An Evolving Hypernetwork Model to Quantify Progress Potential of Emerging Research Topic

Jia Liu, *Kewei* Yang, *Jianguo* Xu, *Yingying* Gao, and *Qingqing* Yang

Department of Systems Engineering, National University of Defense TechnologyChangsha, Hunan, China

Abstract. There is considerable and growing interest in the emergence of research topics. However, current methods to detect the emergence are still problematic mainly due to information loss and aging effect. In this study, we show three intrinsic mechanisms including preferential attachment, exponentially growth and heterogeneous fitness values that decay with time. Depending on the input assumptions, all topics tend to follow a universal temporal pattern according to our model which results in strongly sufficiency to quantify progress potential.

1 Introduction

There is considerable and growing interest in the emergence of research topics, both from the policy-making and academic perspective. Numerous studies and projects have proposed definitions and models to detect, track, and predict emerging topics. For instance, the Comprehensive Strategy on Science, Technology and Innovation of Japan in 2017 (STI 2017) and the Annual Report on the Top 10 Emerging Technologies of the World Economic Forum [1]. Evaluating innovation potential is essential, not only for promoting efficiency of scientific research but also for detecting emerging research topics with strategic significance [2]. The intrinsic improvability of a new research area allows decision makers to track its evolutionary trajectory, or progress potential. It mainly includes two reasons. For one, the studies always attempt to deal with a practical or scientific problem. For another, the current study is affirmably carried out on the basis of existing one, and trying to improve the science and technology level through cooperation.

We contrast this problem of quantifying the innovation potential of emerging research topic based on its intrinsic improvability, which generally focuses on exogenous, market-driven factors or experts factors, for which bibliometrics and Delphi have played a key role. Many researchers utilize various types of relations among publications to establish network based models of bibliometrics, such as bibliographic coupling, co-citation analysis and others [3]. Different citation relations can all be used to aggregate publications. In the case of text-based approaches [4], some studies have established clusters based on the co-occurrence of terms [5], using a large-scale database [6]. However, the effectiveness of any approach for identification is extremely difficult to verify. And such approaches tend to focus on measuring the attributes of novelty and fast growth, and to ignore the consideration of other potentially important attributes, such as impact [7]. After the small-world model proposed by Watts and Strogatz, Barabási and Albert (BA) proposed the scale-free network [8]. The academic community has set off an upsurge of research on complex networks. Despite its different applications, it is still hard to depict some real-life systems.

All in all, traditional homogeneous complex networks have following limitations. There is no consensus on the concept of emergence especially the impact of co-operation. What's more, bibliometrics indicators such as the number of papers or citations are poor prediction of topic's future [9] in Fig 1. The number of scientific literature for each research point relies on the item's age [10]. As a result, the older it is, the more likely it is to be favored, which violates to the intuition that people prefer the emerging things [11]. In the experiment, we regarded Medical Subject Headings (MeSH) term as research topic. PubMed databases of the NCBI（National Center for Biotechnology Information）provide freely source in a well formatted structure. (which is publicly available at http://www.ncbi.nlm.nih.gov/pubmed/advanced) [12]. Moreover, it is also short of predictability: a subset of research points that collect nearly 500 papers in 1966 own widely different evolutionary potential (Fig. 1, inset). To avoid information loss and aging effect, it naturally leads to the emergence of hypernetworks. The concept of hypernetworks is a natural multidimensional generalization of networks and represents n-dimensional relations. The evolving hypernetworks in the existing literatures are almost all uniform, except Guo and Zhu [13] considering the characteristics of non-uniformity and weight of hyperedges. In the non-uniform model, at each time step, both the size of new nodes and the randomly selected existing nodes in one hyperedge are random variables.

