Detecting Structural Changes in the Nanocarbon Domain Based on the Time Distribution of Text Information of Academic Papers

Yuko Nakashio[1], Tadayoshi Hara[2], Junichiro Mori[2], Ichiro Sakata[2]
[1]The University of Tokyo, Department of Systems Innovation, Faculty of Engineering, Tokyo, Japan
[2]Policy Alternatives Research Institutes, The University of Tokyo, Tokyo, Japan

***Abstract*--In recent years, there has been an increasing need for the early detection of emerging research fronts. Research in this field usually employs citation networks, but this methodology does not address the citation lag problem. Text information is required to solve the time gap in citation networks because text information is available immediately when papers are published. However, text information has an inherent domain dependency problem. To address this, we introduce the ”Dynamic Topic Model” (DTM). In a DTM, text information is represented in an abstract "topic" form and text information is captured as an increase or decrease in topics. We apply a DTM to the nanocarbon domain, which has experienced significant structural changes. We note that the choice of a suitable number of topics for the DTM requires further research. In this paper, we show that the proposed methodology, text information analysis with a DTM, can detect emerging research fronts earlier than the citation network technique.**

I. INTRODUCTION

In recent years, the innovation cycle has accelerated and science linkage has grown. As with knowledge expansion, there is an increasing demand for techniques to manage this large volume of data, as the reliance on expert knowledge alone is insufficient. It is important to detect emerging research fronts at an early stage, as this facilitates managers and stakeholders in making timely investments and monitoring competitors' research. Furthermore, decision-makers and policy-makers can allocate budgets more efficiently.

The citation network has been employed in previous research related to the detection of emerging research fronts. [15]. A range of network features of the citation network have been exploited in the literature: Shibata et al. [17] used the average publication years of clusters, Sakata et al. [14] used bibliometric information and network indicators, Fujita et al. [4] used different types of weighted citation networks, and Iwami et al. [8] used dynamics of network indicators. The main limitation with detecting emerging research fronts using citation networks is citation lag. By way of example, consider a scenario where paper A is published in year X and paper B cites paper A. If paper B is published one year later, there will be a one year delay in detecting the citation link between paper A and B. This effect is known as citation lag. Text information is required to solve the time gap in citation networks because text information is available immediately when papers are published.

A disadvantage of using text information is domain dependency. For example, Kajikawa et al. [9] surveyed the sustainability domain using text information to track the change in the meaning of the word sustainability. However, the methodology and knowledge of that research cannot be applied to other domains because word usage differs between domains. To solve the domain dependency problem, text information needs to be abstracted to an objective value. In citation networks, cita- tions are abstracted using network features, such as closeness centrality and betweenness centrality. In this study, we employ a machine learning technique, the Dynamic Topic Model (DTM) [1], to describe text information as an objective value. In a DTM, text information is regarded as an abstract topic, and the model captures the growth or decline in the number of topics.

There are some previous research on bibliometric analysis with a DTM. One is about tracking the development of ideas [6]. This is case study in the field of Computational Linguistics. Another one is using the model in order to measure the importance of the papers [5]. We focus on one research field, not a paper.

In this research, we show that text information can overcome the citation lag problem. We evaluate our approach using the literature of the nanocarbon domain, as it is multidisciplinary and has experienced significant structural changes. Structural changes in the nanocarbon domain are caused by innovations in the field [3] and, therefore, it is both difficult and necessary to predict structural changes in this domain. In the last ten years, the number of papers in the nanocarbon domain has doubled, from approximately 20,000 to more than 40,000. During this period, certain subdomains have grown significantly, while others have diminished.

For context, we summarize the four main breakthroughs that have occurred in the nanocarbon domain: in 1990 Prof. K atschmer developed fullerene [7], in 1991 Prof. Endo developed the carbon nanotube [11], in 2004 K. S. Novoselov and A. K. Geim et al. developed graphene [13] and in 2010 K. S. Novoselov and A. K. Geim, the developers of graphene, won the Nobel Prize in Physics. In this study, we focus on the third breakthrough: the development of graphene in 2004. To summarize, using graphene as the main criterion, we show that the proposed methodology, text information analysis with DTM, can detect emerging research fronts earlier than the citation network technique.

