technology, the topic tag group corresponding to each answer can be used to build the technological topic network. We use the standard adjacency list approach (Bonacich, 1987) to build the topic network. Specifically, topic tag groups attached to each question-and-answer are paired and placed into an adjacency list in turn. Complex network

analysis software is used to generate the corresponding network structure diagram based on the complete adjacency list. For example, if the topic tag group of a particular answer is (Alipay, face recognition, information security), then (Alipay, face recognition), (Alipay, information security), and (face recognition, information security) are added to the adjacency list. According to this principle, when all the answers are updated to the adjacency list, the technology topic network can be generated based on the adjacency list. Fig. 4 shows the partial area slicing of the constructed technology topic network.

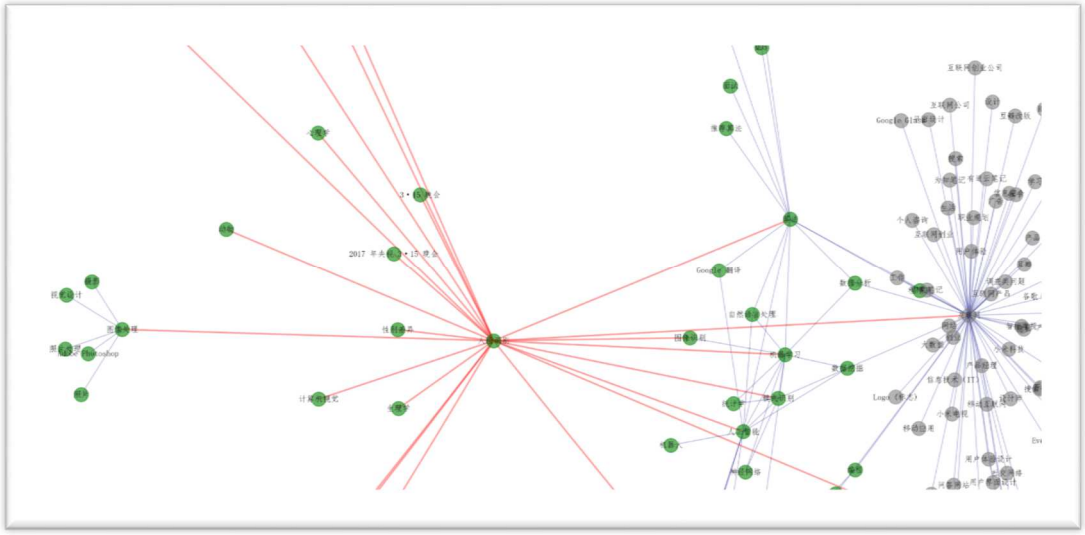

Fig. 4. Partial area slicing of the constructed technology topic network

Specifically, the network constructed in Fig. 4 is an undirected weighting network. In other words, the order in the topic tag group is not distinguished and the topic pairs added to the adjacency list based on each answer will carry different weights.

In complex network systems, the importance of a node is usually described by the concept of "centrality." There are many ways to measure centrality, and they can be roughly divided into two categories: the centralization principle, which reflects the node's local connection density; and the betweenness principle, i.e., the number of nodes that the shortest path between any two nodes in the network needs to pass through, which reflects the role played by the node in the network connectivity. Since this paper studies the popularity of a technological topic node, it is more suitable to adopt the centralization principle to measure centrality. Therefore, the network centrality calculation method based on eigenvalues is used to calculate the centrality of each node.

Assume that the centrality c_i of a node i is equivalent to the weighted sum of other nodes (Bonacich, 1987), namely:

$$\beta c_i = \sum_{j=1}^{n} w_{ij} c_j = W_i C \tag{1}$$

where w_{ij} is the connection strength between node j and node i; β is a constant; W is the adjacency weight matrix of the technology topic network; W_i is the row vector of adjacency weights corresponding to node i in the adjacency weight matrix; and C is the centrality column vector of all topic tags. It is expressed in the form of a matrix as follows (Bonacich, 1987):

$$\beta C = WC \tag{2}$$

Obviously, the centrality vector C of all technological topic tags is the eigenvector corresponding to the maximum eigenvalue of the adjacency weight matrix W.

For this reason, the measurement results of the centrality of technological topics based on social networks can be given according to formula (2). However, due to the long tail characteristic of socially derived data, the direct use of the absolute value of derived data as the connection weight in the topic network will lead to the centrality estimation being accidentally high and does not have a stable trend. Therefore, we need to construct a reasonable connection weight in the topic network.

3.2 Calculation of adjacency weight using Shannon information entropy

In the centrality calculation mentioned in the previous section, the critical step lies in obtaining the adjacency weight. Thus, from the perspective of information theory, information entropy (Shannon, 1948) is introduced to calculate the topic network's adjacency weight.

The technology-related data from social networks reflect the degree of attention given to certain technologies; however, since these observation data are derived from social networks (Barabási and Albert, 1999), they are inevitably scale-free, which is the essential feature of social network data. Specifically, scale-free refers to the severe heterogeneity of social network data, in which very few nodes obtain a large number of connections and most nodes only connect with a small number of nodes. Fig. 5 shows the elemental distribution of social network-derived data (taking the number of words in answers, the number of likes, and the number of comments as examples). It can be seen that the data distribution presents a serious imbalanced state. The samples with low observations account for a high proportion, while the samples with extensive observations account for a low proportion but have typical long-tail characteristics.

According to the above analysis, when evaluating topic popularity based on social network data, it is necessary to strip the social network attributes and identify the core information.

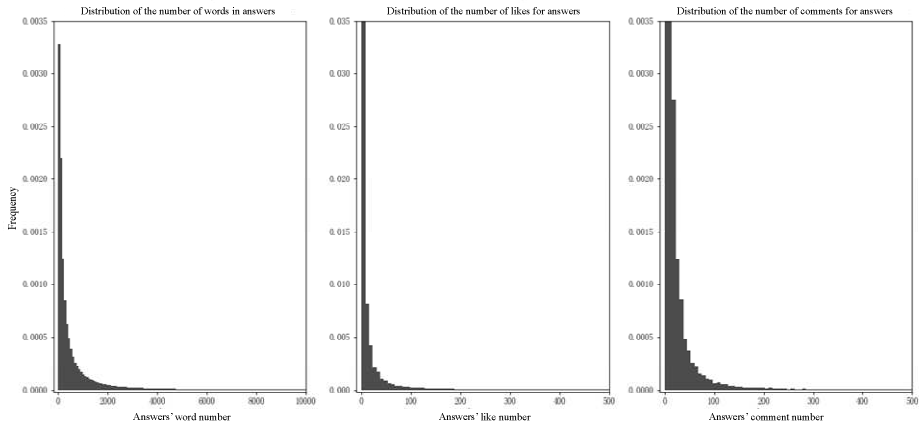

Fig. 5. Distribution characteristics of data derived from social networks

Shannon's information entropy theory establishes a complete set of information evaluation methods for random observation events. According to this theory, the information amount of a certain observed state is a negative logarithm of the probability that the state occurs; that is, if

a certain state x_i occurs, then the information amount it carries is $-\log(\mathrm{P}(x_i))$ (Shannon, 1948). In other words, the lower the probability that a certain state will occur, the greater the amount of information it carries. Therefore, according to information entropy theory, the measurement of the amount of information carried by a specific observation in social network data does not depend on the absolute amount of the observation but the probability of its occurrence. For the probability calculation, the probability distribution of observations should be described. For this purpose, a typical long-tailed distribution, i.e., the Weibull distribution, is selected to describe the data derived from social networks and then the distribution characteristics of the information occurrence probability are shown in the following subsection.