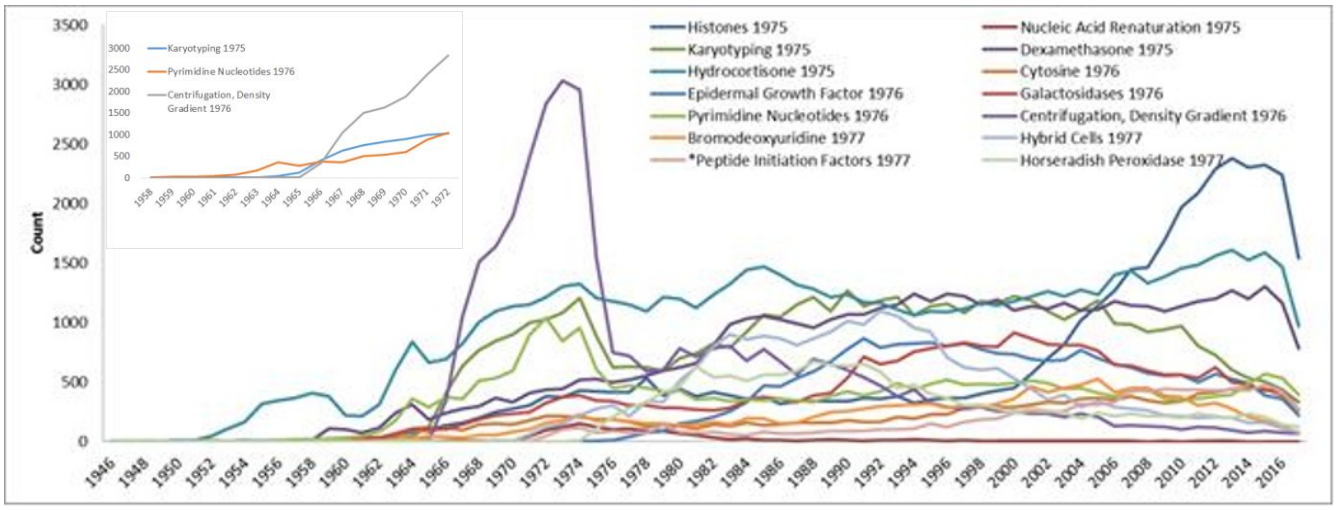

Fig. 1. The life curves of 14 topics selected from 1975 to 1977 in 'CELL'.

2 The evolving hypernetwork model

2.1 Fundamental mechanisms

We started by hackle intrinsic mechanisms to drive new research topic evolving. Intuitively, the life curves of different MeSH terms vary widely in Fig. 1. Some topics are highly valued when they are first proposed. While, some of them have been dormant for a long time before they catch the attention of people, that is, the curve rises out rapidly after a period of time. Some have been developing steadily, that is, the curve flattens out. In addition, in terms of total quantity, some of the hot topics exist in dozens or even hundreds of times of literature, compared to the general topic.

Preferential attachment captures the fat-tailed nature of the publication distribution which several hotspots are more likely to be used and finally result in 'the rich get richer' 14. Despite its rationality in comparing the empirical data and theoretical models, the first-mover advantage of preferential attachment results from strong time bias in the system.

Secondly, unlike the traditional uniform network, the total amount of publications each year is not the same, but increases exponentially year by year15.

$$N(t) \sqcup \exp(\beta t) \tag{1}$$

where $N(t)$ is the amount of documents at the time t, β is the speed of the literature growth. Then the scholars can noticed that the proportion of the literature must be reduced rapidly. Furthermore, the evolving network in the existing publications is non-uniform. That is, the number of research points of each paper is not a constant which is determined by a given distribution function like Poisson function. There are approximately 20 MeSH terms in each publications of 'CELL', new nodes entering into the network or previously existing nodes selected may not be the same in Fig 2.

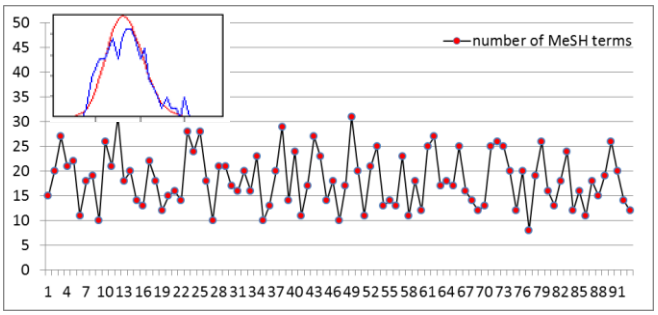

Fig. 2. The number of MeSH terms of 100 papers.

