Selecting Most Influential Topics In The Future

Jinghao Zhao, Hao Wu
Shanghai Jiao Tong University
Email: zhaojinghao@sjtu.edu.cn, albert-wuhao@sjtu.edu.cn

***Abstract*—Every year many new topics appear in the academia and industry, and many topics appeared years ago may be rediscovered because of new technology or new uses, so how to predict the trend of a certain topic and find the most potential topics in the future is an important work for the scientific research and IT industry. Using real scholarly datasets *Microsoft Academic Graph* [1] with more than 12000 topics, 14.4 million authors and 30 million papers from the computer science domain, we perform these tasks to solve this question. First, we draw the Topic Map which show the relationships between topics and use k-core analysis to insight the skeleton of the whole CS field. We also find topics having high growth rates are more likely to have strong relationship with other hot topics.Second, we proposal a series of inner and inter factors that can determine the state of a topic. We examine the co-evolving relations between each part of topics, especially the author factor's influences. Finally, we obtain the time series for recent 50 years of each factor of all topics, and predict the topics scale in five years with an R_2 value of 0.96. Then we use LSTM (Long Short-Term Memory) to improve the performance of our prediction. We forecast the top 100 fastest growing topics and largest 100 topics in the 5 years. Furthermore, we deploy our findings on-line to help the users to know the tendency of computer science domain and find the most potential research directions.**

I. Introduction

In recent years, many topics emerged, with many new technologies and new ideas. Some of the most popular directions are being constantly focused by academia and industry, and have shown great potential in many domains such as Deep Learning, AI and so on. So how to discover next potential technology that can be widely used is an important task. However, for the rapid change of topics, new topic appears, and old topic become out of date, so it is hard to find the formula to know the development of a certain domain and find the next hot point to prepare for the next technological revolution. Most of previous work concentrate on the social network. In this paper, we studied the characters of academic network, especially the relationship between different topics and get some interesting findings.

To know the next hot point, the first step is to seize the basic patterns or the structures of the whole academic field. If we catch the core of the field, we can better understand the foundation and know the structure directly. A k-core is the maximal subgraph where all vertices have degree at least k. In the academic network, k-core is usually employ on co-author network to clustering different authors, finding hierarchical structures and making visualization. In this paper, we use k-core to excavate the core of the computer science domain which mean the basic topics in this field. From these topics we can get the skeleton of the whole computer field, and have a more intuitive understanding of the application in the industry. In addition, the relationship between topics is also particularly important. If we know the relationship between different topics appeared in this field, we could get inspiration from the topics which has tight relationship with our research fields and find some new ways to overcome the obstacles in the research. In order to solve these problems, we propose TopicMap, which show the structure of the CS domain, interrelationship between topics, keywords clustering, and popular areas of the whole field.

Based on the empirical observation, the positive correlation always exists between the number of authors and papers in one topic, which means if more authors are active in one topic this year, in the meantime, the paper number this year or in the near future is always higher as well. And in this paper, what we are eager to find out is whether the number of authors with some special characteristics can predict the number of papers in the near future accurately, and on this purpose, we try to divide authors into four categories reasonably according to pareto principle with two dimensions, the number of papers and the number of citations, higher or lower. For the author is an important part of determine the state of a topic, this part provide insight to author structure and help the following prediction.

The scale of topics, which we determine it as the number of papers containing in the topic, is considered to be a dynamic because of the mix of parameters (Time FactorPaper Factor, Author Factor and etc.) A keen researcher can accurately find new directions for research, or innovations in existing ideas. Though it is very hard to replace the expertise that an experienced researcher has gained, good prediction results can make the research results more breakthrough. In this paper, we use the machine learning techniques which have been applied for time-series prediction to predict the growth rate of the number of paper in the topics.

First, we investigated the factors that might affect the future size of topics. According to the interdependency between these factors and target, we screened some of the most influential factors,which containing some unexpected results that have obvious impact on the growth of topics. Second, we obtain the time series of these factors for recent 50 years of more than 12000 topics, which is a very miscellaneous work in terms of data integrity and data volume. Then we

take the number of papers contained in topic, which means the topics' size, as the target value of the forecast, and use the time series of each factor to predict the future trend of this topic. We show the predict result of different models including Linear Regression, Support Vector Regression, Radial Basis Function, Decision Tree, Random Forest, Extremely Randomized Trees, Gradient Boosted Regression Trees, Bagged Decision Trees. We also introduce the parameters and variables that can be used in order to predict the size of topics which can be helpful in the future prediction.