3.3 Weibull distribution

When specific information spreads from a source node in the social network, whether it can reach the next node is unclear. The information jumps from the source node to the first-level child node, from the first-level child node to the second-level child node, and then to the third-level child node and fourth-level child node. The success rate of the jump gradually changes, and the success rate sequence can be recorded as $\{p_i\} = \{p_0, p_1, p_2, ...\}$, wherein p_0 is the success rate to the source node and p_i is the success rate to the *i*-th child node. In this way, the spread of information on a social network can be regarded as its "survival." The more node jumps, the longer the corresponding "survival time". Finally, the amount of information spread is equal to its "survival time."

It is noteworthy that in the whole process of information jumping, three situations apply to the failure probability of jumping: the first is that the probability of failure does not change as the number of jumps increases; the second is that the failure probability initially increases and then decreases as the number of jumps increases; and the third is that the failure probability gradually decreases as the number of jumps increases. Objectively, there is no situation in which the failure probability continues increasing because information cannot be disseminated on a large scale. The above three cases correspond to the three probability distributions of survival analysis theory.

(1) Exponential distribution:

$$f(t;\lambda) = \lambda e^{-\lambda t}, t > 0, \lambda \geq 0$$

(2) Lognormal distribution:

$$f(t;\mu;\sigma) = \frac{1}{t\sqrt{2\pi}\sigma} e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}, t > 0, \mu > 0, \sigma > 0$$

(3) Weibull distribution:

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}, t > 0, \lambda > 0, k > 0$$

The failure rate functions of these three distributions are shown in Fig. 6:

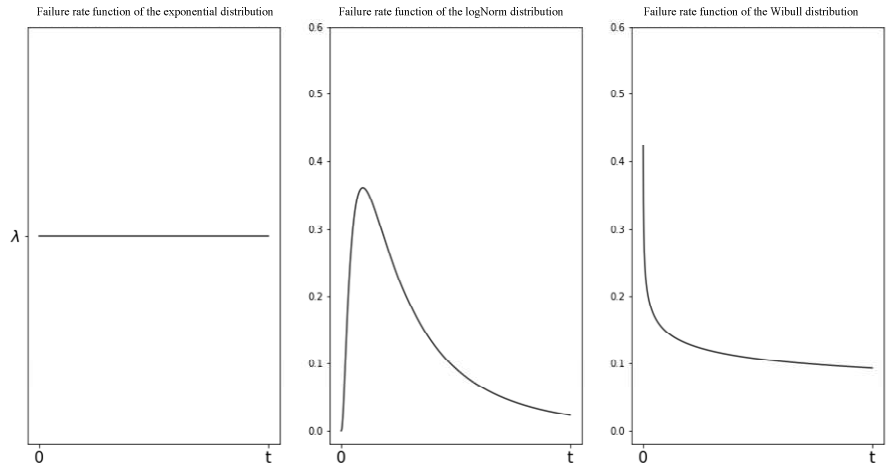

Fig. 6. Failure rate functions of the three distributions

According to Fig. 6, since the information is spread in social networks, the first jump is the most difficult. After each jump, the number of potential successor nodes that the information can reach greatly increases; therefore, the failure rate of the next jump will decrease, which is consistent with the long-tail distribution characteristics of information dissemination. In such a case, the Weibull distribution (the shape parameters are between 0 and 1) is the most appropriate benchmark distribution for characterizing information dissemination. Although the failure rate of the lognormal distribution is also decreasing in the tail, there is a very low failure rate interval at the head, which indicates that the information dissemination phenomenon described by it can complete at least the first few jumps. This finding is not consistent with the common phenomenon of zero interaction information in social networks. Similarly, the exponential distribution obviously cannot reflect this characteristic of social network information dissemination.

Therefore, according to the survival analysis theory, the data derived from social networks will follow a Weibull distribution (Weibull, 1951):

$$f(x;\lambda,k)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^{k}},x>0 \tag{3}$$

where λ is the scale parameter, which is used to adjust the dimension of random variables, and k is the failure parameter, which is used to measure the success rate in the process of information jumping. When $k>1$, the success rate of the information jump will gradually decrease as the number of jumps increases, whereas when $0<k<1$, the success rate of the information jump will gradually increase as the number of jumps increases.

In terms of the information dispersal mechanism, the data derived from the online information obey a Weibull distribution. Nevertheless, whether the Weibull distribution is long-tailed must be validated for two reasons. First, the scale-free property and the corresponding long-tailed distribution are core features of social network data. Second, the logarithmic form of probability needs to be removed when calculating information entropy. If the probability value is too small, a tremendous amount of information will be estimated, and such an extreme value will destroy the robustness of the estimated result.

The long-tailed distribution is a family of distributions with a broad definition. The probability density function of the distribution is required to converge to 0 at the power exponent level. Under this requirement, a long-tailed distribution subfamily with better performance can be defined based on Definition 1:

Definition 1 (Weibull, 1951; Almalki and Yuan, 2013): If a certain distribution *F* belongs to the $\mathcal{L}$ distribution family, then for any $y > 0$, we can obtain the following:

$$\lim_{x\to\infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = 1 \tag{4}$$

where $\bar{F}(x) = 1 - F(x)$ is the tailed distribution of distribution *F*.

In Definition 1, the probability of selecting any two points is the same on the infinite tail. In the social network scenario, Definition 1 means that regardless of the extension, the distribution may still occur; thus, the probability of occurrence remains unchanged. This property is significant for ensuring the robustness of Shannon's information entropy estimation.

Then, it is further proven that the Weibull distribution belongs to the $\mathcal{L}$ distribution family:

Proof:

The tailed distribution of the Weibull distribution is as follows:

$$\bar{F}(x) = \begin{cases} 1 & x < 0 \\ e^{-(x/\lambda)^k} & x \geq 0 \end{cases}$$

Then:

$$\lim_{x\to\infty} \frac{\bar{F}(x-y)}{\bar{F}(x)} = e^{(\frac{1}{\lambda})^k[x^k-(x-y)^k]}$$

where $x^k - (x-y)^k$ can be viewed as the difference between the function $f(x) = x^k$ and a fixed distance *y*.

When 0<*k*<1, the first derivative $f'(x) = kx^{k-1}$ of the function $f(x) = x^k$ converges to 0 from x→∞; thus, the difference in the finite distance is 0.