Old research points will be replaced by new ones eventually. Therefore, in the long run, the new research subject will dies slowly and the total amount of relevant literature will also tend to be stabilized, such that the trend is similar to the shape of the normal distribution. But it is difficult to determine when this distribution begins, due to the difference in the innovative characteristic of the new research point, i. e. Is it a standard normal distribution, left or right?

In the beginning, the research pioneers creatively propose a new research idea and relative researchers with the early derivation of theories. Once forming the theory and receiving more attention of experts in the field, which usually takes a long time, much more researchers, on the basis, start a large number of extension and application researches, until the hotspot is replaced by another. For instance, there is a new MeSH term in 2015 called latent autoimmune diabetes of adults in Fig. 3(New appearance in 2015 in left subfigure and emerging in 2016 on the right).

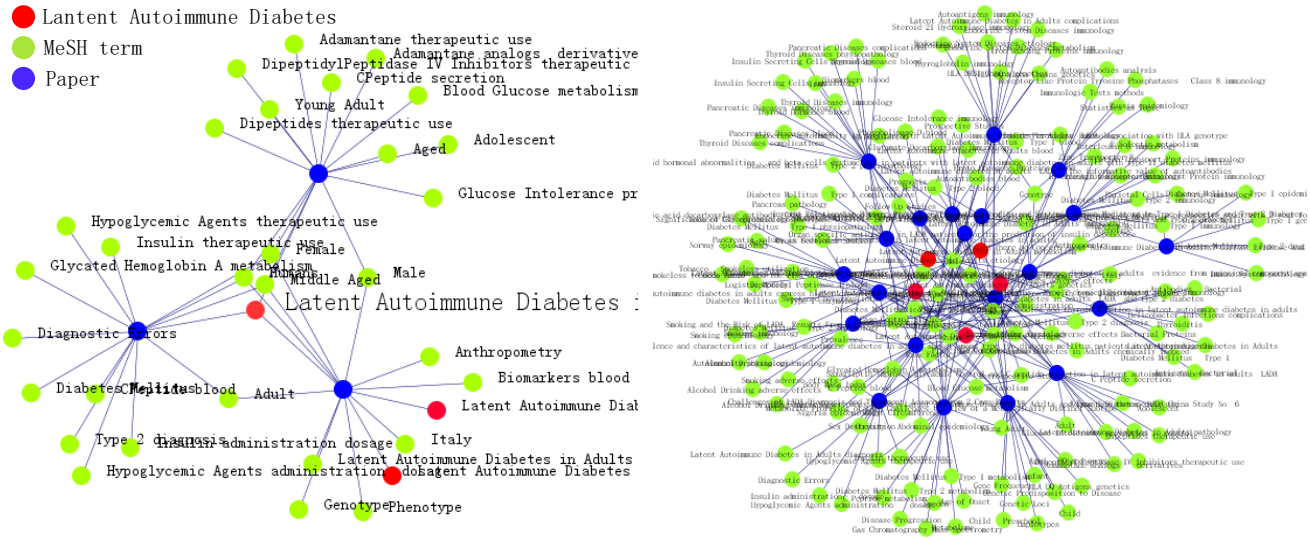

Fig. 3. A sample hypernetwork of latent autoimmune diabetes of adults.

2.2 Theoretical analysis

Now we proceed to an evolving hypernetwork model based on the above-reported empirical observations. Initially, we assume the hyperdegree h_i^{t0} initially equals to 1. The hyperdegree is the number of hyperedges that connect to the node, and the number of related papers in the real world. At time *t*, $\exp(\beta t)$ new papers are introduced. The number of each paper *k* is determined by probability $p(k)$. The time evolution of expected number of h_i^t satisfies

$$\frac{dh_i}{dN} = (p_k k)\frac{h_i R(\Delta t_i)}{\sum_{j=1}^{N} h_j R(\Delta t_j)} \tag{2}$$

where $R(\Delta t_j)$ is the relevance of time which is similar to normal distribution.

$$R(\Delta t_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(\Delta t_i^{\omega} - u_i)^2}{2\sigma_i^2}\right) \tag{3}$$

where u indicates the time for a paper to reach its citation peak, σ is longevity, capturing the decay rate, ω is the degree of skewness.