(ADD LSTM PART.)

II. Related Work

Because of the time series of all the features of an academic topic is a huge and complicated job and the datasets which contain papers topic information are not very common, there is very little work about the prediction of academic topics heat. However, there are some related tasks that can be divided into the following areas: k-core analysis of scholarly network, topic influential factors, and time series prediction.

A. K-core Analysis of Scholarly Network

In the field of k-core analysis of scholarly network, existing researchers always focus on the coauthor relationship and use k-core to do the community detection to the papers. The k-core algorithm can extract a group of authors who have cooperative relationship, or the citation relationship between papers so that cluster the papers to different topics. For instance, Qian $et\ al.$ [2] studied the k-core relation of papers in one topic and the core-groups life circle. Emre $et\ al.$ [3] proposed how centrality in the coauthorship network differs between high impact authors and low impact authors and deploy a classifier to predict the papers citation.

We can see that all the research targets are single items in the scholarly network. In contrast, we extend k-core analysis to topic analysis, and the get the basic structure of the academic domain which support the development of the entire Computer Science field. From higher level of combinations of lots of single items, we can build the whole hierarchy structure of this field.

B. Pareto Principle for Author Analysis

Based on pareto distribution and power distribution, Pareto principle states that for many events, roughly 80% of the efforts come from 20% of the cases, is always used in many field like software engineering applied to optimization efforts, Ankunda [4], or education, Hctor $et\ al.$ [5]. It helps a lot to realize that often a minority of inputs can cause the majority of results.

Based on pareto distribution and power distribution, Pareto principle states that for many events, roughly 80% of the efforts come from 20% of the cases, is always used in many field like software engineering applied to optimization efforts, Ankunda [4], or education, Hctor $et\ al.$ [5]. It helps a lot to realize that often a minority of inputs can cause the majority of results.

Matthew $et\ al.$ [6] gets the conclusion that the citation number of authors obeys the power laws, and only small part of them published papers with high quality, in other words, with high citations. So it's reasonable to implement Pareto principle to identify who is the author with high citation number and who with low citation number. Meanwhile, we will take this method to recognize authors with high or low paper number.

C. Time Series Prediction

To the scholarly network, lots of work has been focus on the prediction of the impact of one paper or one author. For example, J. Gehrke $et\ al.$ [7] focus on how to predict the future citation number of a paper according to its present citation. Xiao $et\ al.$ [8] proposed a model to predict the individual paper citation count over time.Some work focus on the authors, such as Dong $et\ al.$ [9] examine the authors h-index in five years and propose a classifier to distinguish whether a previously (newly) published paper will contribute to the authors' future h-index. They all pay attention to small items of the network and lack of overview of the whole structure of the field. In our work, we propose unique features to describe the development of topic and use these features to predict the future trend of the topic itself.

There are some other work about social network, such as Saha $et\ al.$ [10].They detect emerging topics in the social network and track the topics' evolving process. However, the social network have many differences from the scholarly network. Topics that appear in the social media are more Transient, and users of the hot topic only gather together for a short time, then the topic may disappear. The scholarly topics have much longer life circle and the inner structures of topics are more stable. Contrast to gathering by users' curiosity, the academic topic have more stable relationship between individuals in it such as stable reference relation and coauthor relationship. So the properties of academic network are different. We get the time series of different factors which can describe a topic and use these factors to predict the topics evolving process.

III. Problem Formulation

We want to get the big patterns of computer science domain and predict the development of this field.In this section, we formally define our work into three tasks as follows:

