Prediction of Emerging Papers in Nanocarbon Materials-related Research Using a Citation Network

Hajime Sasaki[1], Tadayoshi Hara[2], Ichiro Sakata[1,2]
[1]Policy Alternatives Research Institute, The University of Tokyo, Tokyo, Japan
[2]Innovation Policy Research Center, Institute of Engineering Innovation,
School of Engineering, The University of Tokyo, Tokyo, Japan

Abstract--**Nanocarbon materials made from graphite are used in diverse applications as semiconductors, fuel cells, optical devices, and structural materials because of their excellent mechanical, electrical, and thermal characteristics. Numerous papers are published annually in this area, and thus it is difficult to assess overall development in the field. Consequently, there is a need for approaches that predict advances from diverse and numerous sources of information. In this study, we used machine learning to examine papers on nanocarbon materials and related topics and to predict papers with emerging ideas that are expected to grow in popularity. We specifically predicted emerging papers that were ranked in the top 5% by number of citations. A total of 411,084 related papers were extracted from the Web of Science Core Collection (Thomson Reuters). A time-expanded network was produced from these data using citation links, and features of each paper were used as explanatory variables to build a prediction model. In this model, 9 of the top 10 papers from 2011 predicted to be emerging satisfied the conditions for emerging papers. These results suggest that the model can predict the direction of nanocarbon materials technology, which is of considerable value for private companies and research institutions.**

I. INTRODUCTION

Nanocarbon material is a general term referring to material made from graphite, carbon nanotubes, graphene, and fullerene. These materials have appeared in succession over a short time, despite the long history of carbon. In 1985, Kroto et al. discovered fullerene, the C60 molecule [1]; in 1991, Iijima discovered nanotubes; and in 2004, Novoselov et al. discovered graphene [2]. These are 0-, 1-, and 2-dimensional nanocarbon materials with characteristic features and many potential applications. Nanocarbon materials are used in diverse applications as semiconductors, fuel cells, optical devices, and structural materials because of their excellent mechanical, electrical, and thermal characteristics, and these materials may be useful in the energy sector [3-7] or as space elevators [8-10].

In this field, the number of related research papers is increasing rapidly, making it difficult to grasp the current landscape and have foresight, especially when restricted by time and resource constraints. One method that is often performed in technology prediction, as represented by the Delphi method [11], is to select experts with detailed knowledge in specific science and technology fields, and ask for their opinions. These methods use interviews and questionnaires answered by experts to gain new insights into the future direction of research and technology. Ever since the Delphi method was proposed in the 1950's, it has been widely utilized to predict trends in science and technology. However, due to quantitative and structural changes in academic knowledge, a number of issues have been raised regarding this method in recent years, including that there are always subsections within a field. Thus, the results can change depending on the selected experts. In addition, there has been a marked increase in the volume of information, and a single person can no longer comprehend everything in a research field.

Increasingly, governments and researchers are demanding procedures through which they can identify advances in emerging research fields using the huge amount of available information. The amount of data and increasingly finely segmented research fields require development of innovative data-oriented techniques. Knowledge on nanocarbon materials technology has expanded, due to the broad range of potential applications in diverse areas, and this has resulted in a complex knowledge structure. In addition, in technology fields with strong science links, such as nanocarbon materials, academic papers are important as information for future technology trends.

In this report, we propose a method that uses machine learning to predict emerging papers in specific areas of nanocarbon research. An emerging paper is defined as one that might develop into an important area of research, but was not in the spotlight at the time of publication. Several procedures have been proposed to identify emerging research that might eventually produce substantial progress in a field.

Previous research

Research related to the prediction of emerging papers has been performed in bibliometrics and library and information science. With increasing awareness of big data in recent years, similar research is being performed in computer science sub-fields, particularly in data mining and information searching, which use research metadata on a large-scale. Winnink & Tijssen examined prediction of emerging research based on bibliographic information for Nobel Prize papers in fields related to graphene research [12]. Goffman & Newill are well-known for their comparison of information propagation to propagation of infectious diseases [13], and Bettencourt et al. used the SIR model, an infectious disease propagation model, to describe the propagation of newly appearing fields in existing fields [14]. Chen et al. used a co-citation network for academic papers and a joint research network to link innovative discoveries to stimulate research that filled structural holes in networks [15]. Young classified technology growth curves into nine types and performed a test to determine a growth curve model that can be fit to multiple datasets and perform the best predictions [16].