The equation represents the evolution of each node's hyperdegree depends not only on itself but also on the current hyperdegree and relevance of all other nodes. Two simplifications could be introduced to solve this partial differential equation. At first, the number of MeSH terms each new paper contains is subordinated to an extremely narrow Poisson distribution

$$p(k) = \frac{\lambda^k}{k!}e^{-\lambda}, (k = 0,1,\cdots) \tag{4}$$

where its expectation and variance are λ. According to the Poisson theorem, $\lim_{t\to\infty} p(k)k \to \lambda$. Secondly, there are millions of nodes with non-negligible infects of infinite series $\sum_{j=1}^{N} h_j R(\Delta t_j)$. Combining with (1) leading to $\Delta t_i = t - t_i = \ln(N/i)$, we get $dN = \beta N d\Delta t_i$. While $R(\Delta t_j)$ decays faster than $\frac{1}{\Delta t_j}$ and $\lim_{\Delta t_j\to\infty} R(\Delta t_j) \to 0$, the finally growth of the infinite series converges to a certain value $\Omega^* = \frac{1}{N}\sum_{j=1}^{N} h_j R(\Delta t_j)$. This allows us to approximate (3) with

$$\frac{dh_i^t}{d\Delta t_i} = \beta\lambda\frac{h_i R(\Delta t_i)}{\Omega^*} \tag{5}$$

With the initial condition $h_i^0 = 1$, assuming $\alpha_i = \frac{\lambda\beta h_i}{\Omega^*}$, thus

$$h_i^t = \exp\left(\alpha_i \Phi\left(\frac{\Delta t_i^{\omega} - u_i}{\sigma_i}\right)\right) \tag{6}$$

When $\omega > 0$ and $\Delta t_i \to \infty$, $\Phi\left(\frac{\Delta t_i^{\omega} - u_i}{\sigma_i}\right) \to 1$. Therefore the ultimate number related MeSH term *i* equals to a remarkably value $h_i^{\infty} = \exp(\alpha_i)$. Else if $\omega < 0$ and $\Delta t_i \to \infty$, then, $\Delta t_i^{\omega} \to 0$, $h_i^{\infty} = \exp\left(\alpha_i \Phi\left(-\frac{u_i}{\sigma_i}\right)\right)$, indicating that the ultimate impact of a research topic is independent with ω.

3 Experiment

3.1 Results

For the purpose of this study, we selects the scientific data of medical field from 1946 to 2017. PubMed databases in the National Center for Biotechnology Information (NCBI) provide free publications data.

Moreover, PubMed comprises more than 28 million records for biomedical literature from MEDLINE, life science journals, and online books. Without loss of generality, we randomly choose fifty papers per year as long as 72 years utility data, granted between 1946 and 2017, downloaded from the NCBI website. 15 articles and their MeSH terms are randomly selected each year, such as ("Dinoflagellida" [MeSH Terms]). It finally provided us (on April 1, 2018) with a retrieval of 75 records.

To illustrate the universality of the model, we choose four data with completely different characteristics in Fig. 4. First, as illustrated, the red line, corresponding to the key word 'Pyrimidine Nucleotides', has grown rapidly in the early period and has also declined significantly in the later period. The blue line, corresponding to the key word 'Hexosyltransferases', has been steadily rising, only to decline sharply in the end. The entire life curve of the purple line is relatively stable, which corresponds to 'Leukemia Virus, Murine', initially maintaining a small increase in the early period, and then gradually decaying. The green line, corresponding to the key word 'Epidermal Growth Factor', is not the same as the above three. There was a long period of incubation in the early period, followed by a rapid increase and then a slow decline. They can basically summarize the developing characteristics of all various data. Simultaneously, the large amount of these four data further enhances their persuasiveness and representation. The Fig.4 Inset shows the cumulative amount of the four life curves.

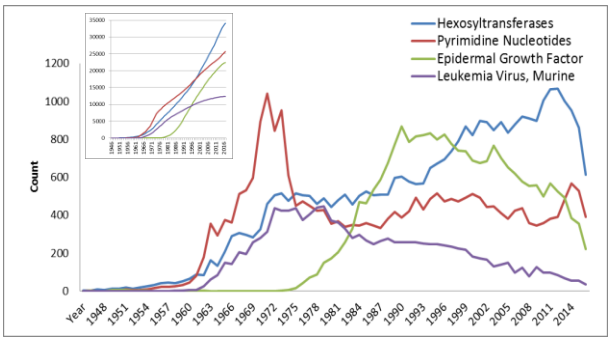

Fig. 4. Four life curves with typical characteristics.