II. METHODOLOGY

A. Detection of emerging research fronts using a citation network

In this subsection, we describe how to use citation networks for the early detection of emerging research fronts. Figure 1 illustrates the methodology of citation network analysis.

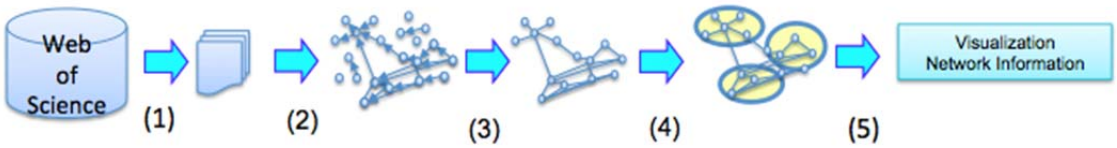


Figure 1: Methodology of citation network analysis [1)]

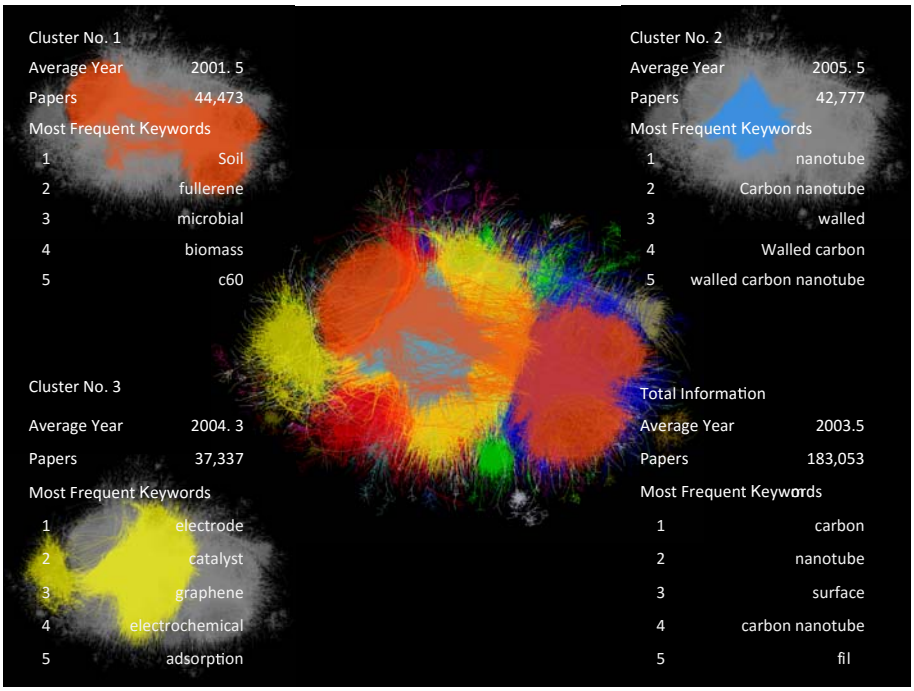


Figure 2: Citation Network of Nanocarbon in 2009

Figure 1 can be described, referring to the numbers in the figure, as follows: (1) we collect citation data from publications from the Web of Science (WoS), (2) we construct a citation network using direct citations [16], (3) we only use maximum components of networks, as papers with no links are not considered relevant to the domain, (4) we perform topological clustering using the Newman-Girvan algorithm [12]. Citation networks are divided into clusters, which extract communities where within-cluster links are dense, (5) after clustering, a spring-model [10] is used to visualize citation networks and assess network features in detail.

We collect bibliographic information from publications in the nanocarbon domain from the WoS, using a query designed by domain experts. The bibliographic information includes: publication year, keyword, authors, citation data, and so on. The query used is: TS=(((carbon and (nano* OR micro*)) or fullerene or Buckminsterfullerene or Buckminster-fullerene or C60 or C-60 or graphene or (lament* and carbon))), where * indicates a wildcard character. Bibliographic information is collected from papers published between 1990 and 2015. The collected data contains 407,336 items, of which 374,482 (91.9%) consist of maximum components.

