International Journal of Advanced Smart Convergence Vol.8 No.2 183-190 (2019)
http://dx.doi.org/10.7236/IJASC.2019.8.2.183

IJASC 19-2-22

A Study on Research Trend Analysis and Topic Class Prediction of Digital Transformation using Text Mining

JeeYoung Lee

Dept.of Software, Seokyeong University, Korea
J.Ann.LEE@skuniv.ac.kr

Abstract

In the era of the Fourth Industrial Revolution, digital transformation, which means changes in all industrial structures, politics, economics and society as well as IT technology, is an important issue. It is difficult to know which research topic is being studied because digital transformation is being studied in various fields. Convergence research is possible because a research topic is studied in various fields such as computer science area and Decision science area. However, it is difficult to know the specific research status of the research topic. In this study, eight research topics were derived using the topic modeling technique of text mining for abstract of academic literature and the trend of each topic was analyzed. We also proposed to create a Topic-Word Proportions Table in the LDA based Topic modeling process to predict the topic of new literature. The results of this study are expected to contribute to advanced convergence research on topic of digital transformation. It is expected that the literature related to each research topic will be grasped and contribute to the design of a new convergence research.

Keywords: *The fourth industrial revolution, Digital transformation, Text mining, Topic modeling*

1. INTRODUCTION

The fourth industrial revolution, triggered by more advanced digital technology since the third industrial revolution, is throwing many agenda with expectation. At its core is digital transformation. Digital transformation is already taking place in each industry, and by utilizing digital technology, it is driving changes in the value chain of existing industries as well as existing business processes. Understanding and studying digital transformation should be an important issue for successful settlement and response of the 4th Industrial Revolution. In response, this study examines the meaning of digital transformations and analyzes the topic and trends related to digital transformations by analyzing the literature using the Topic modeling technique of text mining. And the topic-word proportion table generated by the topic modeling is used to classify the literature by predicting topic for new literature.

Manuscript Received: May. 25, 2019 / Revised: June. 3, 2019 / Accepted: June. 10, 2019
Corresponding Author: J.Ann.LEE@skuniv,ac,kr
Tel: +82-2-940-7520, Fax: +82-2-940-7521
Dept. of Software, Seokyeong University, Korea

2. RELATED WORKS

2.1 The Fourth Industrial Revolution and Digital Transformation

The Fourth Industrial Revolution is a concept presented by Schwab at the World Economic Forum (WEF) in Davos, Switzerland in 2016. Schwab defined the fourth industrial revolution as a convergence of technology in the fields of physics, digital, and biology based on digital technology, which is more sophisticated and advanced since the third industrial revolution [1]. And that the fourth industrial revolution will have a significant impact on the industrial structure and market economy model around the world.

Digital transformation is a change in industrial structure, politics, economy and society caused by digital technology. Digital transformations are already taking place in various industries, utilizing digital technology to transform existing business processes as well as existing value chains. Digital transformation, in particular, makes it possible to leverage digital technologies such as big data, mobile, cloud and social to optimize and reconfigure business processes based on changes in processes that increase operational efficiency and competitiveness. Looking at existing literature, digital transformation is defined at various levels of analysis. Martin [2] analyzed the digital literacy stage at the individual level and classified it as digital competence - digital usage - digital transformation. Digital transformation can be understood as a level of realizing creative innovation using digital technology based on digital knowledge as the final stage of digital literacy at the level of personal analysis. Westerman et al. [3] defined digital transformation as a strategy for companies to integrate digital and physical components to transform business models and establish new directions for the industry, in relation to the use of digital technologies to improve the performance of companies such as efficiency and productivity from an enterprise strategic point of view. Matt et al. [4] defined digital transformation as a reshape of business structure using digital technology considering interaction with customers. Piccinini et al. [5] define digital transformations as a strategy to create new business value in all areas of technology and non-technology in terms of the relationship between producer and consumer. Collin [6] and Khan [7] defined digital transformation as a result of digitization at the social level as a global facilitation process of technology adaptation by individuals, businesses, societies and nations, Møller et al. [8] described market value as an advanced digital technology and described the process of innovation as a way of transcending the existing way of thinking through speed and scale. Ershova [9] classified the key components of the ecosystem of digital transformation and emphasized the importance of coherence between components.