Therefore, it can be proven that when 0<*k*<1, the Weibull distribution belongs to the $\mathcal{L}$ distribution family. According to the fitting of actual data in the empirical results, as shown in Table 1, the estimated *k* is in the interval (0,1). Therefore, the Weibull distribution applied to social network data is a long-tailed distribution.

3.4 Measuring the amount of information derived from online social networks

After proving that the Weibull distribution can effectively describe the distribution characteristics of social network-derived data, we combine it with Shannon information entropy to construct a computing method that can measure the information amount contained in socially derived data as follows.

Step 1: Organize the processed data into a data matrix $\{d_{i,j}\}$, where *i* stands for the index of socially derived data (such as "thumbs up", "comments" and "forwards"), and j represents the ordinal number of the index. Because the definitional domain of the Weibull distribution does not include 0, the Laplace smoothing method is used, by which the value of all observed data points is added by 1.

Step 2: Take the column items in the data matrix in turn $f_j = \{d_{*,j}\}$ (* is a wildcard). Then, the following information conversion operation is performed.

(1) First, assume that the data points in a column item of features are sampled from the Weibull distribution and the distribution parameters λ, k are unknown.

(2) Then, according to the principle of maximum likelihood estimation, the negative logarithm likelihood function based on the characteristic data points of this column is constructed as follows:

$$\log P(d_{*,j};\lambda_j,k_j) = -N\log\frac{k_j}{\lambda_j} - (k_j - 1)\sum_{l=0}^{N}\log\frac{d_{l,j}}{\lambda_j} + \sum_{l=0}^{N}\log(\frac{d_{l,j}}{\lambda_j})^{k_j} \tag{5}$$

where N is the total number of records in the data set.

(3) The gradient descent method is used to solve the optimal solution of the parameters λ_j, k_j in the above equation as follows:

$$\hat{\lambda}_j, \hat{k}_j = \arg\min_{\lambda_j,k_j}\left[-\log P\left(d_{*,j},\lambda_j,k_j\right)\right] \tag{6}$$

(4) According to the optimal solution of the parameters, Shannon information entropy is used to convert each data point in this column into the corresponding information amount $I_{i,j}$ as follows:

$$I_{i,j} = -\log P\left(d_{i,j},\hat{\lambda}_J,\hat{k}_J\right) \tag{7}$$

Step 3: Each column of original data in the data matrix is converted into the amount of information, and the corresponding information amount matrix $\{I_{i,j}\}$ is obtained.

3.5 Topic network centrality algorithm framework based on information measurement

Based on information measurement, we summarize the discussed methods and propose the topic network centrality algorithm to effectively identify and monitor the development trend of hot technical topics. Assume that every piece of data can be recorded as $Item_i = [(k_{i,1}, k_{i,2}, \cdots, k_{i,n}), (d_{i,1}, d_{i,2}, \cdots, d_{i,m})]$, where $(k_{i,1}, k_{i,2}, \cdots, k_{i,n})$ represents the topic tag group and $(d_{i,1}, d_{i,2}, \cdots, d_{i,m})$ represents the corresponding features. A complete process to calculate the "activity" of each topic in the topic network during a period is as follows.

Step 1: Collect all the feature data $\{d_{*,j}\}_t$ in this period and fit the feature with a Weibull distribution to obtain the most distribution parameters matching the data. N_t represents the data point number in this period, and λ_j, k_j represents the parameters of the Weibull distribution as follows.

$$\begin{cases} \log P(d_{*,j};\lambda_j,k_j) = -N\log\frac{k_j}{\lambda_j} - (k_j - 1)\sum_{l=0}^{N}\log\frac{d_{l,j}}{\lambda_j} + \sum_{l=0}^{N}\log(\frac{d_{l,j}}{\lambda_j})^{k_j} \\ \hat{\lambda}_j, \hat{k}_j = \arg\min_{\lambda_j,k_j}\left[-\log P\left(d_{*,j},\lambda_j,k_j\right)\right] \end{cases} \tag{8}$$

Step 2: Build the Weibull distribution based on distribution parameters, convert the corresponding derived characters in each answer into its corresponding probability, and then calculate the Shannon information entropy as the information measure. The formula for measuring the information amount is as follows:

$$I_{i,j} = -\log P\left(d_{i,j}, \hat{\lambda}_J, \hat{k}_J\right)$$

Step 3: Average the amount of information for the four derived data characters, which is called the average information, and regard it as the link weight of the topic tag group attached to the answer as follows:

$$\begin{aligned}\left[\left(k_{i,1}, k_{i,2}, \cdots, k_{i,n}\right), \left(d_{i,1}, d_{i,2}, \cdots, d_{i,m}\right)\right] &\Rightarrow \left[\left(k_{i,1}, k_{i,2}, \cdots, k_{i,n}\right), \left(I_{i,1}, I_{i,2}, \cdots, I_{i,m}\right)\right] \\ &\Rightarrow \left[\left(k_{i,1}, k_{i,2}, \cdots, k_{i,n}\right), \frac{1}{m}\sum_{j}^{m} I_{i,m}\right]\end{aligned} \tag{9}$$

Step 4: Pair the topic tags of each answer and attach the average information of the answer to the network's adjacency list. The form of the adjacency list is $\{(nd_1, nd_2, w_{12}), \cdots\}$, where nd_1 and nd_2 represent the nodes in the network and w_{12} represent the weight between two nodes. The calculation process is as follows:

$$\left[\left(k_{i,1}, k_{i,2}, \cdots, k_{i,n}\right), \frac{1}{m}\sum_{j}^{m} I_{i,m}\right] \Rightarrow \left\{\left(k_{i,1}, k_{i,2}, \frac{1}{m}\sum_{j}^{m} I_{i,m}\right), \left(k_{i,1}, k_{i,3}, \frac{1}{m}\sum_{j}^{m} I_{i,m}\right), \cdots\right\} \tag{10}$$

Step 5: Generate an undirected weighting network based on the constructed adjacency list.

Step 6: Calculate the centrality of each topic node in the undirected weighting network with the centrality calculation method based on the eigenvalue, i.e., the maximum eigenvector of the adjacency weight matrix *W.*

Step 7: Arrange each topic node's centrality in different periods of time series to construct a topic popularity trend curve.

Step 8: Calculate each topic node's month-on-month growth rate and select the two fastest growing topics in each period as the hot topics in that period.

4. Empirical analysis and results

This section studies the topic of "artificial intelligence" and its subtopics in the Zhihu community to verify the proposed topic network centrality algorithm for hot technological topic discovery and monitoring. Approximately 700,000 pieces of relevant "question and answer" (Q&A) data are collected to evaluate each child node's popularity fluctuation trend under the technology category.

According to the answers' date, the 700,000 answers are divided into 15 different periods from January 1, 2012, to June 30, 2019, with a temporal interval of six months. We use the topic network centrality algorithm to process the data as follows.