While emerging papers, by definition, do not receive much attention immediately after publication, they do contain great possibilities for the future. Adams showed a correlation between citation numbers from one to two years and three to ten years after publication of a paper in the life and physical sciences [17]. Li & Tong [18] formulated an optimization problem that predicted paper citation numbers using 500,000 papers from computer science, and predicted the number of citations they would receive after 10 years based on information three years after publication. In this research, citations in the three years after publication were shown to influence the number of citations after ten years. Based on analysis of two million computer science papers, Dong et al. [19] showed that the author's h-index five years after publication of a paper could be predicted, with the impact of the paper defined by six factors: author, joint authorship, content, publisher, citation count, and time lineage. Davletov et al. [20] predicted citation numbers five and ten years after publication, using time series information on the citation number several years after publication and information on the citation network structure, using 27,000 arXiv energy physics papers, 1.5 million computer science papers (ArnetMiner), and 2 million additional papers (CiteSeerX). The time series of the citation number in the two years after publication was found to be important for prediction [20]. Chakraborty et al. [21] showed that the number of citations after five years can be predicted from citation numbers several years after publication, together with data for the author, academic association, and keywords, based on 1.5 million computer science papers. The number of papers by the authors and the citation number one year after publication were found to be important to predict the future influence of the paper [21]. Wang et al. [22] used the power law for a paper's citation number, and created a formulation predicting the future citation number from time series information for the citation number five years after publication. Papers from Physical Review B, PNAS, and Cell were used to predict the citation number for a paper 25 years later, with a prediction accuracy of 90% [22].

These studies of prediction of emerging papers, and particularly prediction of the citation number or impact, have mainly used predictions based on time series of citation numbers several years after publication. This indicates that citation trends several years after publication are a good indicator of citation numbers after that period, but also shows that observation over many years is required for prediction. This limits early detection, but is useful for research that is focused on several years after publication. In contrast, Rogers defined “early adopters” or an “early majority” in “diffusion of innovation” as papers with information that immediately induced action in their fields [23]. To become an “innovator”, judgment is needed immediately after publication. In this report, we describe a model to predict papers with a high probability of increased citation in the near future (three years after publication) solely from information found immediately after publication (less than one year). We used this method to predict emerging papers in the nanocarbon materials field.

II. METHODOLOGY

The methodology is divided into four categories: data acquisition, construction of features, construction of the prediction model, and evaluation of the prediction model, as described below.

A. Data acquisition

Papers incorporating "nano*” and “carbon*" in the paper title, abstract, or keywords were extracted from the Thomson Web of Science Core Collection. We targeted journal papers and did not include international conference proceedings. Papers were extracted based on target fields, including title, abstract, author name, year of publication, and citation-related information, over the period from January 1901 to November 2015. We also collected information on all citation data on papers not indexed in the Web of Science, via API. This information is handled as correct data in construction of prediction models.

B. Construction of features

From the extracted data, we created a citation network for each year up to the present year, with all papers as nodes. From the created time expanded network, we extracted features for the following classes in each paper in each year. Here, the constructed features are used to express learning data for prediction of emerging papers. The constructed features used in the prediction model can be divided broadly into four classes: network features, cluster features, centrality features, and citation-related features. A summary of the respective features is shown in Table 1. The network feature expresses the general features of the citation network in which the papers are incorporated. The cluster feature expresses the cluster features in which the papers are incorporated. A cluster is a group of papers that are cited within a specific region of the citation network. We constructed a community extraction algorithm that has been modularity-maximized in relation to the network [24]. The centrality feature expresses the degree of centrality of papers in the citation network structure, and the degree of centrality can be quantified at multiple viewpoints [25-31]. The citation-related features set the statistical summary values (maximum, minimum, average, and total) of the papers cited by the targeted paper. A total of 63 features were used. All features were incorporated into the maximum component of the time expanded network created each year.

TABLE 1. FEATURE VALUES FOR PREDICTION OF EMERGING PAPERS

Classification of features	Name of feature	Description	Ref.
Network		Target data set, target fiscal year network feature	
	NW_NODES	Number of papers incorporated into network	
	NW_EDGES	Number of citation links incorporated into network	
	NW_MAXQ	Maximum value of cluster Q value incorporated into network	[24]
Cluster		Feature value resident in cluster where target paper resides	
	CL_QMAX	Maximum value of Q value in cluster where target paper resides	[24]
	CL_NODES	Node number of cluster where target paper resides	
	CL_RANK	Ranking of cluster where target paper resides	
Centrality		Network centrality of target paper	
	CNT_DEGRE	Degree centrality	[25]
	CNT_BETWE	Betweenness centrality	[26]
	CNT_CLOSE	Closeness centrality	[25]
	CNT_EIGEN	Eigenvector centrality	[27]
	CNT_NETWO	Network constraint	[28]
	CNT_CLUST	Clustering coefficient	[29]
	CNT_PAGER	PageRank	[30]
	CNT_HUBSC	Hub Score	[31]
	CNT_AUTHOR	Authority Score	[31]
Citation		For papers cited by the targeted paper, each feature is tabulated	
	CITING_MAX-[*feature*]	Maximum value of the feature in the cited papers	
	CITING_MIN-[*feature*]	Minimum value of the feature in the cited papers	
	CITING_AVG-[*feature*]	Average value of the feature in the cited papers	
	CITING_SUM-[*feature*]	Total value of the feature in the cited papers	