2.2 LDA based Topic Modeling

Text mining, which was first introduced by Feldman et al. [10] is also known as knowledge discovery from text (KDT), and is a field of data mining that utilizes large amounts of data efficiently [11]. Text mining is accomplished through a series of text analysis and processing that extracts meaningful information using natural language processing (NLP) in unstructured text. Topic modeling is a text mining method that identifies and classifies topics that are latent in a document. Topic modeling inferences topics by clustering words with similar meanings to find topics in a large unstructured document set [12, 13]. The most popular topic modeling technique is the potential probability estimation technique developed by Blei et al. [14]. The LDA assumes that a single document contains multiple topics or that multiple documents can share a common theme. The LDA uses the Dirichlet distribution to calculate the probability that individual documents and words will be included in a particular topic and the probability that individual words derived from the entire document will be included in a particular topic. Figure 1 shows the LDA as a graph model [12, 14]. N is the number of words per document and indicates the length of the document. D represents a set of documents. $W_{d,n}$ represents the n-th word of d document, which is determined by Z and β. Z represents a

topic by word with a per-word topic assignment and is determined by the value θ_d, which is the subject distribution of the document. θ is the weight of each subject obtained using the parameter α and the Dirichlet distribution.

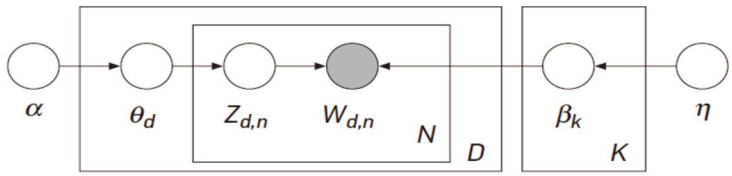


Figure 1. LDA graphical model for Topic modeling [12]

In this study, LDA based topic modeling for abstracts is performed to extract topics, and keywords and dominant documents of each topic are derived by analyzing weights for topics in words and documents.

3. RESEARCH METHODS

3.1 Research Framework

The research framework for this study consists of two parts as shown in Figure 2. First, Topic Modeling derives topics and key words by performing LDA based topic modeling on collected research data. Then, Topic-word Probability Table is created. In the second Topic Class Prediction, topic class is predicted based on topic-word proportions table for the new text.