Step 1: Collect Q&A records from each period and summarize the socially derived data features, including "number of browses", "number of thumbs-up", "number of comments", and "number of words" on each Q&A record. According to the topic network centrality algorithm based on information measurement, we obtain the estimation results of the parameters $\hat{\lambda}, \hat{k}$ in the Weibull distribution social sentiment data, as shown in Table 1.

Table 1 Estimation results of parameters $\hat{\lambda}, \hat{k}$ in the Weibull distribution

	browse	like	comment	number of words

$\hat{k}$	0.75	0.57	0.70	0.65
$\hat{\lambda}$	2575	6.04	3.47	471

Regarding "like" and "comment", we draw the corresponding Weibull distribution probability density curve based on the parameters $\hat{\lambda}, \hat{k}$ and compare the curve to the statistics histogram of the original data set in Fig. 7. The probability density curve of the Weibull distribution and the histogram fit well, which shows that the Weibull distribution can better reflect the distribution characteristics of the social network-derived data.

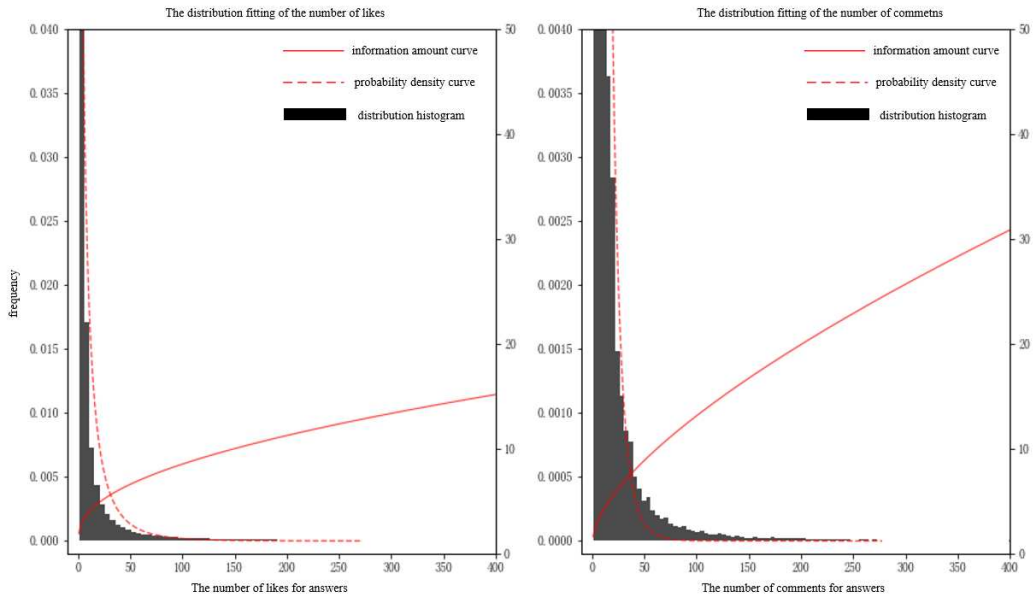

Fig. 7. Comparison between the Weibull distribution probability density curve and statistical histogram

Step 2: Based on the Weibull distribution parameter on each column of socially derived data, we convert the socially derived data on that column into the information amount and provide two examples from the Q&A records in Table 2.

As shown in Table 2, we can find that the absolute value of "thumb-up" in Case 1 is 35, and the value of that in Case 2 is 1. The information intensity of Case 1 is 35 times higher than that of Case 2, which is unreasonable. However, let us analyse this scenario from the perspective of information amount, which is 17.3 in Case 1 and 13.5 in Case 2. This finding indicates that the information intensity of Case 1 is only approximately 21% higher than that of Case 2, which is more reasonable and in line with the actual situation.

Table 2 Comparison of the information amount extracted from an absolute value

	absolute values	information amount
Case 1	(35, 25,2501, 2942)	(17.3, 19.1, 16.5, 13.9)
Case 2	(1, 2, 400, 637)	(13.5, 14.9, 13.9, 13.0)

The growth rate of the information amount in the socially derived data is different from the

growth rate of the corresponding absolute values. For example, a comparison of the growth curve of the information amount on two different columns of socially derived data, namely, "thumbs up" and "comments" in Fig. 7, we found that "comments" are more informative than "thumbs up" for the same absolute value.

Step 3: Calculate the average information amount for each corresponding Q&A record on the topic of "artificial intelligence" according to the information amount of each column of socially derived data, that is, the average information amount of four different socially derived data points mentioned in Step 1 is as follows:

$$[(Alipay, Face\ Recognition, Information\ Security), (35,25,2501,2942)] \Rightarrow$$
$$[(Alipay, Face\ Recognition, Information\ Security), (17.3,19.1,16.5,13.9)] \Rightarrow$$
$$[(Alipay, Face\ Recognition, Information\ Security), 16.7]$$

Step 4: Pair tag groups and then fill out the adjacency list with the average information amount that each pair of tags will carry as follows:

$$[(Alipay, Face\ Recognition, Information\ Security), 16.7] \Rightarrow$$
$$\{(Alipay, Face\ Recognition, 16.7),$$
$$(Alipay, Information\ Security, 16.7),$$
$$(Face\ Recognition, Information\ Security, 16.7)\}$$

Step 5: Utilize complex network software to generate the weighted topic network based on the adjacency list, as shown in Fig. 8:

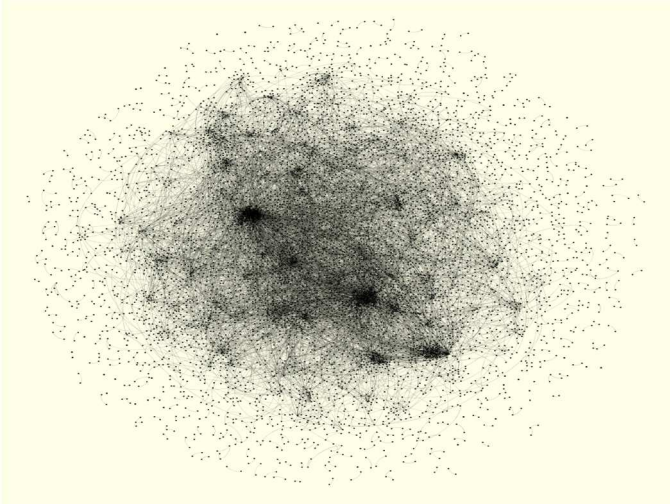

Fig. 8. Weighted topic network of "artificial intelligence" based on the adjacency list

Step 6: Calculate the centrality degree of each topic node in this period based on the eigenvalue and according to the constructed weighted topic network. Table 3 shows the calculation results of some nodes from January 2016 to June 2016.

Table 3 The centrality degree of some nodes from January 2016 to June 2016

	Face Recognition	Speech Recognition	Man-Machine Battle	Machine Learning
2016.1~2016.6	8.6	12.3	28.6	339.8

Step 7: Arrange each topic node's centrality degree in chronological order to form the corresponding topic heat trend curve. We select four different topic nodes for presentation, as shown in Fig. 8.