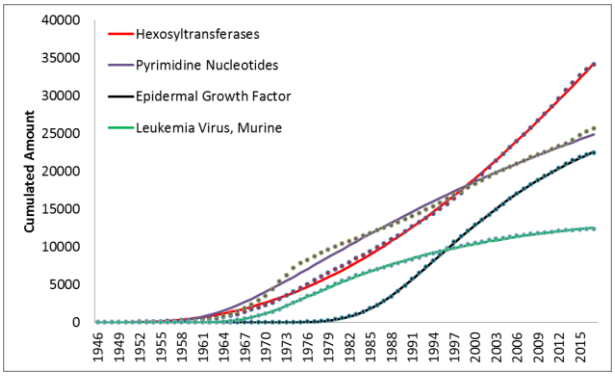

Fig. 5. The fitting results of model Eq. 5.

We have obtained a model suitable for this paper. The fitting results of this model we got are shown in Fig. 5. The solid line corresponds to our model fitted with the data. The dashed line denotes the life curve of the four real data, respectively. Obviously, the degree of fit between the solid and dashed lines is high, especially the blue line (corresponding to 'Epidermal Growth Factor') and the green line (corresponding to 'Leukemia Virus, Murine'), which are almost perfectly fitted. The second is the red line(corresponding to 'Hexosyltransferases').

The α and δ value of 'Hexosyltransferases' are highest for its continuously innovative ability, It has been rising from 1950 and taking a long time. 'Epidermal Growth Factor' has the smallest ω value, and its influence is mainly concentrated in the early stage since its appearance.

Table 1. The results of parameter estimates

	CASE1	CASE2	CASE3	CASE4	AVG
α	27.96	12.9	10.71	9.77	15.335
ω	-0.22	-0.76	-1.75	-0.51	-0.81
u	0.11	0.14	0.002	0.3	0.138
σ	0.41	0.13	0.001	0.1	0.16025

CASE1: Hexosyltransferases; CASE2: Pyrimidine Nucleotides; CASE3:Epidermal Growth Factor; CASE4: Leukemia Virus Murine

3.2 Comparison and analysis

The observed accuracy prompts us to compare with other models. We therefore identified several models that the others have been used in the past to fit histories the Model (Wang, 2013), the Logistic model and Gompertz model as shown in Table 2.

Table 2. Three models for comparison.

Model	Formula	Instruction
Model(Wang) 13 (Science, 2013)	$h_i^t = m(e^{\lambda_i \Phi(\frac{\ln t-\mu_i}{\sigma_i})}-1)$	m measures the average references each new paper contains. λ_i is the relative fitness, μ indicates the time for a paper to reach its citation peak, σ is longevity.
Logistic Model	$h_i^t = \frac{a}{1+be^{-cx}}$	a , b is two constants. c is longevity. x corresponds to immediacy of paper.
Gompertz Model	$h_i^t = ae^{-e^{(p+qt)}}$	a is a constant. p sets the displacement in h_i^t . q characterizes the growth rate of citations.

Correlation coefficient is the state or relation of being correlated. Specifically**,** a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone. In Fig. 6, the correlation is between the model fitting and the actual trends of data. That reflects the approximate degree of the model and the actual situation. This shows that the stronger the correlation is, indicating that the model is more accurate. From the Fig.6, we can clearly see that the correlation of our model, Model(Wang) and Gompertz Model is relatively close,

but our model is still better than the other two models. And, the effect of Logistic Model is not satisfactory.

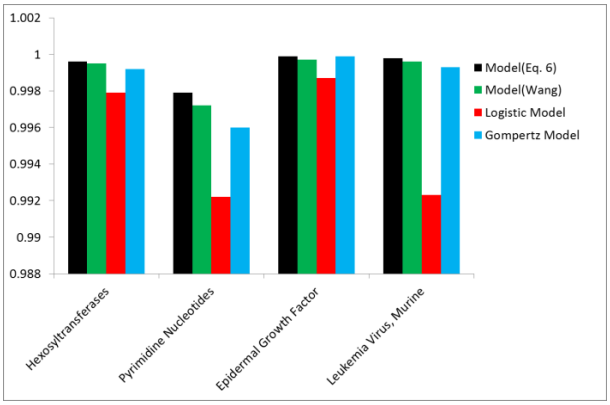

Fig. 6. Correlation coefficient among four cases.