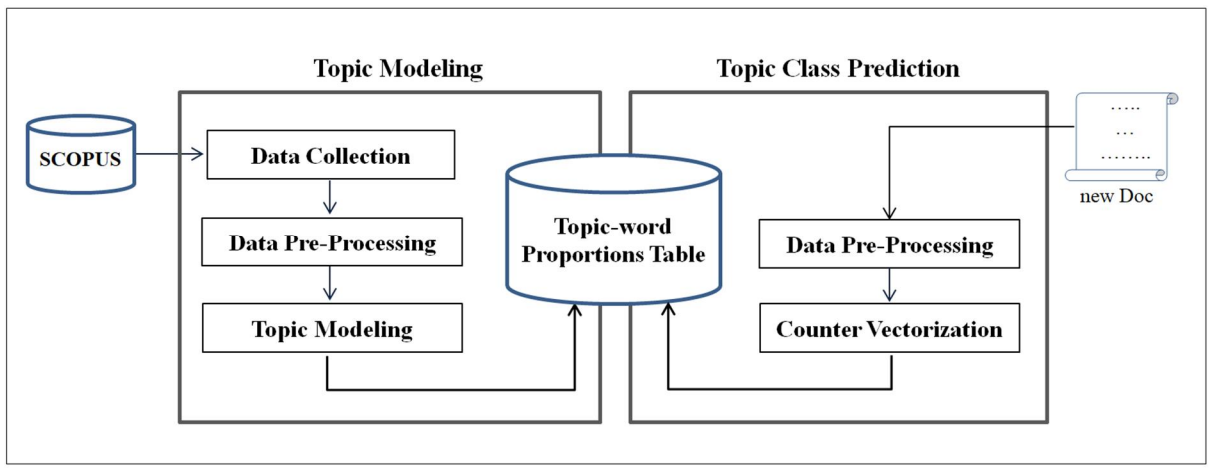


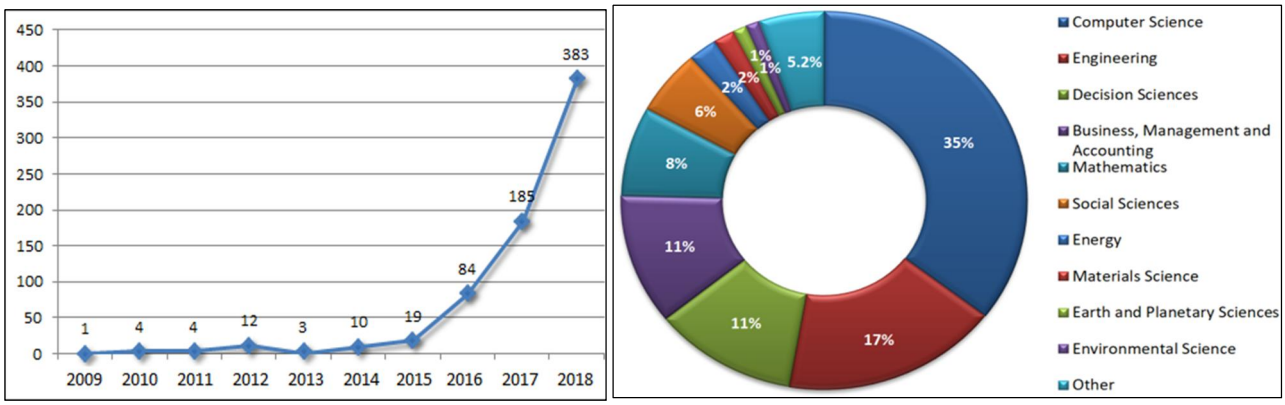
Figure 2. Research Framework

In this study, we used Python 3.6 and Anaconda 3 for text data processing and text mining, and Python gensim package, sklearn package and NLTK package were used.

3.2 Data Collection

In this study, we analyze research trends of digital transformation using abstract texts of academic documents. In order to collect research data, 705 research data were collected for 10 years from 2009 to 2018 on articles and conference papers that include "digital transformation" as keywords in SCOPUS, an academic database. The document numbers were assigned from 0 to 704 in the order collected. Figure 3. (a) shows that the literature has been increasing rapidly since 2016. This shows that the interest in digital transformation has surged as the fourth industrial revolution was discussed in the WEF in 2016. Figure 3. (b) shows the subject area of the journal in which literature using the keyword digital transformation is published. As digital

transformation is a phenomenon caused by evolving digital technology, the Computer Science area is the largest with 35%, and is studied in various areas such as engineering, decision sciences, business, and social science.



(a) documents by year (2009~2018) **(b) documents by subject area**

Figure 3. Data collection results

3.3 Data Pre-processing

We have preprocessed the collected text data for analysis. We converted the words in the document into lowercase letters, and then performed tokenization to separate them into words. We also removed stop words, which are not necessary for analysis. And we performed lemmatization to extract the lemma for words used in various forms in the sentence.

3.4 Topic modeling

To analyze the text, we first performed vector space modeling to generate a counter vector, and constructed a DTM (document term matrix) to perform LDA based topic modeling. To perform LDA based topic modeling, you must determine the number of topics that are hyper parameters. In this study, topic coherence, which is a technique to evaluate the performance of topic modeling, was used to derive the optimal number of topics. Topic coherence is a performance evaluation method proposed by Newman et al. [15]. The better the topic modeling, the more semantically similar words are gathered within the topic, which increases the coherence of the topic. The similarity calculation between words uses PMI (pointwise mutual information) index. The higher the PMI value, the higher the relevance between words [16].