C. Construction of the prediction model

We defined emerging papers as those for which citation three years after (t_0+3) publication at t_0 is in the top 5% of all papers published in that year (t_0). Li and Tong showed that information three years after publication is useful for prediction of the number of citations after 10 years. Therefore, we assumed that a three-year window is sufficient to predict future citations. Since many papers are published yearly in the nanomaterial field, 10% will exceed 5,000 papers in a few years. Therefore, if the emerging paper cut-off was set at 10%, there will be too many emerging papers for assessment. Additionally, in a statistical context, 5% probability is well-known clinically. We constructed a model that extracts features of emerging papers. In the model, data at three years after publication (t_0+3) are used as correct data and applied for prediction four years later (t_0+4=t_1). Data for year t_1 is referred to as prediction target year data. To evaluate the performance of the model, the citation number at three years after the prediction target year (t_1+3) is required. Figure 1 shows the relationship between the learning target period and prediction target period.

For example, if 2011 was the prediction target year (t_1), model construction requires features data based on the citation network up to the year 2007 (t_0), and the correct data at t_0+3. This is called the “2007 model”. We apply this “2007 model” to the paper dataset published in 2011 (t_1=t_0+4) to perform prediction. Evaluation of this prediction model is conducted using the citation number information at the end of 2014 (t_1+3). Table 2 shows the data for emerging papers for the prediction target year and the prediction results verification year from the “2002 model” to the “2007 model”.

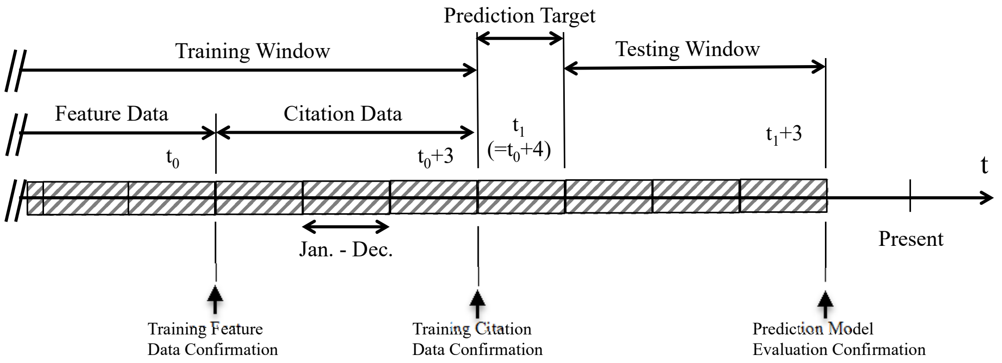

Figure 1. Model training and prediction

TABLE 2. MODEL TRAINING YEAR AND CORRESPONDING TARGET AND EVALUATION YEARS

Training window		Testing window	
Model training year t_0	Correct data (training citation data) confirmation year t_0+3	Prediction target year t_1	Prediction model evaluation year t_1+3
2002	2005	2006	2009
2003	2006	2007	2010
2004	2007	2008	2011
2005	2008	2009	2012
2006	2009	2010	2013
2007	2010	2011	2014

To construct the model, we used statistical machine learning. This prediction task can be explained by the supervised machine learning approach, in which a computer receives a series of input data (features) and corresponding correct data, and detects errors by comparison between output from an algorithm and correct data. In this study, the input data are features and correct data are data indicating whether the paper is emerging. Using knowledge from the data confirmation year, items that become emerging papers are assigned a flag "1" and treated as positive. Papers with citation numbers in the bottom 50% are assigned a flag "0" and treated as negative. These binary assignments were used as correct data (in other words, response variables). We constructed a model to describe the response variables using the features (in other words, explanatory variables) calculated as shown above.

In the analysis, we utilized logistic regression, a linear classifier, and LIBLINEAR in the analysis package [32]. A logistic regression model provides a probability that a response variable will be “1”. That is, the probability indicates whether a paper will be emerging or not. The prediction model is constructed using learning weights for each feature in linear equations. Utilizing the logistic regression model, the algorithm detects features with high impact for prediction of emerging papers. The inner mechanism of the constructed models can be understood from these weights. A support vector machine is not suitable for this purpose because the method cannot provide such weights, although it is a well-known classifier method in machine learning. However, in logistic regression, regularization parameters are set as initial values, and these parameters affect the prediction performance. To explore appropriate parameters and avoid overfitting, we divided learning data into 5 sets in each model, with 4 used for learning the weights and 1 for validation (5-fold cross-validation) in our framework.

D. Evaluation of the prediction model

As shown in Table 2, the model was created using networks constructed each year from 2000 to 2007, and the results were verified based on the actual citation numbers three years after the networks are fully established, from 2004 to 2011. An F-value was used to evaluate the model. The F-value is an index obtained using the harmonic mean of the conformance rate and the recall rate. The conformance rate is defined as the number of papers that are actually emerging, compared to the number of papers predicted to be emerging. The recall rate describes the rate of papers predicted to be emerging out of those that are actually emerging. The F-value is normally used as an index for evaluation of a prediction model.