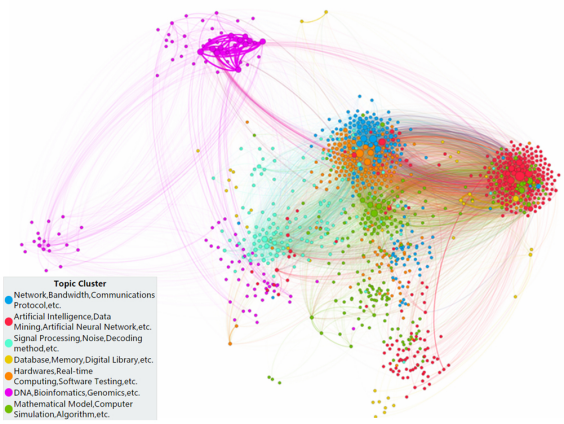

Fig. 1. Topic Map

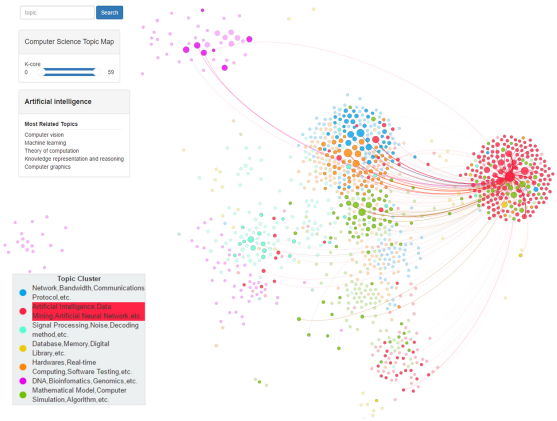

Fig. 2. Click Effect of Topic Map

Task 1: Topic Map. The goal is to visualize the basic structure of the cs field and show the relationship between different topics. Including topic clustering, k-core analysis, and heat representing.

Task 2: Topic Factor Extraction. The goal is to examine factors that can influence the future development or the factor that can show the present state of a topic.Including the co-evolution relation between the different factors and how these factors influence the growth rate of the topic.

Task 3: Topic Scale Prediction. The goal is to regard the scale prediction as a regression problem. Given the factor matrix M of topic T at time t, the problem is to predict the paper number N, which means the size and scale of this topic, at the time $t + \Delta t$.

IV. Topic Map

In this section, we will describe the detail of Topic Map, including the data and we used, and the information we can get from this graph.

The data we used come from the *Microsoft Academic Graph* [1]. It provide the information of more than 30 million papers, 14.4 million authors and their citation relationship. Unlike common datasets, this dataset provides topic information for each paper and the topic hierarchy structure. Because of the topic information is hard to get, other researchers can only get the topic info through a number of ways, such as latent Dirichlet allocation (LDA) [11]and other text cluster or topic discovery algorithm. There is a problem that the topics extracted from the document text may not very meaningful and it's hard to deploy on millions of paper. It's may be a reason why related researches to our work are vary rare.

The hierarchy of the topics contains 3 level in the dataset, $L0$, $L1$,$L2$ and $L3$. $L0$ level represent the basic domain of the whole academia, such as *Computer Science, Mathematics, Biology etc.* We choose computer science as our research object. The $L0$ topic contains $L1$ topics. To the computer science, $L1$ levels contain some basic topic of cs field, such as *Network etc.* The $L2$ and $L3$ topics are not totally parents and children relationship, for some $L3$ topics are directly belong to the $L1$ topic, but most of the $L2$ topics are bigger than the $L3$ topics.$L2$ topics contain some big topics such as *Data Mining, Machine Learning etc.* and $L3$ topics are more specific academia areas such as *5G, Topic Model etc.*

In the Fig. 1, we draw the top 10% large topics (according to its paper number) of $L2$ and $L3$ topics. Each circle represents a topic and the radius of the circle represent the paper number in this topic. If two topics contain a same paper, an edge will form between two topics, and the weight of this edge means the total number of papers shared by two topics. For there are too many edges if we don't set a weight threshold, we filter the edges whose weight is less than 500, and the left edges mean a strong relation between the topics.

We cluster the topics according their similarity and paint the different themes with the different color. We can see that the topics form communities and some core topics are at the centers of each community. We select the top size topics of each community to represent its theme. Just like the Fig. 2, When we click one topic on the TopicMap, the select part and the related topics with the selected one will be highlighted. The name of selected topic,and the top 5 topics which have closest relation with the selected topic will also be listed at the left. We can also search the topic that we are interested in th search box which will show the topic's position in the TopicMap. From this graph we can determine the relationship between each topic directly and clearly.