Next, we show a citation network focusing on graphene. Figure 2 is a citation network in 2009, which is the first time a cluster related to graphene is detected. The sub-picture in the center of this figure is total information of the nanocarbon domain, covering the entire academic landscape. In this figure, “Average Year ” refers to the average year of publication and “ Papers ” refers to the number of papers in the cluster. “Most Frequent Keywords” refers to the keywords characteristic to each cluster. The upper left cluster (colored orange and named No.1 Nanobiotechnology) is the largest cluster. The upper right cluster (colored blue and named No.2 Multi-Walled Carbon Nanotube) is the second largest cluster. The lower center cluster (colored yellow and named No.3 Graphene) is the third largest cluster. In the No. 3 cluster, we have recognized the word “graphene” in the “Most Frequent Keywords” in 2009 for the first time. In the remainder of this study we focus on the graphene cluster. Furthermore, this No.3 cluster has separated from No.1 and the average year of No.3 cluster about Graphene is still older than that of the No.2 cluster about Multi-Walled Carbon Nanotube, although graphene is the newest and the most

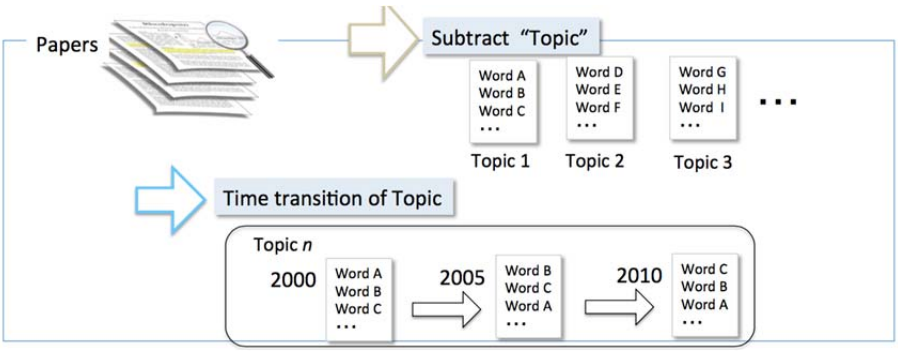


Figure 3: Conceptual Diagram of DTM [1]

remarkable material developed in 2004 and it is already more than 15 years since Carbon Nanotube was developed in 1991. It is worth noting that we only detect the graphene cluster 5 years after its discovery in 2004. This long delay is caused by citation lag.

B. Detecting emerging research fronts using text information

In this subsection, we introduce a new methodology for detecting emerging research fronts using text information. Using citation networks, there was a seven-year delay to detect the emerging research front of graphene. The Dynamic Topic Model is derived from the Topic Model [2]. In the Topic Model method, documents are considered to contain topics, which follow a probability distribution. In reality, we can only observe terms from text information in documents, and we arbitrarily decide the number of topics. Then, with time transition, the model automatically learns the topic distribution using given parameters. By calculating the term frequency with DTM, we can extract and evaluate topics important and prior in academic research fields. This procedure is depicted in Figure 3.

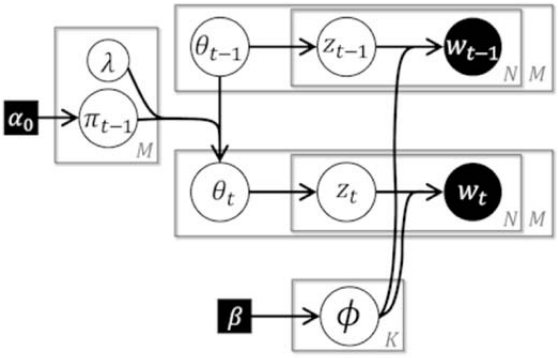


Figure 4: Graphical Model of DTM [1]

Dynamic Topic Model doesn't mean that we apply Topic Model every year. In Figure 4, we show the graphical model of Dynamic Topic Model. Parameters colored with black are observable. For example, α_0 and β are initial values and w_t is words in the documents. N (the number of words) and M (the number of documents) are also observable and K (the number of topics) is the parameter which we set arbitrarily. λ represents time transition. From the given parameters, this model learns the distribution of topics in each document, φ(the distribution of words in each topic), and Z (the distribution of topics using each word).