III. RESULTS

A. Dataset retrieval and feature creation

Papers published between January 1, 1901 and November 31, 2015 with “nano*” and “carbon*” in the title or keywords were extracted from the Web of Science database. A total of 411,084 papers were extracted that fit these criteria. Figure 2 shows a plot of the number of qualifying publications in each year since 1901. This figure shows that the number of publications suddenly rose in 1991, with an inflection. Over 45,000 papers were published in 2014.

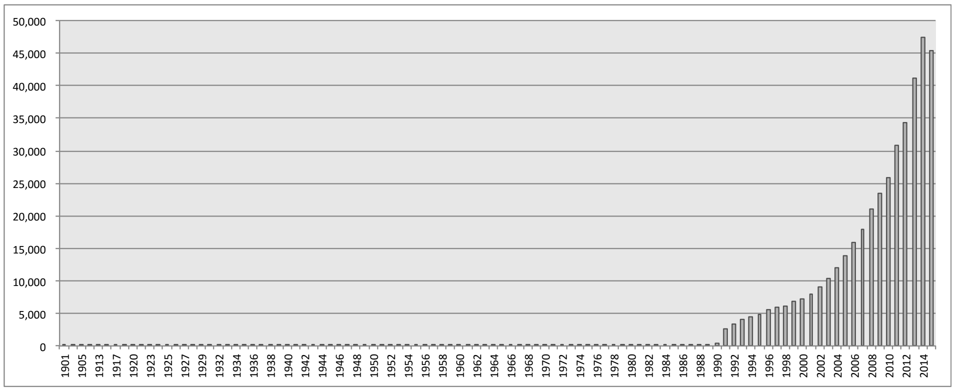

Figure 2. Number of publications in each year
(x-axis: Year of publication, y-axis: Number of publications in each year)

TABLE 3. TOP FIVE CONTRIBUTING FEATURES IN EACH MODEL

2002 model for 2006		2003 model for 2007		2004 model for 2008	
CNT_PAGER	20.5	CNT_PAGER	22.3	CNT_PAGER	27.1
CNT_AUTHO	9.4	CNT_AUTHO	10.3	CNT_AUTHO	11.2
CITING_MAX-CNT_DEGRE	5.3	CNT_DEGRE	8.0	CNT_DEGRE	9.0
CNT_DEGRE	5.3	CITING_MAX-CNT_DEGRE	5.4	CNT_CLOSE	5.5
CITING_SUM-CL_RANK	4.2	CNT_CLOSE	4.3	CITING_AVG-CNT_CLOSE	4.5
2005 model for 2009		**2006 model for 2010**		**2007 model for 2011**	
CNT_PAGER	23.3	CNT_PAGER	25.8	CNT_PAGER	33.1
CNT_AUTHO	9.7	CNT_AUTHO	18.3	CNT_AUTHO	14.9
CNT_DEGRE	6.1	CNT_DEGRE	8.2	CNT_CLOSE	9.3
CITING_SUM-CL_RANK	3.6	CNT_CLOSE	5.7	CNT_DEGRE	8.9
CITING_SUM-CL_QMAX	3.5	CITING_SUM-CL_RANK	4.6	CITING_AVG-CNT_CLOSE	5.2

A network was constructed based on the direct citation coefficient between papers. There were 379,044 resulting papers that were incorporated into the maximum connected component, which is the element comprising the largest islands in the network groups. For papers residing in the maximum component, we calculated feature values at the time of publication. We also calculated the predicted citation number three years after publication for all papers.

B. Model development

A model was constructed for predicting a paper published between 2006 and 2011 that would be designated as a emerging paper after three years. As explained above, models were built using data from 2002 to 2007 to perform feature value learning. Table 3 shows the top five rankings from the feature value weighting.

In Table 3, the feature with the highest contribution was PageRank (CNT_PAGER). PageRank is based on a weighting defined by citation relationships between papers. This index identifies a paper cited by papers that are themselves frequently cited. High PageRank suggests that an author who wrote a paper cited papers in the reference list that are frequently cited in other papers, and therefore the author must have surveyed the related area very well. In this sense, it is reasonable that such a paper is regarded as a paper that will have many future citations. Furthermore, this reduces the relative importance of papers that contain mutual citations. In all models, the feature value with the second highest contribution was Authority Score (CNT_AUTHO). This index rates papers based on their span across multiple clusters of research. Higher scores are given to papers that have a bridging role among cluster members. We included this property based on the assumption that new fields arise from papers that transcend fields, hinting at the formation of a emerging new area.

Degree centrality (CNT_DEGRE) also had a high weight in several models. The more a paper is cited in reference lists, the higher the index. As for PageRank (CNT_PAGER), an author who wrote a paper citing many papers in the reference list must have surveyed the related area very well. In all models, network centrality features dominated the top rankings. Based on this observation, it may be possible to predict emerging papers based only on network centrality. Aside from the network centrality feature, the paper's group feature value from its reference list was also valid.