To know the most basic structure of computer science domain, we use the k-core analysis to extract the skeleton. A k-core is the maximal subgraph where all vertices have degree at least k. In each cycle, the vertices whose degree

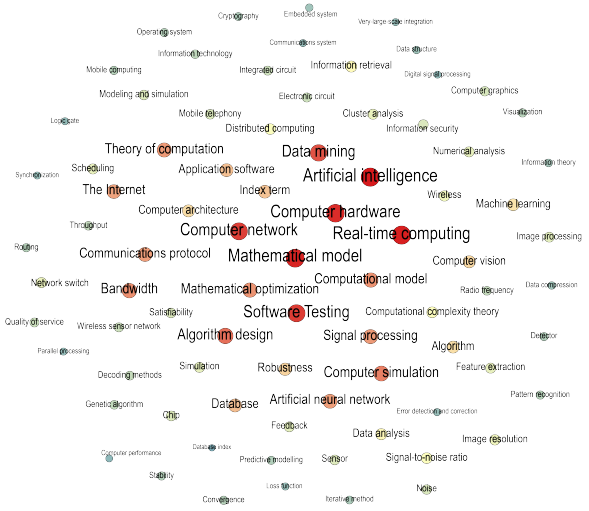

Fig. 3. Core Topics of Computer Science

is less than threshold k will be deleted from the graph, and edges connected to this vertex will be deleted at the same time. After many cycles filtering, the graph will be stable and doesn't change any more. The number of topics after k-core processing changes with the value of threshold k. We can adjust the slider in the 2 to filter out the topics in a specific k-core range. The max value of k we can set to our Topic Map is 58, which means if k is larger than 58, all vertices will be filtered. The subgraph after processing is as shown in the Fig. 3. We can see that these topics are the foundations to support the entire CS field.

(core-ratio graph)

V. Topic Factors

The development of academic topics is influenced by many factors, and there is a mutual influence between these factors.In these section, we focus on what factor can influence the future trend of a topic and how these factors effect on each other. Furthermore, we want to determine how to accurately describe the present state of a topic, and make insight of the different topics.

A. Paper Factor

To predict the future trend of a topic, paper is the essential factor. The formal paper make the base of its topic and attract more attention from researchers to focus on this topic, and the new paper attracted by the old paper give this topic more impact. This forms a circle to make the large and important topics have higher growth rates than the smaller topics.The number of paper in this topic is the fundamental elements of the topic.We define *paper-num* to represent the number of papers in this topic at a certain year. Papers' citation information can also represent a topic's impact. One highly cited paper may lay a foundation to the related topic and open a new era of technology wave, so we

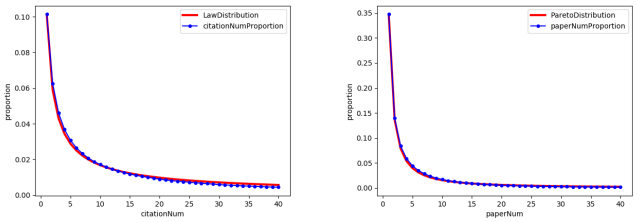

Fig. 4. Real Distribution of authorCitation-num proportion and paper-num proportion based on the total information of authors and papers from 1900 to 2016 in whole CS field

define *citation-max* of a topic as the max citation number of papers in this topic. For the similar reason, if papers' average citation in one topic is higher than another topic, this topic will obviously obtain more attention, which will increase the size of this topic in the future, so we define *citation-ave* to calculate the average citations of all the papers in this topic.

B. Author Factor

The relationship between the number of author and the development of academic topics can not be ignored.Owing to the increasing participation of authors who focus on this topic, this topic could be more intriguing and be more influential in near future.And based on our assumption, for example, if some authors with higher citation number and paper number, who is also more insightful in our opinions, take part in one topic, it means they admit the potential or the importance of this topic. Therefore, in this part, we mainly focus on the real distribution of *authorCitation-num* as citation number and *authorPaper-num* as paper number of each author and define four kinds of author due to the distribution, *PHCH*(with high paper number and high citation number), *PHCL*(with high paper number and low citation number), *PLCH*(with low paper number and high citation number), *PLCL*(with low paper number and low citation number).

In this part, we utilize the dataset about information of all papers and authors in CS field provided by *Microsoft Academic Graph* [1]. For obtaining the break point of authorCitation-num and authorPaper-num, we take three steps as follows, drawing the real distribution of both two indexes, looking for the probability density function function, using pareto principle to get two break points.