In Equation (1), PMI(w_i, w_j) is calculated by using the probability of word w_i, probability of word w_j and the probability of a word pair (w_i, w_j) appearing at the same time.

$$\mathrm{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{1}$$

In this study, the topic coherence was calculated while the number of topics was varied from 2 to 40, and the number of topics with the highest coherence score was found. As a result, the number of topics in the LDA model was set to 8.

The Topic-Word Proportion Table is generated using the per-word topic proportions obtained from the LDA algorithm. The value of the Topic-Word Proportions Table is the product of per-word topic proportions and topic weight.

4. RESULTS AND DISCUSSION

4.1 Topic Analysis Results

Table 1 summarizes the results of topic modeling. We have compiled five words that show high weight for topic among words of analytical significance, except for the words "digital" and "transformation," which have a high incidence of appearance since we used it as search terms. The three dominant documents with high weight for topic and the probability of inclusion in topic are shown in parentheses. We also summarized three dominant documents with high weight for topic. The number in parentheses is the probability that the document will be included in the topic. The topic labels were selected based on the top words and dominant documents in each topic.

Table 1. Extracted topics for Digital Transformation

Topic #	Top 5 words	Numbers of documents	Top 3 dominant doc #	Topic label
0	innovation, framework, performance, capability, organization	48	doc_602(0.96) doc_306(0.94) doc_263(0.92)	Digital Innovation
1	company, production, project, industry, manufacturing	154	doc_156(0.98) doc_167(0.97) doc_52(0.96)	Digitalization methodology
2	business, model, risk, management, ecosystem	83	doc_118(0.98) doc_222(0.98) doc_373(0.96)	Risk management
3	enterprise, architecture, service, environment, cloud	28	doc_381(0.98) doc_640(0.95) doc_646(0.94)	Adaptable enterprise architectures
4	information, organization, strategy, communication, life	151	doc_480(0.97) doc_351(0.96) doc_387(0.95)	Social impact
5	internet, customer, application, communication, market	15	doc_72(0.97) doc_253(0.96) doc_486(0.63)	Services application
6	service, government, sector, quality, experience	85	doc_232(0.96) doc_631(0.96) doc_525(0.96)	Public Service Transformation
7	industry, economy, design, identify, requirement	141	doc_186(0.98) doc_117(0.96) doc_194(0.95)	digital economy

Topic 0 is a study of the relationship between digital transformation and IT capability, corporate performance and innovation. Topic 1 is a methodology study for digitalization in various industries such as manufacturing, media and entertainment, and education. Topic 2 is a study on the implications of digitization on risk management. Topic 3 is flexible and adaptable service-oriented enterprise architecture mechanisms in the IT environment (micro-services, Internet of Things components, enterprise social networks, and cloud environments, mobility systems, big data and adaptive case management) for supporting digital transformation. Topic 4 is based on the research on the improvement of social sustainability, the guarantee of

social participation and the provision of welfare services by utilizing big data produced through digitalization. Topic 5 is a study on the function and development of service applications which will be provided much more broadly due to digital transformation in advanced network environment such as mobile network, IOT, 5G. Topic 6 is the study of the application of digital transformation to the public service sector in order for e-governance to evolve into smart governance. Topic 7 shows research on regulatory policy, investment environment, block chain technology and so on to activate digital economy.

4.2 Trend Analysis Results

The research trends by topic were analyzed using the frequency of each year 's publications. As Figure 4 shows, we are seeing an increase in research on all topics since 2016. In particular, Topic 7's "digital economy", Topic 4's "Social impact", and Topic 1's "Digitalization methodology" have been analyzed as hot topics.