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values. So the smaller the RMSD, the better. As illustrated from Fig.7, the RMSD among our model on the four samples (H,P,E&L) are obviously smaller than those of the other three models (Model(Wang), Logistic Mode and Gompertz Model). That means, our model clearly outperforms the other three. At the same time, we can also see that the second-optimal model is Wang's model. The effects of the remaining two models are slightly worse.

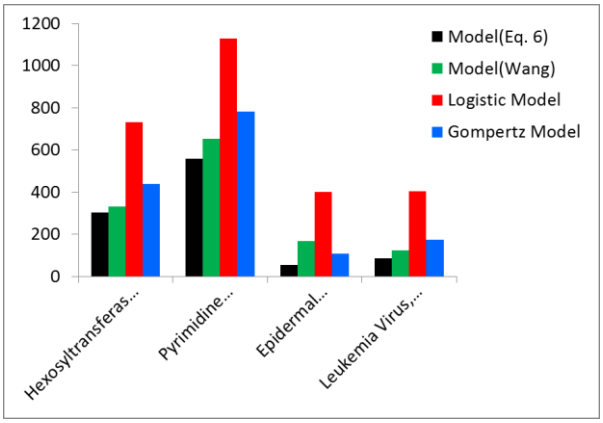

Fig. 7. RMSD among four cases.

4 Discussion

Under the premise of the same number of parameters, the fitting effect of our model is generally superior to others. It benefits from the understanding of the evolutionary mechanism of research topics and the restoration of information as possible. The innovation α, duration σ, and interval of the topic u will affect the ultimate development of itself. Therefore, the best choice for new researchers is the research point with both the larger α and σ, which means large research scale and continuous innovation.

And it is worth emphasizing that although ω does not affect the total number of final documents, it still is of great significance. ω can be a measure of the time efficiency of this topic being recognized. The bigger ω is, the shorter time it takes. When ω is more than 1, it can be thought that a number of researchers followed suit as soon as proposed. The smaller ω is, the more gradually it increases at first. When σ=0.2 and μ=0, the effect of ω is shown in the Fig. 8. Based on comprehensively analyzing of the literature amount of Moore's law, the influence of time affect is neither exponential decline nor left normal distribution.

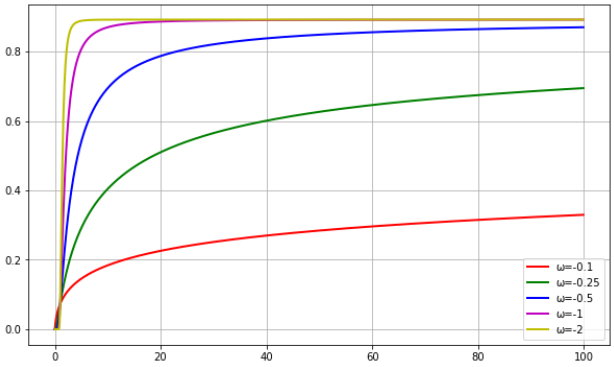

Fig. 8. The impact analysis of the parameter ω.

Citation-based measurement, which used to gauge impact, including analysis of impact factors to citation of short-term papers, always lacks predictability. In this paper, we propose a citation model with long-term predictability. Meanwhile due to little is known about the mechanism of time evolution of individual papers in the past research. Here, we derive a mechanistic model for the dynamics of citations in individual papers, allowing us to collapse the citation history of papers from different journals and disciplines into a single curve, indicating that all papers tend to follow the same universal time model.

The observed patterns help us to discover the basic mechanisms of management science impact. We can use this universal model to quantify progress potential of emerging research topic to predict the emergence of emerging technologies within a certain discipline. In combination with expert experience, we will select emerging technologies that will have the most development impact in the future. Through the government's vigorous policy support and social resources support, it will lead to the emergence of emerging technologies in advance. Therefore, it is doubtless to have tremendous boost to the development of the discipline, and even the entire history of science and technology.