Step 8: Calculate each topic node's heat growth ratio in each period, which is the ratio of the centrality of the topic node in the current period to the centrality of the previous two periods. We summarize the two topic nodes with the fastest heat growth rate in different periods in Table 4.

According to the results shown in Fig. 9 and Table 4, the following analysis is performed.

Four AI technologies with high social cognition, i.e., face recognition, speech recognition, human-computer fighting, and machine learning, are selected to display their popularity trend curves, as shown in Fig. 9.

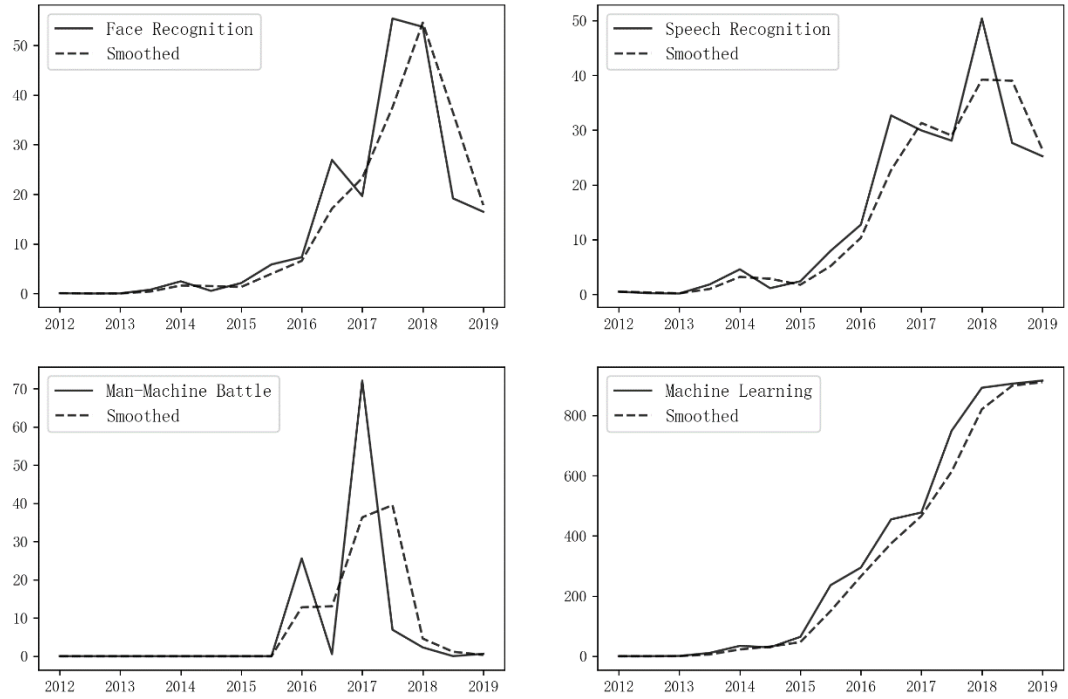

Fig. 9. Popularity trend curves of four AI technologies

One example is the Google computer program's victory over Go champions twice in 2016 and 2017, which triggered the public's attention to the technical topic of "man-machine competition" and dramatically increased the popularity of related topics. Nevertheless, this technology application occurred long before its commercialization; therefore, the topic popularity rapidly dropped to a low level after 2017. In reality, there is no other essential or noticeable research output in the field of "human-computer fighting" after Google AlphaGo. Other interesting points to consider are "face recognition" and "speech recognition", which are the most attractive applications in the field of artificial intelligence in recent years. These technologies had the highest popularity in the period of 2017-2018 when the commercial application of the two technologies was expanding quickly. However, the topic popularity has begun to decline since several US cities banned facial recognition technology due to security concerns in 2019. The last example is "machine learning", which has the highest popularity and is much more popular than other social networks. As the most successful research

paradigm and the largest branch in artificial intelligence, machine learning was widely recognized in 2014, and its popularity has been on the rise for the last five years.

In addition, by examining the rising topic popularity, it is possible to discover emerging and hot technological topics in real time. In this case, the instant popularity to average popularity ratio of the previous two periods is used as the measurement index of the instant relative popularity of a certain topic, and it screens out the two most popular artificial intelligence subtopics in each period, as shown in Table 4.

Table 4 Relative hot topics of artificial intelligence at different times

Time	First level of relative hot topics	Second level of relative hot topics	Time	First level of relative hot topics	Second level of relative hot topics
2013.7~ 2013.12	machine learning	artificial intelligence	2016.7~ 2016.12	target detection	convolutional neural network
2014.1~ 2014.6	artificial intelligence	robot	2017.1~ 2017.6	human-computer fighting	convolutional neural network
2014.7~ 2014.12	deep learning	machine learning	2017.7~ 2017.12	intensive learning	convolutional neural network
2015.1~ 2015.6	deep learning	image recognition	2018.1~ 2018.6	intelligent transportation	intelligent robot
2015.7~ 2015.12	deep learning	computer vision	2018.7~ 2018.12	target detection	domestic robot
2016.1~ 2016.6	human-computer fighting	artificial intelligence	2019.1~ 2019.6	intensive learning	text mining

As shown in Table 4, in the early days of the artificial intelligence wave (2013-2014), the relatively popular topics on social network platforms were relatively broad and lacked specific application scenarios, and the understanding of artificial intelligence was still associated with robots. Then, from the second half of 2014, deep learning emerged radically, and it was accompanied by the first clear-cut application scenario: computer vision. In 2016, the man-versus-machine AlphaGo match drew public attention to the field of "human-computer fighting", and convolutional neural networks, which is the most typical deep learning, continued to be popular. In the second half of 2017, intensive learning, as the representative of the next stage of the AI wave, represented a relatively hot topic for the first time, and AI applications close to social life, such as intelligent transportation and household robots, emerged in 2018. Finally, intensive learning and text mining, which showed the highest relative popularity in 2019, represents the first wave brought by deep learning and starts the second new wave.

From the above analysis, the four technology popularity trend curves based on the proposed method, namely, face recognition, speech recognition, human-computer fighting, and machine learning, are consistent with the actual development situation and public attention. This finding advocates that the method proposed in this paper effectively identifies and monitors hot topics in an online forum.

5 Discussion

This section will demonstrate the superiority and rationality of our method through a comparative analysis of different methods and then discuss the universality of the method’s application.

5.1 Comparison of prediction results based on the absolute value and information amount of social derivative data

The empirical results show that the topic network constructed with information amount as the weight can better reflect each topic node's importance and has high consistency with the public perception. However, what would be the impact of using the absolute value of social derivative data as the topic network connection weight instead of the information amount? To answer this question and highlight the superiority of the proposed method, we compare the prediction results based on absolute values and information amount, as shown in Fig. 10 and Table 5.