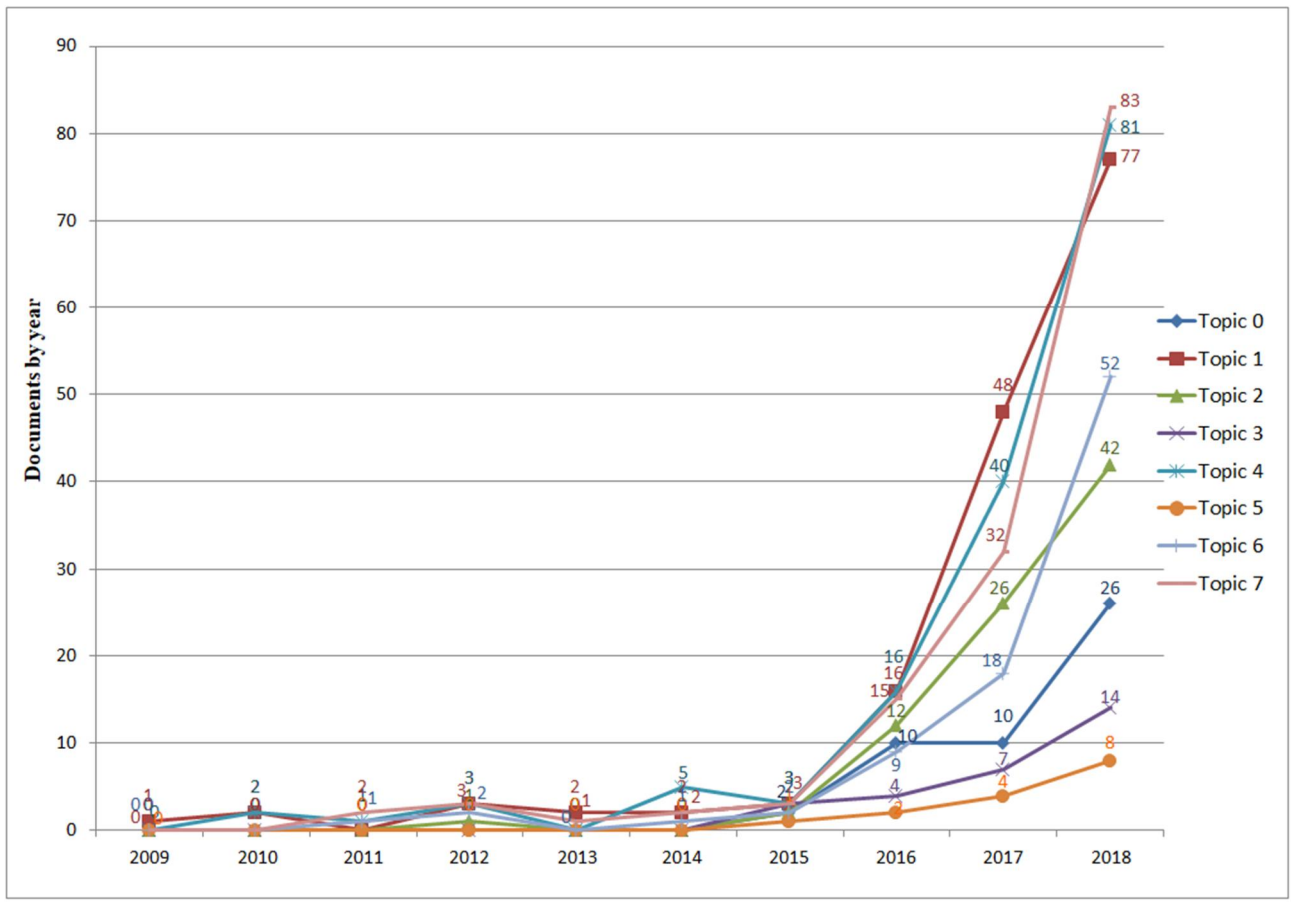


Figure 4. Time trends of 8 topics from 2009 to 2018.

4.3 Topic Class Prediction

For new documents, we can predict the topic for new documents by using the LDA model that we created, without re-doing the whole topic modeling process. If a new document is input, we can derive the keywords and find the most similar topic in the Topic-word Proportions Table generated by using per-word topic proportions and topic weights obtained from LDA based topic modeling.