Step 1: Real distribution As Fig. 4 shows, we draw two graphs, including the real proportion of authorCitation-num with citation number increasing and standard pareto distribution, and the real proportion of authorPaper-num with paper number increasing and modified pareto distribution based on the total citation-num and author-num from 1900 to 2016.

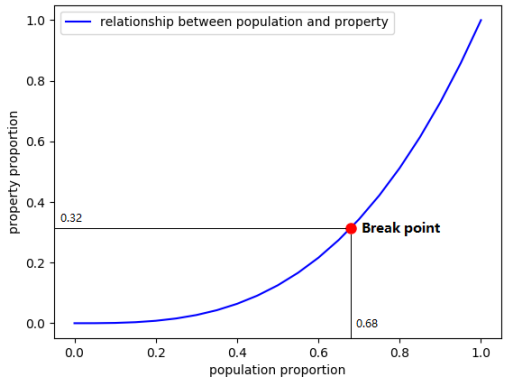

Fig. 5. Relationship between the proportion of population and property

Step 2: Probability density function The standard pareto distribution is to describe the distribution of a random variable, the probability that X is greater than some number x is given by

$$\overline{F}(x) = Pr(X > x) = \begin{cases} (\frac{x_m}{x})^{\alpha} & x \geq x_m \\ 1 & x < x_m \end{cases}$$

where x_m is the minimum possible value of X, and α is a positive parameter called pareto index. So the probability density function of X followed is

$$f_X(x) = \begin{cases} \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} & x \geq x_m \\ 0 & x < x_m \end{cases}$$

As Fig. 4 presents, we get the conclusion that probability density function of paper-num obeys the pareto distribution with pareto index α_{paper} = 0.347. And for the probability density function of citation-num, it obeys the law distribution with attenuation coefficient $\beta_{citation} = 0.782$

$$f_X(x) = \begin{cases} (\frac{x_m}{x})^{\beta} & x \geq x_m \\ 0 & x < x_m \end{cases}$$

where x_m is the minimum possible value of X, and β is a positive parameter called attenuation coefficient.

Step 3: Break point Owing to the distribution of citation-num and paper-num obeying standard pareto distribution and law distribution, the pareto principle can be utilized to calculate break point.The pareto principle is to describe a phenomenon that for many events, most of the effects come from little of the causes. For example, as Fig. 5 presents, let's define function f(x) as the proportion of total property and x as the proportion of total population form the poor to the rich, and the break point x_{point} satisfies that

$$x_{point} + f(x_{point}) = 1$$

After these three steps, as Fig. 6 denotes, we calculate the two break points for each year from 1950 to 2016.

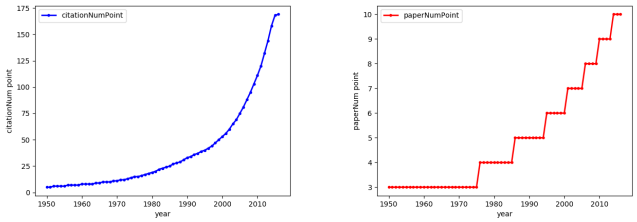

Fig. 6. The break points of citaitonNum and paperNum for each year from 1950 to 2016

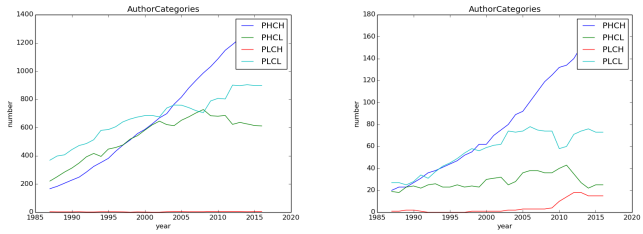

Fig. 7. The development f four types of authors about topic *Coupling coefficient of resonators* and topic *HAMP domain*

Based on this result, we classify the authors into four types, PHCH, PHCL, PLCH and PLCL, and count the number for each type in each year, which takes us to the conclusion that the number of PHCH is increasing steadily, the number of PHCL and PLCH is increasing with complementary trend, which means when one is increasing, the other is declining relatively. We plot two graphs about the development of four types of authors in Fig. 7 to illustrate the phenomenon more persuasively and clearly.