To summarize, text information is abstracted as a topic in a DTM-based approach. This abstraction overcomes the domain dependency problem, which is one of the disadvantages of using text information. Text information is described as an increase or decrease in the size of each topic and as a word possibility distribution in each topic, allowing text information to be abstracted. In the next section, we show the result of using the proposed DTM-based approach and analyze the results.

III. EXPERIMENT

In this research, we apply two methodologies (citation network analysis and the proposed DTM technique) to the same corpus.

In citation network analysis, we used five characteristic keywords in each cluster, which have the highest term frequencyinverse document frequency (tf-idf). In this research, "Documents " in tf-idf mean clusters. Tf-idf is generally used when extracting characteristic keywords in each document.

In text information analysis, we use Keyword Plus Information as the text information, which is collected from WoS using its API. Keyword Plus Information refers to the keywords that Thomson Reuter give each paper; these are extracted automatically from references and generally different from the keywords given by the author. Because authors' keywords often contains spelling inconsistency including stopwords and tend to be subjective, we regard Keywords Plus as more objective and controllable. In this experiment, we arbitrarily decide that the number of topics is twenty. In a DTM, documents consist of topics and topics are learned from the associated term frequency. We set time transition interval as one year. Except the number of topics and time transition interval, the other parameters remain default.

The results of each methodology are investigated and compared using two metrics: semantic similarity and the time to detect emerging research fronts. Citation networks have been already evaluated when capturing entire academic

trends in each domain. If the proposed methodology has semantically similar results to the citation network results, then the new methodology will also be useful. In other words, we evaluate the validity of the proposed methodology by comparing against the semantic results of citation networks. To assess semantic similarity, each topic is labeled by domain experts. The second metric used to compare the two sets of results, is when the method detects emerging fronts of research. Here we assess whether the proposed method can detect emerging research fronts earlier than citation networks. To summarize, we compare both methods using semantic similarity and the time to detecting emerging research fronts.

IV. RESULTS

In this section, we present the results of the proposed methodology, DTM. First, we compare the semantic results of topics with text information and compare clusters with the citation network in Table 1.

Table 1 (a) shows topics with text information, using the new methodology and Table 1 (b) shows clusters using the citation network methodology. In Table 1 (b) cluster number x-y indicates that x is the result by clustering once, and that y is result of clustering x again, in more detail. Cluster numbers are provided in descending order, following to the number of papers comprising each cluster. In other words, if x and y are smaller, the cluster size is bigger. We observe that we get semantically similar results using the new methodology with text information. For example, bio sensor in Cluster No. 3-2 is also detected in Topic 17. Lithium ion battery in Cluster No. 3-1 corresponds to Topic 18. Nano composites in Cluster Nos. 1-2 and 4-1 are also appeared in Topic 0 and 14. Solar cell in Cluster no. 2-5 is also detected in Topic 7. Adsorption in Cluster 3-4 and 4-7 also appeared in Topic 10 and 16, where drug delivery is based on adsorption. These examples all reflect current trends in the nanocarbon domain. Nanobiotechnology based on fullerene and the application of carbon nanotubes to a variety of disciplines have already generated a vast amount of knowledge, and the relatively new material graphene is now transitioning from basic research to applied research. Graphene only appears in Table 1, whereas, in the citation network, Clusters Nos. 3-x are related to graphene. Table 1 shows that the results of text information analysis by Dynamic Topic Model are semantically similar to the citation network analysis results.