C. Model evaluation

The citation network expands over time, and performance in each model varies with network growth. Therefore, we built several models over time to evaluate the stability of the prediction performance. We evaluated the results from target years 2006 to 2011, with citations (impact) measured in 2009 to 2014 (prediction evaluation year (t_1+3)). Based on the citation number, we are able to verify whether the predictions of emerging papers were correct. Table 4 shows the model performance for each year. We randomly extracted the same data volume for negative examples as that for positive examples several times to create multiple balanced datasets for each year, and calculated the performance average. Since the F-value exceeds 80 in later years, we consider the most recent models to be most reliable.

TABLE 4. AVERAGE VALUE OF RESULTS IN EACH BALANCE SET IN EACH YEAR

Model training year t_0	Prediction target year t_1	Prediction model evaluation year t_1+3	Prediction target paper number	Emerging paper number	Precision	Recall	F-value
2002	2006	2009	2,990	1,495	91.48	53.77	67.51
2003	2007	2010	3,598	1,799	95.80	48.14	63.77
2004	2008	2011	3,990	1,995	89.56	63.66	74.26
2005	2009	2012	4,664	2,332	92.68	40.43	55.54
2006	2010	2013	4,994	2,497	83.39	89.13	86.15
2007	2011	2014	5,830	2,915	83.54	87.16	85.30

TABLE 5. TOP 10 PAPERS PUBLISHED IN 2011 PREDICTED TO BE EMERGING PAPERS.

	Authors	Title	Journal	Times Cited (2011)	Times Cited (2014)	Ref.
1	Zhang, Q; Huang, JQ; Zhao, MQ; Qian, WZ; Wei, F	Carbon Nanotube Mass Production: Principles and Processes	Chemsuschem	0	84	[33]
2	Lan, YC; Wang, Y; Ren, ZF	Physics and applications of aligned carbon nanotubes	Advances in Physics	0	35	[34]
3	Lee, SH; Lee, DH; Lee, WJ; Kim, SO	Tailored Assembly of Carbon Nanotubes and Graphene	Advanced Functional Materials	6	82	[35]
4	Das Sarma, S; Adam, S; Hwang, EH; Rossi, E	Electronic transport in two-dimensional graphene	Reviews of Modern Physics	51	664	[36]
5	Huang, X; Yin, ZY; Wu, SX; Qi, XY; He, QY; Zhang, QC; Yan, QY; Boey, F; Zhang, H	Graphene-Based Materials: Synthesis, Characterization, Properties, and Applications	Small	26	587	[37]
6	Saito, R; Hofmann, M; Dresselhaus, G; Jorio, A; Dresselhaus, MS	Raman spectroscopy of graphene and carbon nanotubes	Advances in Physics	0	98	[38]
7	Li, YD; Li, DX; Wang, GW	Methane decomposition to COx-free hydrogen and nano-carbon material on group 8-10 base metal catalysts: A review	Catalysis Today	5	46	[39]
8	Yan, LA; Zhao, F; Li, SJ; Hu, ZB; Zhao, YL	Low-toxic and safe nanomaterials by surface-chemical design, carbon nanotubes, fullerenes, metallofullerenes, and graphenes	Nanoscale	5	67	[40]
9	Singh, V; Joung, D; Zhai, L; Das, S; Khondaker, SI; Seal, S	Graphene based materials: Past, present and future	Progress in Materials Science	7	506	[41]
10	Leary, R; Westwood, A	Carbonaceous nanomaterials for enhancement of TiO_2 photocatalysis	Carbon	11	223	[42]

The top 10 ranked papers predicted to be emerging papers among papers published in 2011 are shown in Table 5. Most of these papers had high citation numbers three years later, as of 2014. The papers that actually attained the conditions by which we defined a emerging paper are underlined in the times cited (2014) column. Of the 10 predicted papers, 9 were emerging papers as of 2014. The paper listed at #2 did not attain the conditions for a emerging paper, but it did have a high citation number. Regardless, 90% of the predicted papers in Table 5 did become emerging papers.

IV. DISCUSSION

The data in Figure 2 show an explosive increase in knowledge in recent years in the nanocarbon materials field. In particular, there was a sudden increase in the number of papers in 1991. The discovery of tubular nanocarbon (carbon filaments) was made by Radushkevich and Lukyanovich in 1952 [43], but the discovery of carbon nanotubes by Iijima in 1991 was the spark for expansion in this field [2].

In this study, we constructed a model that predicts whether papers will become emerging papers three years after publication. As shown in Table 4, the F-value for the model in 2011 (stemming from the 2007 model) was over 80; therefore, it appears a balanced model was constructed for the conformance rate and recall rate. For models from 2006 to 2011, the greatest contributing feature was PageRank. This indicates that citing a paper that has a high citation number increases the probability of a high citation number for the paper. In addition, degree centrality indicates that a paper that references many papers will eventually also have a high citation number. Of the top 10 papers (Table 5) published in 2011 that were predicted to become emerging papers, 9 actually became emerging papers. This indicates that our constructed model is valid. We summarize these papers below. All of these papers are reviews.