C. Growth Factor

The growth trends of an academic topics are not as volatile as stocks, so the growth rate in the past several years may effect the future trend. We define the *increase-num* to represent the growth of paper number in a topic between current year and last year. We also calculate the average growth of the past 5 years as *increase-num-ave* to show the growth constancy of this topic. Furthermore, if a topic suddenly get a lot of attention for the new theory come out in this field, obviously the topic will grow very fast in the following years.For this reason, we calculate the max value of the growth of past 5 years in this topic and define it as *increase-num-max*.

D. Venue Factor

Most of papers are published on the various journals, conferences and so on, so the venue information is a very important factor of a topic.. First, for each topic, we obtain the venues that appears in this topic, which means that at least one paper in this topic appeared in these venues before. We

TABLE I
TOPIC FACTORS AND CORRELATION COEFFICIENTS BETWEEN THIS ELEMENT AND TOPIC SCALE AFTER t YEARS

Factor	Element	Definition	cc_1	cc_5	cc_{10}
Paper	*paper-num*	The number of papers in this topic	0.9927	0.9861	0.9570
	citation-ave	The average value of papers' citations in this topic	-0.0103	-0.0007	-0.0029
	citation-max	The max value of papers' citations in this topic	0.3368	0.3413	0.3373
Author	*author-num*	The number of authors in this topic	0.9618	0.9532	0.9371
	author-hindex-ave	The average value of authors' h-index in this topic	0.0688	0.0629	0.0637
	author-hindex-max	The max value of authors' h-index in this topic	0.3580	0.3691	0.3811
	author-hindex-var	The variance of authors' h-index in this topic	0.0542	0.0486	0.0500
Growth	*increase-num*	The growth of paper number between current year and last year	0.8885	0.9432	0.9438
	increase-num-ave	The average value of growth number in the past five years	0.9487	0.9586	0.9558
	increase-num-max	The max value of growth number in the past five years	0.9381	0.9385	0.9294
Venue	*venue-num*	The total number of venues in this topi	0.7054	0.6767	0.6511
	venue-distinct-num	The number of distinctive venues in this topic	0.5669	0.5616	0.5550
	venue-index-ave	The weighted average of the *venueIndex* of venues appeared in this topic.	0.0123	0.0280	0.0528
Interaction	*interaction-growthnum-ave*	The average value of *increase-num* of neighboring topics	0.0291	0.0356	0.0290
	interaction-growthnum-ave	The max value of *increase-num* of neighboring topics	0.0331	0.0381	0.0290

define *venue-num* to represent the total number of venues in this topic.Among these venues, some venues are in the same series of conferences or journals.For example, ICDM2015, ICDM2016 and so on are all in ICDM series. So we remove the duplicate venues belonging to the same series, and we get the *venue-distinct-num* as the number of distinctive venues in this topic to show the diversity of this topic. The more different venues appear in this topic, the more wide-ranging this topic may be, and it may get more attention in the future.

Generally speaking, the influence of the paper is proportional to the influence of the conference or the journal. We want to distinguish the impact of papers from what venues they appear, so the first task is to determine the impact of a certain conferences or journal. To measure the impact of a venue V, we quantify the impact as *venueIndex* by

$$venueIndex(V) = \frac{\sum_{p \in V} citations(p)}{N_V} \tag{1}$$

where N_V is the total number of papers in this venue and $citations(p)$ is the citation number of paper p. One venue's *venueIndex* shows the average citations of all papers in this venue and show the impact of this venue at the same time. Then we define the *venue-index-ave* of topic T as following:

$$\textit{venue-index-ave(T)} = \frac{\sum_{p \in T, V} venueIndex(V)}{N_T} \tag{2}$$

where N_T is the total number of papers in this topic, and the *venue-index-ave* of topic T is the weighted average of the venueIndex of venues appeared in this topic.This factor can help us to quantify the venues' impact to a topic and help us to predict the topic's future trend.

E. Interaction Factor

From the Figure(...), we can see the topics' relationship can influence the topic a lot.The high growth rate topics are more likely to have closer relationship with each other and low rate topics also form communities in the Figure(...). So the interaction is also an important part of factors to show the topics' future trend. We have got the weights of edges between any two topics which equals to the number of common papers these two topics containing. After screening out the top 5 topics which have closest relationships with the present topic, we regard these topics as neighboring topics of the present one.Then we define the *interaction-growthnum-ave*, which is the average of *increase-num* of neighboring topics. The *interaction-growthnum-ave* can represent the community's growth rate and show the growth potential of this small community. Furthermore, if a topics in the neighboring topics suddenly get much attention, this effect may radiate to the present topic and make it get attention, too. So we calculate the max *increase-num* of the neighboring topics as *interaction-growthnum-max* and this element can quantify the driving effect of a neighboring hot topic.