TABLE 1: TOPICS AND CLUSTER LABELED BY EXPERTS

(a) Topics with Text Information

No.	Label
0	Carbon Fiber
1	Environmental Science
2	Nano Sheet
3	Evaluation of Basic Physical Property
4	Fullerene
5	Pollution Control
6	How to Make Nanocarbon
7	Solar Cell
8	Graphene
9	Environmental Impact Assessment (Ecosystem)
10	Adsorption
11	Environmental Impact Assessment (Component)
12	Bio Medicine
13	Graphene
14	Alloys with Carbon Nanotube
15	CO_2 Capture
16	Drug Delivery
17	Bio Sensor
18	Power Storage Device
19	Basic Physical Property

(b) Clusters using Citation Network

No.	Label
1-1	Multi walled CNT
1-2	Multi walled CNT and Nano Composites
1-3	Multi walled CNT
1-4	Field Emission
2-1	Soil
2-2	Diamond CNT
2-3	Titanium Oxide
2-4	Catalysis and Redox
2-5	Solar Cell
3-1	Lithium Battery
3-2	Biosensor and Electrode
3-3	Nano Ribbon
3-4	Adsorption
4-1	Carbon Fiber and Nano Composites
4-2	Carbon Steel
4-3	Oxidation Catalyst
4-4	Fermentation Medium
4-5	Supercritical Carbon Dioxide
4-6	Nanocarbon Thin Film
4-7	Adsorption
4-8	Corrosion of Carbon Steel

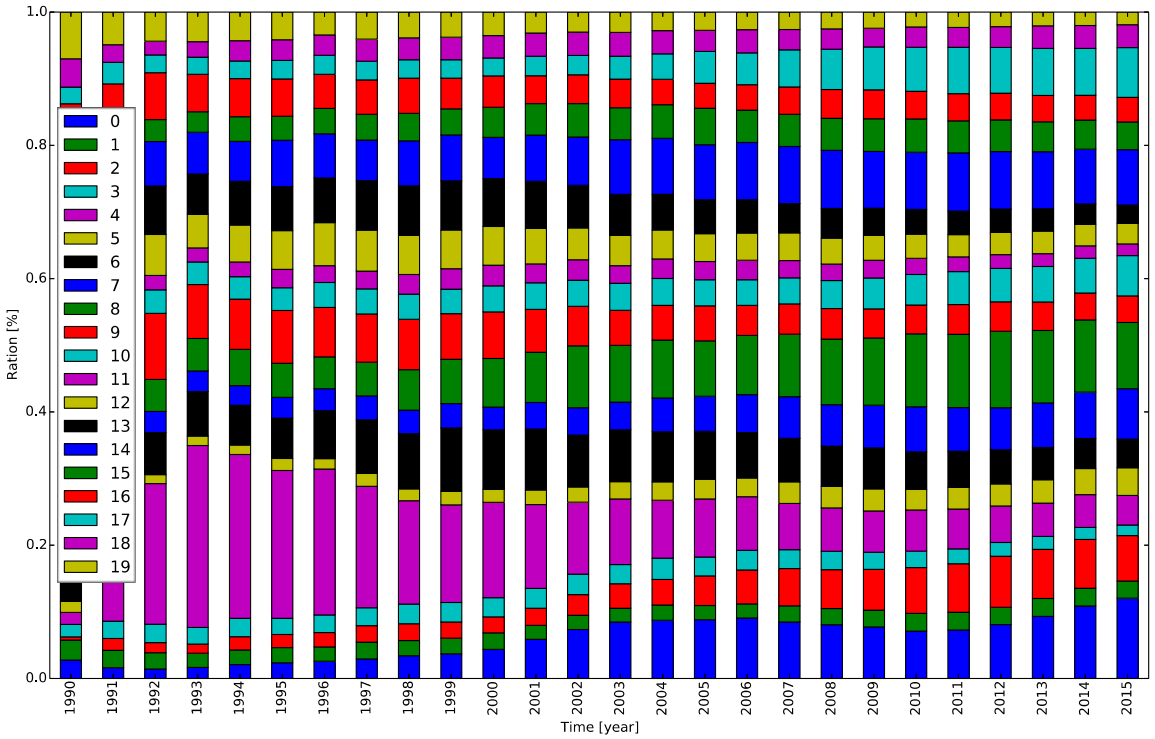
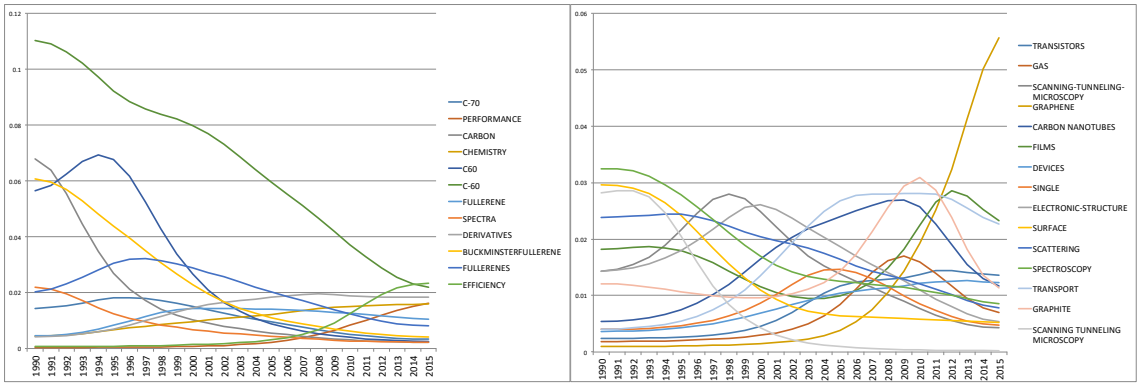