Paper #1 focuses on carbon nanotube (CNT) mass production [33]. The paper discusses the arc discharge and laser vaporization methods, which were previously used in CNT manufacturing. The arc discharge method can be used to manufacture quality CNTs with few defects, but does not allow manufacture at a reasonable industrial volume. The laser vaporization method also produces CNTs of relatively high purity, but is not considered to be suitable for industrial-level manufacturing. After giving background on these methods, the paper focused on the chemical vapor deposition (CVD) method, which is reported to be suitable for mass-volume synthesis. Based on ideas such as "carbon multiwall nanotubes" at the Hyperion Company by Professor Endo of Shinshu University and the "CoMoCAT Process at SWeNT" at the University of Oklahoma, several manufacturing technologies have been pioneered and are being used. This paper provided a general introduction and discussion of research related to methods for CNT mass production. This paper had 84 citations in a three-year period.

Paper #2 discusses the current status, physical characteristics, and applications of aligned CNTs [34]. Among methods for CNT manufacturing, CVD is important due to its high orientation ability. These aligned CNTs have potential applications in a wide range of areas, including in field emission, optical antennae, subwavelength light transmission in CNT-based nanocoax structures, and nanocoax arrays for novel solar cell structures. This paper did not satisfy our criteria for a emerging paper, but it did receive a high number of citations.

Paper #3 focuses on tailored assembly of CNTs and graphene into three-dimensional architectures [35]. The paper describes how to achieve scalability for practical mass production of these materials.

Paper #4 reviews the properties of graphene obtained from the one-atom thick surface of graphite crystals [36]. In 2004, Geim et al. reported use of adhesive tape to separate the surface of highly oriented pyrolytic graphite (HOPG) and then extract a flake of graphene. Subsequently, the electrical, electronic, mechanical, and scientific properties of this material have been defined [44]. Graphene has high electron mobility, a measure of the speed of electrons within a solid. A mobility of 2,000,000 cm^2/Vs was predicted theoretically [45] and later experimentally verified [46]. This value is more than 100 times greater than the mobility of electrons in silicon, which is 1000 cm^2/Vs at its highest. High electron mobility is important in creation of high-speed transistors and other related technologies. This paper describes the electrical characteristics and potential applications of the high electron mobility of graphene. As of 2014, it had attained 664 citations.

Paper #5 considers additional properties of graphene [37]. In addition to high electron mobility, this material also offers thermal stability and excellent mechanical strength. The paper describes the importance of these physical characteristics and the possible applications of grapheme, including in FETs, memory, photovoltaic devices, and sensing platforms. The paper has received 587 citations, which makes it one of the most cited of the 10 top papers.

Paper #6 focuses on structural analysis of carbon nanotubes and graphene, and summarizes related research [38]. Raman spectroscopy of carbon materials gives peaks that indicate structural properties. G-band peaks derived from graphite structure and D-band peaks due to defects reveal information on graphene and nanotube quality. These peak comparisons are useful in the evaluation of crystalline purity or defect concentrations in nanocarbon materials.

Paper #7 describes a method through which steam breaks down methane using a catalyst in a high-temperature environment and generates hydrogen and carbon [39]. The generated hydrogen can be used as fuel for fuel cells. This method has attracted attention as a potential approach for hydrogen generation. In addition, since the generated carbon can be used directly in a carbon fuel cell, the method may allow generation of nanocarbon materials. This review paper summarizes the leading literature on formation of nanocarbon materials generated from the principles of methane catalytic decomposition.

Paper #8 shows that carbon nanotubes have a structure similar to asbestos and are similarly toxic to humans, indicating that reduction of the toxicity is important for utilization of nanocarbon materials [40]. This paper provides systematic nanotoxicology knowledge and discusses approaches that can be used to achieve low toxicity through chemical modifications in design and changes of the biological and toxicological properties of carbon nanomaterials.

Paper #9 summarizes the history of graphene, its characteristics, how it is formed, and the impact and application of graphene in electronic and optoelectronic devices, chemical sensors, nanocomposites, and energy storage [41]. As of 2014, this paper had 506 citations.

Paper #10 discusses titanium oxide (TiO_2), which is used as a photocatalyst, but has low efficiency and a narrow response range [41]. Combining nanocarbon materials with TiO_2 enables significant changes in these characteristics. This review considers nanocarbon-TiO_2 as a photocatalyst, with guidelines for generation, characteristics, and future directions. As of 2014, the paper had 232 citations.