F. Time Serialization

These factors we mentioned above are shown in Table I. We want to know the development of each topic, and the topic's states are describe by the factors we proposal, so we do the time serialization to each factor of 12464 topics from 1950 to 2015.First, we extract the paper list of each topic and split the papers into many parts by their published years. When we calculate each factor at time t, we use the subset of papers published earlier than t. Finally we get the 12464 topics' information of each year. We also calculate the correlation coefficients of each element and the size of topics after t years.The results are listed at Table I.

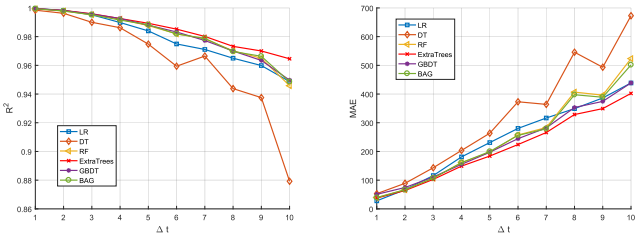

Fig. 8. (a)Predictive performance (R^2) (b)Predictive performance (MAE)

VI. Experiment

In the previous section, we examine the different factors' effect on the future topic trend. The results show that these factors can be the features to help us predict the future trend. In this section, we employ a series of models to predict the scale of topics in the future.First, we use some traditional models and some ensemble models. Then we use the LSTM to improve our prediction and predict the top 100 hot topics in 5 years.

A. Predicting topics with ensemble model

We use several models to predict the topic size in the future, including linear regression(LR), Decision Tree Regression(DT), Random Forest Regression(RF), Extremely Randomized Trees Regression(ExtraTrees), Gradient Boosting Regression(GBDT) and bagged decision trees(BAG). To evaluate the prediction accuracy, we compare these models by the coefficient of determination (R^2) and the mean absolute error (MAE). Fig. 8 shows the performance of difference models in terms of R^2 and MAE. We can see that for all the models, R^2 will decrease as the prediction gap become larger, and the MAE will increase at the same time, which both mean that our prediction have better performance in the shorter time interval.From the Fig. 8 we can obtain that the Extremely Randomized Trees Regression(ExtraTrees) shows the best performance among these models and achieve the R^2 of 0.9893 when $\Delta t = 5$ and 0.9646 when $\Delta t = 10$. The detailed performance are showed in Table II.

From the Fig. 9 we can have an intuitive perception of the results of the forecast. We choose the ExtraTrees, which has the best performance among the models, to predict a topic's size after different years. The x axis represent the true values of samples in the test dataset and the y axis represent the predicted values. The red line denote $y = x$ which means the predicted results fit perfectly with the true values. Each point represents a test sample. We can see that the accuracy will be higher with the less forecast interval and our prediction perform well both on the big topics and small topics.

B. Factor Importance Comparison

C. Improve the performance

VII. Conclusion

In this paper, we present how to select most influential topics in the future. First we draw the Topic Map which show the relationships between topics. We also find topics having high growth rates are more likely to have strong relationship with other hot topics, which help us to screen out the potential influential topic from an overall level. We also quantify this inter effect and use it to improve prediction accuracy in the following part.Then we proposal a series of inner and inter factors that can determine the state of a topic. These factors are closely related to the future trend of topics. In addition, we examine the co-evolving relations between each part of topics, especially the author factor's influences. Finally, we obtain the time series for recent 50 years of each factor of all topics, and predict the topics scale in five years with an R_2 value of 0.96.Then we use LSTM (Long Short-Term Memory) to improve the performance of our prediction. We forecast the top 100 fastest growing topics in the 5 years. Furthermore, we deploy our findings on-line to help the users to know the tendency of computer science domain and find the most potential research directions.

References

[1] (2016) Microsoft academic graph. [Online]. Available: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[2] T. Qian, Q. Li, B. Liu, H. Xiong, J. Srivastava, and P. C. Y. Sheu, "Topic formation and development: a core-group evolving process," *World Wide Web*, vol. 17, no. 6, pp. 1343–1373, 2014.