Figure 5: Time transition of diminution and growth of each topic

Next, we assess how quickly the proposed method detects emerging research fronts. Figure 5 shows the time transition of topic proportions, where some topics have increased in proportion, while other topics have decreased. Two topics which show dramatic changes in these results are: Topic No.4 ”Fullerene”, Topic No.8 ”Graphene”. In Figure 1, No.4 ”Fullerene” is pink and No.8 Graphene is green. No.4 Fullerene shows that four or five years after the discovery of fullerene in 1990, the proportion of No.4 rapidly diminished. Experts interpret that this corresponds to the fact that the basic research of fullerene’s physical properties converged earlier than expected, contrary to expectations. No.8 Graphene shows that after graphene’s discovery in 2015, its topic share increases up to two times. In a broader context, the graphene topic is still small, but the topic has grown rapidly. In 2015, the share ratio of No.8 ”Graphene” doubled compared with its ratio in 2004.

The word possibilities and topics are shown in Figure 6 and fullerene and graphene are given individual subplots. In No. 4 Fullerene (Figure 6(b)), as observed in Figure 5, the possibility of words, such as fullerene, buckminster-fullerene, C-60 drops after its discovery in 1994. In fact, re- search of the basic physical property of fullerene diminished sooner than expected, because fullerene is a simple molecule. In No. 8 Graphene (Figure 6(c)), very soon after graphene was developed in 2004, the word possibility of graphene starts to grow. This word possibility of graphene has the highest rate of growth among all word possibilities.



(a) No. 4 Fullerene (b) No. 8 Fullerene

Figure 6: Time Transition of word possibilityes of each topic

2008	2009	2010	2011
TRANSPORT	GRAPHITE	GRAPHITE	GRAPHITE
CARBON NANOTUBES	TRANSPORT	TRANSPORT	TRANSPORT
GRAPHITE	CARBON NANOTUBES	CARBON NANOTUBES	FILMS
GAS	FILMS	FILMS	GRAPHENE
FILMS	GAS	GRAPHENE	CARBON NANOTUBES
FIELD-EFFECT TRANSISTORS	GRAPHENE	GAS	TRANSISTORS
ELECTRONIC-STRUCTURE	FIELD-EFFECT TRANSISTORS	TRANSISTORS	GAS
SCATTERING	TRANSISTORS	FIELD-EFFECT TRANSISTORS	DEVICES
TRANSISTORS	SCATTERING	DEVICES	FIELD-EFFECT TRANSISTORS
SPECTROSCOPY	ELECTRONIC-STRUCTURE	SCATTERING	SCATTERING

Figure 7: Most Frequent 10 words in No. 8 Graphene

Figure 7 shows the time transition of the ten most frequent words in No. 8 ”Graphene ”. In 2008, graphene was not among the most frequent ten words, but by 2009 graphene was ranked 6th. In other words, we detect ”graphene ”in 2009. In subsequent years, graphene becomes more frequent, and after 2012, graphene becomes the most frequent word, with the highest growth rate.