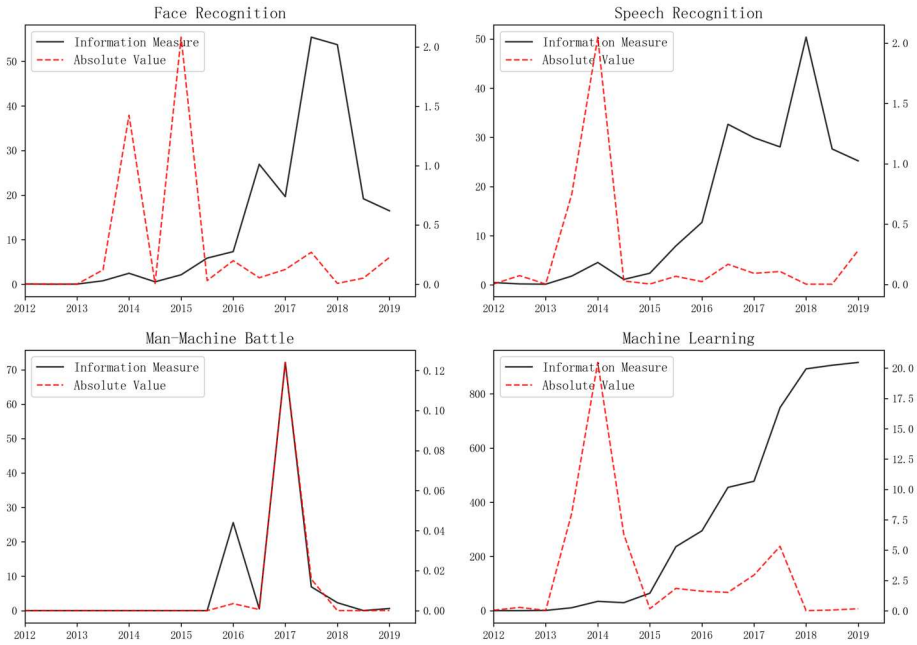

Fig. 10. Comparison of the topic heat trend curves based on information amount and absolute value

Table 5. Relative hot topics of artificial intelligence at different times based on absolute values

Time	First level of relative hot topics	Second level of relative hot topics	Time	First level of relative hot topics	Second level of relative hot topics
2013.7~ 2013.12	data mining	machine learning	2016.7~ 2016.12	machine learning	deep learning
2014.1~ 2014.6	artificial intelligence	robot	2017.1~ 2017.6	artificial intelligence	data analysis
2014.7~ 2014.12	machine learning	data mining	2017.7~ 2017.12	artificial intelligence	machine learning
2015.1~ 2015.6	deep learning	face recognition	2018.1~ 2018.6	artificial intelligence	robot
2015.7~ 2015.12	deep learning	robot	2018.7~ 2018.12	robot	face recognition
2016.1~ 2016.6	artificial intelligence	augmented reality	2019.1~ 2019.6	speech recognition	data analysis

The information amount weights the solid black line, and the absolute value weights the red dotted line. The findings show that the estimation results weighted by absolute values exhibit random volatility and cannot be correlated with the actual development trend of artificial intelligence-related technologies. Table 5 lists hot topics weighted by absolute values. We find that these topics are repeated with some conceptual and broad topics, such as "artificial intelligence", "machine learning", "deep learning", "data analysis", and "robot." Only "face recognition" and "speech recognition" are involved in the application technologies in Table 5. In the discussion of "artificial intelligence" generated by the two AlphaGo games in 2016 and 2017, the critical technology topic of "human-computer fighting" was not identified. However, the hot topics in Table 4 based on the information amount reflect the AI field's complete development history starting from the paternal concept of "artificial intelligence", "machine learning", and "robot", then gradually developing to "face recognition", "target detection" and other specific application-oriented technologies, and then further expanding to "intensive learning", "text mining" and other new AI development directions. The hot topics also monitored in real time the eruption of hot topics, such as "human-computer fighting".

Consequently, the topic network centrality algorithm based on information measurement using Shannon information entropy and the Weibull distribution can tailor the long tail inherent in the social sentiment data and accurately calculate the weights of randomly generated connections between nodes in the social network. Therefore, the proposed method effectively supports the calculation process of the centrality of topic nodes in topic networks and provides accurate results for forecasting technology trends.

5.2 Comparison analysis of technology popularity trend results based on different long-tailed distributions

To further demonstrate the superiority of the method proposed in this paper, the influence of different long-tailed distributions on the calculation results of the technology popularity trend index is compared. The classical Pareto distribution and lognormal distribution are selected as the comparison objects, the three long-tailed distributions are used for dataset grouping, characteristic information identification and transformation, network graph model construction, node centrality calculation and centrality information identification in sequence, and then a technology popularity trend index based on the amount of information is constructed. Fig. 11 compares the technology popularity trend curves based on three different long-tailed distributions.

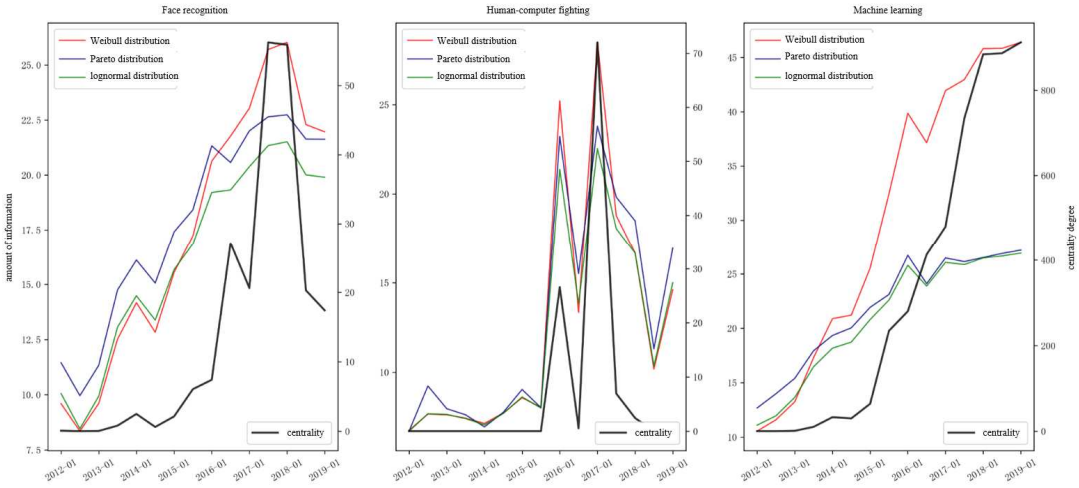

Fig. 11. Technology popularity trend curves based on three different long-tailed distributions

First, the comparison of Fig. 11 shows that compared with the original centrality trend curve (black line), the fluctuation of the trend curve based on the three long-tailed distributions is greatly reduced after the centrality is converted into the amount of information. The calculation results of the popularity trend curves of technological topics based on the above three long-tailed distributions are reasonable. For example, in "face recognition", although the centrality has dropped significantly in the last two time units, this decline is a general decline caused by the expansion of the network; therefore, the amount of information has not decreased much. Second, in the case of low centrality (less than 50 units), the estimated results of the amount of information given by the three long-tailed distributions are roughly equivalent. When the centrality increases to a higher level (more than 100 units), the trend curve of the information amount based on the Pareto distribution and lognormal distribution reaches the "peak" and no longer increases with increasing centrality; however, the trend curve based on the Weibull distribution does not show the same results. The reason for these differences may be based on the difference in the probability density function of the three long-tailed distributions.