The model proposed in this paper can fit the data of different characteristics well. In particular, it solves the problem of 'Sleeping Beauties' mentioned by Wang, whose reason is that the Model (Wang) limit the time to the early stages of publication with the Log function. The fundamental reason why logistic regression fails to fit

well is that it's characteristics of 'fast growth and fast disappearing'. Although our model is highly applicable, it is not completely universal, such as neural network, which experienced several degrees of sinking and developing. At the same time, we also observed that the early prediction is not good enough when the study was not followed for a long time. Therefore, it's better to analyze the data which has been accumulated for a period of time. This paper involves only the validation and analysis of the model, and uses only one type of data. Next step, the model will continue to be expanded in terms of data types. At the same time, we will focus on the predictability of the model to explore the future research.

References

1. M. Halaweh, "Emerging technology: What is it," Journal of technology management & innovation, vol.8, nov. 2013, pp. 108-115, doi:10.4067/S0718-27242013000400010.
2. D. Rotolo, D. Hicks, and B.R. Martin, "What is an emerging technology?" Research Policy, Dec.2015, vol. 44, p.1827-1843, doi:10.1016/j.respol.2015.06.006.
3. H. Noh, Y.K. Song, and S. Lee, "Identifying emerging core technologies for the future: Case study of patents published by leading telecommunication organizations , " Telecommunications Policy, vol. 40, Oct. 2016, pp. 956-970, doi:10.1016/j.telpol.2016.04.003.
4. H. Small, H. Tseng, and M. Patek, "Discovering discoveries: Identifying biomedical discoveries using citation contexts", Journal of Informetrics,vol. 11, Feb. 2017, p. 46-62, doi:10.1016/j.joi.2016.11.001.
5. J. Joung, and K. Kim, "Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data," Technological Forecasting & Social Change, Jan. 2017, vol. 114, pp. 281-292, doi:10.1016/j.techfore.2016.08.020.
6. Zhang Y, Zhang G, Chen H, et al. Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research[J]. Technological Forecasting & Social Change, April 2016, vol. 105, pp.179-19, doi:10.1016/j.techfore.2016.01.015 1.
7. S. Zhang, and F. Han, " Identifying emerging topics in a technological domain," Journal of Intelligent & Fuzzy Systems, vol. 31, Sep. 2016, pp.2147-2157, doi: 10.3233/JIFS-169054.
8. M. Medo, G. Cimini, and S. Gualdi, "Temporal effects in the growth of networks," Physical review letters, vol. 107, Dec 2011, pp. 956-970, doi: 10.1103/PhysRevLett.107.238701.
9. I. Park, K. Lee, and B. Yoon, "Exploring Promising Research Frontiers Based on Knowledge Maps in the Solar Cell Technology Field" Sustainability, vol. 7, Oct. 2015, pp. 13660-13689, doi:10.3390/su71013660.
10. D. Rotolo, I. Rafols, M.M. Hopkins, and L. Leydesdorff, "Strategic intelligence on emerging technologies: Scientometric overlay mapping," Journal of the Association for Information Science & Technology, vol. 68, Dec. 2015, pp. 214-233, doi:/10.1002/asi.23631.
11. J.L. Guo, X.Y. Zhu, Q. Suo,and J. Forrest , "Non-uniform evolving hypergraphs and weighted evolving hypergraphs," Scientific reports, Nov. 2016, doi:10.1038/srep36648.
12. Leydesdorff L, Comins J A, Sorensen A A, et al. "Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level", Scientometrics, vol.109, March 2016, pp. 2077-2091,doi: 10.1007/s11192-016-2119-7.
13. D. Wang, C. Song, and A.L. Barabási, "Quantifying long-term scientific impact," Science, vol. 342, Oct. 2013, pp.127-132, doi: 10.1126/science.1237825.
14. G. González-Alcaide, P. Llorente, and J.M. Ramos, "Bibliometric indicators to identify emerging research fields: publications on mass gatherings," Scientometrics, Nov. 2016,vol. 109, pp. 1283–1298, doi: 10.1007/s11192-016-2083-2.
15. Q. Wang, "A bibliometric model for identifying emerging research topics," Journal of the Association for Information Science and Technology, vol. 69, Feb. 2018, pp. 290-304, doi:10.1002/asi.23930.