In Section 2.1, in citation networks, only after part of networks which consist of papers about graphene separated from the networks of fulleren in 2009, we could detect the cluster of graphene and regard it as important. On the other hand, in text information analysis with dynamic topic model, the same as in citation network analysis, we could recognize the word “graphene” with high visibility in 2009 when the word “graphene” ranked in top 10 words. However, careful observation shared us the fact that No. 8 topic “graphene” started increasing the rate to its all before 2009 and in No.8 the possibility of “graphene” has dramatically grown soon after 2004. Considering these facts, we could suppose that we could detect the increase of graphene topic before 2009 and text information analysis could detect emerging research fronts with high visibility, at en earlier stage than citation network analysis.

V. DISCUSSION

A Dynamic Topic Model used for text information analysis works well for two reasons. First, the resulting text information analysis is semantically similar to that of citation network analysis. Second, text information analysis can detect emerging research fronts at an earlier stage than citation networks. However, a fundamental limitation is the arbitrary choice of the number of topics. There is a criterion ”perplexity” to evaluate the number of topics [2], but for this research, what ”most suitable” means depends on the purpose of experiments. In short, we have yet to overcome domain dependency, one of the disadvantages of text information.

We introduce two additional experiments, related to the number of topics. In the first additional experiment, we apply the DTM approach to three other domains: Aviation, Gallium Nitride (GaN), and Solar photos. Similar to the nanocarbon domain, the number of topics is also arbitrarily chosen to be twenty. Compared with the nanocarbon domain, we detect few major changes in topic distribution because these domains have experienced less dramatic structural changes. One possible explanation for this is that the choice of twenty topics is too small for these domains. In the other experiment, we set the number of topics in the nanocarbon domain to be one hundred. Except for the number of topics, we conduct the experiment under the same conditions as the original experiment. The authors of this paper, who are not domain experts, classified the word possibilities of each topic. The one hundred topics chosen include: fourteen are related carbon nanotube, five are related fullerene, eight are related graphene, and others are related biosensor, battery, and solar cells. More than half the topics are related to a physical or chemical background. Words from topics related to a physical and chemical background are too professional in each discipline to judge the context of each topic. However, there is possibility that we could observe most recent scientific trends in more detail in this additional experiment.

To summarize, we can detect structural changes in more detail if we choose a larger number of topics. Conversely, if we choose a smaller number of topics, we acquire comprehensive information about the topic time transition. Therefore, when using DTM for text information analysis, the main problem is to decide the suitable number of topics, which is domain dependent and varies from purpose to purpose.

We consider followings as future works. Adding to the number of topics, we have to investigate about multi-label problem, because DTM is based on probability model. When the number of topics are large, this problem could be solved. In this experiment, we used “Keywords Plus” as text information, which Thomson Reuter provides. Because

"Keyword Plus" is based on terms automatically extracted from the titles and the list of reference of articles, there remained some possibility that "Keywords Plus" doesn't reflect accurately the meanings of the papers. Therefore, we will conduct experiments on data directly extracted from titles and authors' keywords of papers, using NLP method.