[3] E. Sarigol, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.

[4] A. R. Kiremire, "The application of the pareto principle in software engineering," *Consulted January*, vol. 13, p. 2016, 2011.

[5] T. S. A. Arias, S. E. M. Martínez, V. Hernandez, J. S. Guerrero, A. Reinoso *et al.*, "A methodology for identifying attributes of academic excellence based on a 20/80 pareto distribution," in *Global Engineering Education Conference (EDUCON), 2016 IEEE*. IEEE, 2016, pp. 1207–1211.

[6] M. L. Wallace, V. Larivière, and Y. Gingras, "Modeling a century of citation distributions," *Journal of Informetrics*, vol. 3, no. 4, pp. 296–303, 2009.

[7] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 kdd cup," *Sigkdd Explorations*, vol. 5, no. 2, pp. 149–151, 2003.

[8] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zha, "On modeling and predicting individual paper citation count over time," pp. 2676–2682, 2016.

[9] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[10] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization," pp. 693–702, 2012.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

TABLE II
COMPARISON BETWEEN MODELS PREDICTING TOPIC SCALE AFTER Δt YEARS

Criteria	Models	$\Delta t = 1$	$\Delta t = 2$	$\Delta t = 3$	$\Delta t = 4$	$\Delta t = 5$	$\Delta t = 6$	$\Delta t = 7$	$\Delta t = 8$	$\Delta t = 9$	$\Delta t = 10$
R^2	LR	0.9997	0.9983	0.9952	0.9898	0.9840	0.9749	0.9711	0.9650	0.9599	0.9487
	DT	0.9984	0.9963	0.9900	0.9862	0.9748	0.9594	0.9665	0.9438	0.9375	0.8792
	RF	0.9994	0.9979	0.9953	0.9921	0.9877	0.9820	0.9784	0.9701	0.9650	0.9458
	ExtraTrees	0.9995	0.9983	0.9960	0.9926	0.9893	0.9852	0.9802	0.9732	0.9700	0.9646
	GBDT	0.9995	0.9982	0.9957	0.9926	0.9880	0.9832	0.9773	0.9700	0.9634	0.9497
	BAG	0.9994	0.9979	0.9952	0.9915	0.9882	0.9819	0.9788	0.9696	0.9663	0.9485
MAE	LR	27.97	66.91	116.03	181.06	230.97	280.29	316.54	349.20	386.38	438.62
	DT	52.90	88.99	143.31	203.51	263.53	373.09	364.01	545.74	493.30	672.88
	RF	39.40	66.56	106.94	160.98	197.03	257.41	283.27	406.80	396.07	523.54
	ExtraTrees	37.41	63.89	102.67	148.97	183.98	224.09	265.97	328.66	349.28	402.29
	GBDT	50.69	74.41	111.07	155.07	197.39	244.56	282.22	353.16	374.38	438.84
	BAG	39.17	66.40	107.03	160.99	199.60	256.13	278.73	397.61	388.92	501.52

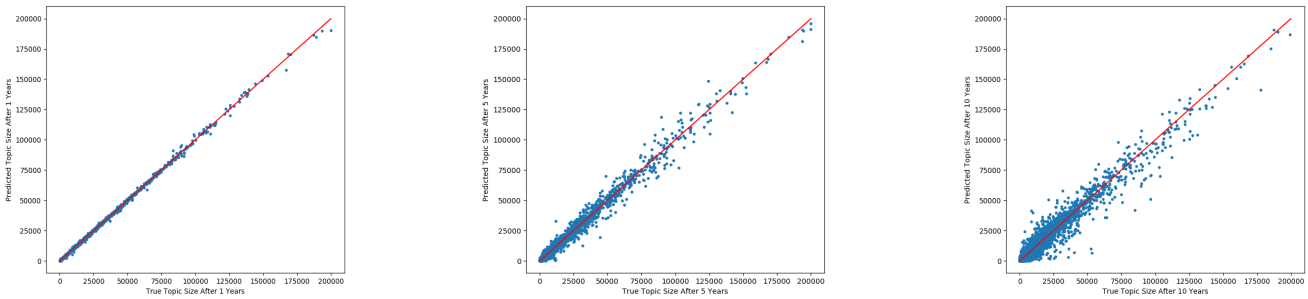

Fig. 9. Comparison of true value and predicted value