Hence, we further compare the probability density function of the Pareto distribution, lognormal distribution and Weibull distribution as well as the difference in tail descent speed, as shown in Table 6.

Table 6 Tail descent speed of different long-tailed distributions

	probability density function	tail descent speed
Pareto distribution (Rodriguez-Dagnino, 2005)	$Pareto(x) = \alpha k^{\alpha} x^{-\alpha-1}, x \geq k > 0$	descent speed with power function level
Lognormal distribution (Fenton, 1960)	$LogNorm(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, x > 0$	descent speed with power function level
Weibull distribution (Weibull, 1951)	$f(x;\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, x > 0$	descent speed with exponential function level

According to Table 6, the estimation method of time complexity can be used to calculate the tail descent speed. Since logarithmic transformation is needed in the process of information identification, if the distribution tail descends at the power function level, then the descent speed will be quite small when the observed value x of the random variable is large enough, thereby leading to the bottleneck of information growth. Thus, when the Pareto distribution and lognormal distribution in Fig. 11 are used as the long-tailed distribution, the increase in the amount of information will cause a "ceiling" phenomenon. The tail descent speed of the Weibull distribution is close to an exponential function and maintains a constant level decline after logarithmic transformation. Therefore, regardless of how large the observed value x is, if it is doubled to $2x$, the amount of information will increase to the original fixed multiple, thus ensuring the global robustness of the information identification algorithm.

5.3 Generalizability of the proposed method

The Zhihu forum in the Chinese environment and the Quora forum (https://www.quora.com/) in the English environment are basically the same in terms of content organization, including the basic structure of questions and answers, with topic tags as the linking method between different question and answer groups. That is, in the Quora and Zhihu forums, the topic network is organized by using topic tags as links and the same weighted topic network can be constructed by using social sentiment data, such as likes and comments, as network weights to analyse the method of this article. On other online forums that do not use clear hashtags, it is possible to add hashtags to each text material through natural semantic analysis technology or keyword query technology. Therefore, with the aid of basic natural language processing technology, this method can be widely used in the discovery and tracking of hot topics in various online technical forums and even further extended to the discovery and tracking of public opinion in other types of websites, such as Yahoo, Facebook, Twitter and other online social communities.

6. Conclusions

User-generated social sentiment information in online professional communities is becoming a vital data resource for identifying technology developments and forecasting trends. For practical mining of online information to support technology forecasting, this paper has developed the topic network centrality algorithm based on information measurement. In addition, as a scientific domain attracting the most attention in recent years, artificial intelligence has a relatively clear development path with a higher degree of recognition. Therefore, by taking advantage of online community information, this paper selected the technology development of artificial intelligence as an application example to verify the effectiveness of the proposed method. The empirical analysis indicates that the algorithm proposed in the present paper accurately identified hot and trending technological topics and was able to construct popularity trend curves of different AI technological topics over the timeline. The findings presented here can provide support for state administrators of science, technology, and innovation in monitoring and evaluating the development status of emerging and frontier technologies, facilitate the preparation of related policies and provide guidance for technology research and development for relevant practitioners.

Overall, the proposed technology trend identification methodology has several features that distinguish it from similar efforts.

First, the social network itself is scale-free and all characteristics derived from the network obey a long-tailed distribution. The best method of building a technology forecasting model based on these characteristics is to convert the absolute value into the information amount by information theory, which significantly improves the model's stability.

Second, when information theory is used to extract the amount of information from data subject to a long-tailed distribution, the optimal choice of the corresponding long-tailed distribution is the Weibull distribution, which ensures better stability and sensitivity of information after conversion.

Third, eigenvalue decomposition-based methods are suitable for calculating the centrality of networked technology nodes, and such methods conform to the main characteristics of the technology system.

Fourth, the data and structure of the Zhihu community are very conducive to constructing the technology network model. The technology popularity trend index constructed based on these data also reflects the actual situation of technological development. In addition, by collecting more comprehensive data, a real-time monitoring platform of the technology popularity trend index can be constructed to provide strong support for the preparation of national industrial policies and industrial investment or related work by scientific research institutions.

Nevertheless, this paper has the following limitations that need to be considered in future work. First, this article focused on social-emotional structured data and professional users' sentiments embedded in the topic text were not involved in the current measuring process. A semantic analysis needs to be conducted in a further study to quantify and measure the answerers' sentiments towards relevant technology topics and further improve the index construction method. Second, this paper only considers the Zhihu online community as the data source platform. However, socially derived data can be mined from a variety of different platforms to build a multisource heterogeneous data-driven technical topic prediction model to improve the prediction accuracy.

Acknowledgment: This research was supported in part by the Natural Science Foundation of China (No. 71704007), the Beijing Social Science Foundation of China (No. 18GLC082), and Social Science Program of Beijing Municipal Education Commission (Grant No. SM202110005012).

Conflicts of Interest: The authors declare no conflict of interest.

References

Alberto, M., Geraldine, J., Christian, M., 2020. Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software. Futures 117, 102511.

Almalki, S.J.; Yuan, J., 2013. A new modified Weibull distribution, Reliab. Eng. Syst. Safe. 111, 164-170.

Bain, L.J., 1974. Analysis for the linear failure-rate life-testing distribution. Technometrics

16(4), 551–9.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286(5439), 509-512

Berg, S., Wustmans, M., Bröring, S., 2019. Identifying first signals of emerging dominance in a technological innovation system: A novel approach based on patents. Technol. Forecast. Soc. Change 146, 706-722.

Bonacich, P., 1987. Power and centrality: a family of measures. Am. J. Sociol. 92(5), 1170-1182.

Breitzman, A., Thomas, P., 2015. The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems, Res. Policy 44(1), 195–205.

Chen, H.S., Zhang, G.Q., Zhu, D.H, et al., 2017. Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 201. Technol. Forecast. Soc. Change 119(7), 39–52.

Chen, X., Zou, D., Cheng, G., et al., 2020. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. Comput. Educ. 151, 103855.

Diego, C.G., Marta, O.U.C., Eva-María, M.V., 2019. Knowledge areas, themes and future research on open data: A co-word analysis. Gov. Inform. 36(1), 77-87.

Dotsika, F., Watkins, A., 2017. Identifying potentially disruptive trends by means of keyword network analysis. Technol. Forecast. Soc. Change 119, 114–127.