VI. CONCLUSION

In recent years, there has been an increasing need for the early detection of emerging research fronts. In previous research, this has normally been achieved using citation networks; however, this technique has a citation lag problem. To solve the time gap problem in citation networks, text information analysis is required because it is available immediately when papers are published. However, text information has an inherent domain dependency problem. To overcome this limitation, we introduce the Dynamic Topic Model. In this model, text information is generalized as an abstract "topic" form, and is captured as an increase or decrease of the topic. We evaluated our DTM approach in the nanocarbon domain, which has experienced significant structure changes. Using the proposed methodology, the graphene topic is detected earlier than citation networks by careful observation. The most suitable number of topics remains to be assessed, which is domain dependent and varies according to purpose. In this paper, we have shown that the proposed methodology, text information analysis with DTM, can detect emerging research fronts with high visibility earlier than the citation network technique.

ACKNOWLEDGEMENTS

This research was supported by grants from the Project of the NARO Bio-oriented Technology Research Advancement Institution (Integration research for agriculture and interdisciplinary fields). I appreciate the feedback offered by Prof. Bunshi Fugetsu[1] and Prof. Mildred Dresselhaus[2]. They are experts in the nanocarbon domain and have evaluated our results of Topics and Clusters.

REFERENCES

[1] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pp. 113–120. ACM, 2006.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, Vol. 3, pp. 993–1022, 2003.

[3] Wilfred Dolfsma and DongBack Seo. Government policy and technological innovation – a suggested typology. Technovation, Vol. 33, No. 6, pp. 173–179, 2013.

[4] Katsuhide Fujita, Yuya Kajikawa, Junichiro Mori, and Ichiro Sakata. Detecting research fronts using different types of weighted citation networks. Journal of Engineering and Technology Management, Vol. 32, pp. 129–146, 2014.

[5] Sean Gerrish and David M Blei. A language-based approach to measuring scholarly impact. In ICML, Vol. 10, pp. 375–382, 2010.

[6] David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas us- ing topic models. In Proceedings of the conference on empirical methods in natural language processing, pp. 363–371. Association for Computational Linguistics, 2008.

[7] Sumio Iijima, et al. Helical microtubules of graphitic carbon. nature, Vol. 354, No. 6348, pp. 56–58, 1991.

[8] Shino Iwami, Junichiro Mori, Ichiro Sakata, and Yuya Kajikawa. Detection method of emerging leading papers using time transition. Scientometrics, Vol. 101, No. 2, pp. 1515–1533, 2014.

[9] Yuya Kajikawa, Junko Ohno, Yoshiyuki Takeda, Katsumori Matsushima, and Hiroshi Komiyama. Creating an academic landscape of sustainability science: an analysis of the citation network. Sustainability Science, Vol. 2, No. 2, pp. 221–231, 2007.

[10] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. Information processing letters, Vol. 31, No. 1, pp. 7–15, 1989.

[11] Wolfgang Kr¨atschmer, Lowell D Lamb, K Fostiropoulos, and Donald R Huffman. C60: a new form of carbon. Nature, Vol. 347, No. 6291, pp. 354–358, 1990.

[12] Mark EJ Newman. Clustering and preferential attachment in growing networks. Physical Review E, Vol. 64, No. 2, p. 025102, 2001.

[13] Kostya S Novoselov, Andre K Geim, SV Morozov, D Jiang, Y Zhang, SV Dubonos, , IV Grigorieva, and AA Firsov. Electric field effect in atomically thin carbon films. science, Vol. 306, No. 5696, pp. 666–669, 2004.

[14] Ichiro Sakata, Hajime Sasaki, Masanori Akiyama, Yuriko Sawatani, Naoki Shibata, and Yuya Kajikawa. Bibliometric analysis of service innovation research: Identifying knowledge domain and global network of knowledge. Technological Forecasting and Social Change, Vol. 80, No. 6, pp. 1085–1093, 2013.

[15] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, Vol. 28, No. 11, pp. 758–775, 2008.

[16] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology, Vol. 60, No. 3, pp. 571–580, 2009.

[17] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, Ichiro Sakata, and Katsumori Matsushima. Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technological Forecasting and Social Change, Vol. 78, No. 2, pp. 274–282, 2011.

[1] Nano-Agri Laboratory, Policy Alternatives Research Institute, The University of Tokyo

[2] Department of Physics, Massachusetts Institute of Technology