This summary covers the top ten papers that were predicted to be emerging papers among all the papers published in 2011. All 10 papers had a high number of citations as of 2014. With the exception of paper #2, all papers satisfied the criteria for our definition of emerging papers. Therefore, we conclude that it is possible for a prediction model to evaluate the future importance of specific research. These ten papers all considered carbon nanotube applications and all were reviews. Degree centrality had a relatively high weight in our model, and thus it follows that many highly ranked papers reviewed a breadth of topics. In general, reviews give a survey of a specific area, with the goal of introducing and describing the topic. Therefore, the reference list of a review paper is often quite lengthy. As a result, our model showed a strong tendency to extract review papers, and future models may require reformulation to avoid this result. Nevertheless, not all papers with long reference lists became emerging papers, and extraction of potential emerging reviews can give perspective on future trends in the field.

V. CONCLUSION

The increase in science and technology knowledge and trends towards complexity and detail make it very difficult for any one person to have a full understanding of every development in any given field. To address this problem, we constructed a model for prediction of papers that are likely to be emerging papers in the future. The goals of the work were to identify emerging papers at an early stage based on information immediately after publication. The defining characteristics were the use of diverse features, network indexes, and clustering results, with the most novel aspect of the prediction model being the use of features at the time of publication to predict the citation number several years in the future. The constructed features used in the model were divided broadly into four classes: network features, cluster features, centrality features, and citation-related features. Linear logistic regression was used in the prediction model algorithm.

We extracted all papers (411,084) from the Web of Science that incorporated "nano*" and "carbon*" in keywords or titles and were published after January 1, 1901. After constructing a model for each year, we predicted emerging papers. Features related to network centrality were particularly in this prediction. This may be because researchers who write papers that cite other suitable papers are likely to have fully surveyed the target field and are able to construct a reference list that is neither excessive nor deficient. The paper can easily be cited by other researchers after publication, and thus attracts a large number of citations.

Some of our models had an F-value of 80 or more, whereas the F-value for a random prediction is generally less than 50. Dong achieved about F=70 in some models, compared with F=38 in a random model. Most previous models do not exceed F=80, and we were able to reach this level of performance without use of citation data for several years after publications. A summary of the top 10 predicted papers was performed to understand the importance of these papers. The predicted emerging papers were all reviews, which may be a limitation of our method. However, not all review papers collect many citations, even those with a long reference list, and many reviews do not have a lot of citations. Therefore, detection of review papers that are likely to be emerging papers is meaningful and the fields mentioned in such papers can be viewed as emerging fields. However, we are also trying to address this limitation. We are also considering broader applications in other fields. Since the citation network is extracted in each target field, prediction models can learn data based on a characteristic citation pattern (knowledge expansion pattern). For a dataset from another field, the model learns based on the characteristics of the pattern of the field. We plan to develop a higher performance and more stable model that can contribute to policy-making and identification of future trends in multiple sectors.

As the amount of information continues to expand and knowledge becomes more complex, it will only become more difficult for corporations and national governments to make decisions on where to focus research and development investments and make budget allocations. Foresight in science or technology trends is challenging and requires a proactive approach. Our prediction model can provide support for government officials and investors engaged in decision-making. We believe that the need for tools that support extraction of useful information from vast pools of papers will increase in the future.

ACKNOWLEDGEMENTS

This research was supported by grants from the Project of the NARO Bio-oriented Technology Research Advancement Institution (Integrated Research for Agriculture and Interdisciplinary Fields).

REFERENCES

[1] Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., & Smalley, R. E. (1985). C 60: buckminsterfullerene. *Nature*, *318*(6042), 162-163.

[2] Iijima, S. (1991). Helical microtubules of graphitic carbon. *nature*, *354*(6348), 56-58.

[3] D'Souza, F., & Ito, O. (2012). Photosensitized electron transfer processes of nanocarbons applicable to solar cells. *Chemical Society Reviews*, *41*(1), 86-96.

[4] Burke, A. (2007). R&D considerations for the performance and application of electrochemical capacitors. Electrochimica Acta, 53(3), 1083-1091.

[5] Suzuki, K., Yamaguchi, M., Kumagai, M., & Yanagida, S. (2003). Application of carbon nanotubes to counter electrodes of dye-sensitized solar cells. *Chemistry Letters*, *32*(1), 28-29.

[6] Kasavajjula, U., Wang, C., & Appleby, A. J. (2007). Nano-and bulk-silicon-based insertion anodes for lithium-ion secondary cells. *Journal of Power Sources*, *163*(2), 1003-1039.

[7] Liu, C., Li, F., Ma, L. P., & Cheng, H. M. (2010). Advanced materials for energy storage. *Advanced Materials*, *22*(8), E28-E62.

[8] Edwards, B. C. (2000). Design and deployment of a space elevator. Acta Astronautica, 47(10), 735-744.

[9] Pugno, N. M. (2006). On the strength of the carbon nanotube-based space elevator cable: from nanomechanics to megamechanics. *Journal of Physics: Condensed Matter*, *18*(33), S1971.

[10] Pugno, N. M. (2007). The role of defects in the design of space elevator cable: from nanotube to megatube. *Acta Materialia*, *55*(15), 5269-5279.

[11] Rescher, N. (1998). *Predicting the future: An introduction to the theory of forecasting*. SUNY press.