Du, J., Li, P., Haunschild, R., et al., 2020. Paper-patent citation linkages as early signs for predicting delayed recognized knowledge: Macro and micro evidence. J. Informetr. 14(2), 101017

Ena, O., Mikova, N., Saritas, O., et al., 2016. A methodology for technology trend monitoring: the case of semantic technologies. Scientometrics 108 (3), 1013–1041.

Erzurumlu, S.S., Pachamanova, D., 2020. Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. Technol. Forecast. Soc. Change 156, 120041

Evangelista, A., Ardito, L., Boccaccio A., et al., 2020. Unveiling the technological trends of augmented reality: A patent analysis, Comput. Ind.118, UNSP 103221.

Fenton, L.F., 1960. The sum of log-normal probability distributions in scatter transmission systems," IEEE Trans. Commun. COM-8(1), 57– 67.

Guo, Y., Sun, G., Zhang, L., et al., 2018. A new model based on patent data for technology early warning research, Int. J. Technol. Manage. 77(4), 210-234

Hong, T., Han, I., 2002. Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks, Expert Syst. Appl. 23(1), 1–8.

Jaewoo, C., Woonsun, K., 2014. Themes and trends in Korean educational technology research: a social network analysis of keywords. Procedia Soc. Behav. Sci. 131, 171–176.

Jeong, Y., Park, I., Yoon, B., 2019. Identifying emerging research and business development (R&BD) areas based on topic modeling and visualization with intellectual property right data. Technol. Forecast. Soc. Change 146, 655-672.

Kim, S., Park, H., Lee, J., 2020. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. Expert Syst. Appl.

152, 113401.
Kim, T.S., Sohn, S.Y., 2020. Machine-learning-based deep semantic analysis approach for forecasting new technology convergence. Technol. Forecast. Soc. Change 157, 120095
Kristjanpoller, W., Minutolo, M.C., 2018. A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis, Expert Syst. Appl. 109, 1-11.
Kwon, H., Park, Y., 2018. Proactive development of emerging technology in a socially responsible manner: Data-driven problem solving process using latent semantic analysis. J. Eng. Technol. Manage. 50, 45-60.
Kyebambe, M.N., Cheng, G., Huang, Y., et al. 2017. Forecasting emerging technologies: A supervised learning approach through patent analysis. Technol. Forecast. Soc. Change 125, 236-244.
Lee, C., Jeon, D., Ahn, J.M., et al., 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. Technovation 5, 102140
Lee, C., Kwon, O., Kim, M., Daeil, K., 2018. Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technol. Forecast. Soc. Change 127, 291-303.
Li, S., Garces, E., Daim, T., 2019. Technology forecasting by analogy-based on social network analysis: The case of autonomous vehicles, Technol. Forecast. Soc. Change 148, 119731
Li, X., Fan, M., Zhou, Y., Fu, J., et al., 2020. Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining. Nano Energy 71, 104636.
Li, X., Xie, Q., Jiang, J., Zhou, Y., Huang, L., 2019. Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. Technol. Forecast. Soc. Change 146, 687-705.
Li, X., Xie, Q., Huang, L., 2020. Identifying the Development Trends of Emerging Technologies Using Patent Analysis and Web News Data Mining: The Case of Perovskite Solar Cell Technology, IEEE T. Eng. Manage. doi: 10.1109/TEM.2019.2949124.
Liu, C.Y., Wang, J. C., 2010. Forecasting the development of the biped robot walking technique in Japan through S-curve model analysis, Scientometrics 82(1), 21-36.
Mariani, M.S., Medo, M., Lafond, F., 2019. Early identification of important patents: Design and validation of citation network metrics, Technol. Forecast. Soc. Change 146, 644-654
Martin, G.M., Hüseyin, C., 2019. Technological speciation as a source for emerging technologies. Using semantic patent analysis for the case of camera technology. Technol. Forecast. Soc. Change 146, 776-784.
Mejia, C., Kajikawa, Y., 2020. Emerging topics in energy storage based on a large-scale analysis of academic articles and patents. Appl. Energ. 2631, 114625.
Merino, D.N., 1990. Assessing technological forecasts for the fiber optic communications market, IEEE T. Eng. Manage. 37(1), 53-55,

Mohammadi, E., Gregory, K.B., Thelwall, M., et al., 2020. Which health and biomedical topics generate the most Facebook interest and the strongest citation relationships? Inform. Process. Manag. 57(3), 102230.

Noh, H., Lee, S., Forecasting forward patent citations: comparison of citation-lag distribution, tobit regression, and deep learning approaches, IEEE T. Eng. Manage. doi: 10.1109/TEM.2020.2978528.

Qiu, Z., Wang Z., 2020. Technology forecasting based on semantic and citation analysis of patents: a case of robotics domain, IEEE T. Eng. Manage. doi: 10.1109/TEM.2020.2978849.

Ravikumar, S., Agrahari, A., Singh, S.N., 2015. Mapping the intellectual structure of scientometrics: a co-word analysis of the journal scientometrics (2005–2010). Scientometrics 102(1), 929–955.

Rodriguez—Dagnino, R.M., 2005. Some remarks regarding asymptotic packet loss in the Pareto/M/1/K queueing system", IEEE Commun. Lett. 9, 927-929.

Rotolo, D., Hicks, D., Martin, B., 2015. What is an emerging technology? Res. Policy 44(10), 1827–1843.

Sakti, A., Azevedo, I.M.L., Fuchs, E.R.H., et al., 2017. Consistency and robustness of forecasting for emerging technologies: The case of Li-ion batteries for electric vehicles, Energy Policy 106, 415-426.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27(3), 379-423.

Shubbak, M.H., 2019. Advances in solar photovoltaics: Technology review and patent trends. Renew. Sust. Energ. Rev. 115, 109383.

Simon, W., Alberto, M., Vera, R., et al., 2019. Future emerging technologies in the wind power sector: A European perspective. Renew. Sust. Energ. Rev. 113, 109270

Wang, X.F., Qiu, P.G., Zhu, D.H., et al., 2015. identification of technology development trends based on subject–action–object analysis: the case of dye-sensitized solar cells. Technol. Forecast. Soc. Change 98, 24–46.

Weibull, W.A., 1951. Statistical distribution function of wide applicability. J. Appl. Mech. 18, 293–6

Xu, S., Hao, L., An, X., et al., 2019. Emerging research topics detection with multiple machine learning models, J. Informetr. 13(4), 100983

Yoon, B., Park, Y., 2007. Development of new technology forecasting algorithm: hybrid approach for morphology analysis and conjoint analysis of patent information, IEEE T. Eng. Manage. 54(3), 588-599.

Zhao, S.X., Chen, D., Huang, M., Chang, Y., 2020. Potential value of patents with provisional applications: An assessment of bibliometric approach, IEEE T. Eng. Manage. doi: 10.1109/TEM.2019.2943135.

Zhou, Y., Dong, F., Kong, D., et al., 2019. Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. Technol. Forecast. Soc. Change 144, 205-220.