Table 2. Topic-word Proportions Table

Topic #	activity	application	architecture	…	tool	work
Topic0	11.20995	13.30997	25.47284	…	0.722809	13.21251
Topic1	6.968422	33.04617	0.730657	…	31.30162	53.48893
Topic2	0.215115	0.726899	0.331796	…	10.07057	3.651611
Topic3	0.129047	0.257729	**151.2431**	…	0.135506	3.872123
Topic4	**64.15116**	28.66665	1.842815	…	**46.00912**	33.66612
Topic5	2.926520	**62.28627**	0.187289	…	0.129502	0.416221
Topic6	0.144695	0.975322	0.147194	…	4.153879	18.31956
Topic7	11.33843	25.47284	0.184595	…	0.395038	**73.51638**

5. CONCULISON

This study derives the topic of digital transformation which shows the change of all industrial structure, politics, economy and society as well as IT technology in the fourth industrial revolution era by using Topic modeling method and analyzed the research trend of topic. We proposed a method to predict and classify the topic class without applying the whole topic modeling process again, when analyzing a topic for a new document. This study has limitations that the text used in the research was performed on the abstract, not the original text of the academic literature. In addition, since the LDA-based Topic Modeling is an unsupervised method, there is a limitation in that it can not quantitatively calculate the accuracy of the derived topic.

This study is expected to contribute to the convergence research by deriving a research topic related to digital transformation, which is an important research issue, and suggesting a method of classifying topic class. For example, Topic 7 "Digital Economy" provides research collections related to the digital economy such as technology development, infrastructure development, block chains, policies and regulations, use intentions and success factors, and investment appraisals. We expect that this study will be useful for advanced convergence research.

REFERENCES

[1] Schwab, K. and F., *The fourth industrial revolution*, World Economic Forum, Geneva, Switzerland, 2016

[2] Martin, A. *Digital literacy and the 'digital society'. Digital literacies: Concepts, policies and practices*, 30, pp. 151–176, 2008.

[3] Westerman, G., et al., *Digital Transformation: A roadmap for billion-dollar organizations*, MIT Center for Digital Business and Capgemini Consulting, pp. 1-68, 2011.

[4] Matt, C., T. Hess, and A. Benlian, "Digital transformation strategies," *Business & Information Systems Engineering*, 57(5), pp. 339-343, 2015.
DOI: https://doi.org/10.1007/s12599-015-0401-5

[5] Piccinini, E., R.W. Gregory, and L.M. Kolbe, "Changes in the producer-consumer relationship-towards digital transformation," *Changes*, 3(4), pp. 1634-1648, 2015.

[6] Collin, J., *IT Leadership in Transition: The Impact of Digitalization on Finnish Organizations*, Aalto University School of Science, pp. 29-34, 2015.

[7] Khan, S., *Leadership in the digital age: A study on the effects of digitalisation on top management leadership.* 2016.

[8] Møller, L., et al., "Characterizing digital disruption in the general theory of disruptive innovation," *ISPIM Innovation Symposium*, 2017.

[9] Ershova, T.V. "Digital Transformation Framework Monitoring of Large-Scale Socio-Economic Processess," *2018 Eleventh International Conference Management of large-scale system development(MLSD)*, 2018.

DOI: 10.1109/MLSD.2018.8551765
[10] Feldman, R. and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," *KDD*, 1995.
[11] Allahyari, M., et al., "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.
[12] Blei, D.M., "Probabilistic topic models," *Commun. ACM*, 55(4), pp. 77-84, 2012.
DOI: 10.1145/2133806.2133826
[13] Steyvers, M. and T. Griffiths, *Probabilistic topic models*, Handbook of latent semantic analysis, 427(7), pp. 424-440, 2007.
[14] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, pp. 993-1022, 2003.
[15] Newman, D., et al., "Automatic evaluation of topic coherence," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* , 2010.
[16] Newman, D., S. Karimi, and L. Cavedon, "External evaluation of topic models," *Australasian Doc. Comp. Symp*., 2009.