[12] Winnink, J. J., & Tijssen, R. J. (2015). Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. *Scientometrics*, *102*(1), 113-134.

[13] Goffman, W., & Newill, V. A. (1964). Generalization of epidemic theory. *Nature*,*204*(4955), 225-228.

[14] Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chavez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495-518.

[15] Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191-209.

[16] Young, P. (1993). Technological growth curves: a competition of forecasting models. *Technological forecasting and social change*, *44*(4), 375-389.

[17] Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, *63*(3), 567-581.

[18] Li, L., & Tong, H. (2015). The Child is Father of the Man: Foresee the Success at the Early Stage. *arXiv preprint arXiv:1504.00948*.

[19] Dong, Y., Johnson, R. A., & Chawla, N. V. (2015, February). Will This Paper Increase Your h-index?: Scientific Impact Prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 149-158). ACM.

[20] Davletov, F., Aydin, A. S., & Cakmak, A. (2014, November). High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 491-498). ACM.

[21] Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014, September). Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 351-360). IEEE Press.

[22] Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127-132.

[23] Rogers, Everett (16 August 2003). *Diffusion of Innovations, 5th Edition*. Simon and Schuster.ISBN 978-0-7432-5823-4.

[24] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577-8582.

[25] Freeman, L. C. (1979). Centrality in social networks conceptual

clarification. *Social networks*, *1*(3), 215-239.

[26] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

[27] Bonacich, P. (1972). Technique for analyzing overlapping memberships. *Sociological methodology*, *4*, 176-185.

[28] Burt, R. S. (2004). Structural holes and good ideas1. *American journal of sociology*, *110*(2), 349-399.

[29] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *nature*, *393*(6684), 440-442.

[30] Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, *56*(18), 3825-3833.

[31] Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*(7028), 895-900.

[32] LIBLINEAR -- A Library for Large Linear Classification. (n.d.). Retrieved Jan. 28, 2016, from https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[33] Zhang, Q., Huang, J. Q., Zhao, M. Q., Qian, W. Z., & Wei, F. (2011). Carbon nanotube mass production: principles and processes. *ChemSusChem*, *4*(7), 864-889.

[34] Lan, Y., Wang, Y., & Ren, Z. F. (2011). Physics and applications of aligned carbon nanotubes. *Advances in Physics*, *60*(4), 553-678.

[35] Lee, S. H., Lee, D. H., Lee, W. J., & Kim, S. O. (2011). Tailored assembly of carbon nanotubes and graphene. *Advanced Functional Materials*, *21*(8), 1338-1354.

[36] Sarma, S. D., Adam, S., Hwang, E. H., & Rossi, E. (2011). Electronic transport in two-dimensional graphene. *Reviews of Modern Physics*, *83*(2), 407.

[37] Huang, X., Yin, Z., Wu, S., Qi, X., He, Q., Zhang, Q., ... & Zhang, H. (2011). Graphene‐based materials: synthesis, characterization, properties, and applications. *Small*, *7*(14), 1876-1902.

[38] Saito, R., Hofmann, M., Dresselhaus, G., Jorio, A., & Dresselhaus, M. S. (2011). Raman spectroscopy of graphene and carbon nanotubes. *Advances in Physics*, *60*(3), 413-550.

[39] Li, Y., Li, D., & Wang, G. (2011). Methane decomposition to CO x-free hydrogen and nano-carbon material on group 8–10 base metal catalysts: a review. *Catalysis today*, *162*(1), 1-48.

[40] Yan, L., Zhao, F., Li, S., Hu, Z., & Zhao, Y. (2011). Low-toxic and safe nanomaterials by surface-chemical design, carbon nanotubes, fullerenes, metallofullerenes, and graphenes. *Nanoscale*, *3*(2), 362-382.

[41] Singh, V., Joung, D., Zhai, L., Das, S., Khondaker, S. I., & Seal, S. (2011). Graphene based materials: past, present and future. *Progress in Materials Science*, *56*(8), 1178-1271.

[42] Leary, R., & Westwood, A. (2011). Carbonaceous nanomaterials for the enhancement of TiO_2 photocatalysis. *Carbon*, *49*(3), 741-772.

[43] Radushkevich, L. V., Lukyanovich, V. I. (1952). Carbon structure formed under thermal decomposition of carbon monoxide on iron, *Zh. Fiz. Khim., 26*(1) 88-95

[44] Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. A., ... & Firsov, A. A. (2004). Electric field effect in atomically thin carbon films. *science*, *306*(5696), 666-669.

[45] Hwang, E. H., Adam, S., & Sarma, S. D. (2007). Carrier transport in two-dimensional graphene layers. *Physical Review Letters*, *98*(18), 186806.

[46] Bolotin, K. I., Sikes, K. J., Jiang, Z., Klima, M., Fudenberg, G., Hone, J., ... & Stormer, H. L. (2008). Ultrahigh electron mobility in suspended graphene. *Solid State Communications*, *146*(9), 351-355.