International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

SUSTAINABLE TECHNOLOGY MINING USING STATISTICAL MODELING

Sunghae Jun

Professor, Department of Big Data and Statistics,
Cheongju University, Chungbuk 28503 Korea

ABSTRACT

Sustainable technology is a technology that can sustain the technological competitiveness of a company continuously. So, it is important to forecast and understand the sustainable technology. This paper deals with statistical modelling for sustainable technology mining. We use association rule mining, social network analysis, and linear regression for constructing statistical model. In this paper, we combine the results of three analytical methods for sustainable technology analysis. To verify the performance and validity of our research, we carry out case study on the technology domain related to artificial intelligence.

Keywords: Social network analysis, Association rule mining, Sustainable technology, Regression, Patent analysis

1. INTRODUCTION

Technology has affected society for many years. At the same time, society has also had a great impact on technological change and development (Roper, et al., 2011). Technology and society have been evolving together, influencing each other. It is very important for us to analyze and understand the technology. So, many studies on technology analysis have been published in diverse fields (Choi, et al., 2015; Choi, et al., 2016; Kim, et al., 2015; Kim, et al., 2018; Lee, et al., 2018). They were based on two approaches, qualitative and quantitative methods. Delphi is one of the most popular qualitative methods for technology analysis (Hung, et al., 2012). In the Delphi process, we conduct questionnaire surveys to a group of experts. In order to identify the future prospects for a particular technology, we investigate the various opinions of the experts and repeat it several times to draw conclusions from them. This approach to technology forecasting can be subjective. This is because the results of the technology analysis depend on the subjective experience and knowledge of the experts. In order to solve such problems, researches on quantitative technology analysis have been actively conducted recently (Park, et al., 2015; Park and Jun, 2017a; Park and Jun, 2017b; Uhm, et al., 2017). They used various

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

objective data including the results of developed technology such as patents and papers. Patent analysis is one of most popular approaches to technology forecasting. This is made up of statistics and machine learning algorithms to build the quantitative models for technology forecasting. Technology forecasting is one of main tasks in the management of technology (MOT) (Roper, et al., 2011). Most companies are paying much attention to the MOT in order to strengthen their technological competitiveness in the market. As a result of MOT, companies develop new products and lead technological innovation. In addition, many companies are investing heavily in sustainable technology development for their continued development. Sustainable technology means technology that can sustain and expand the competitiveness of the company (Jun, 2018). Many companies want to know about sustainable technologies that can drive their technology in the future. Therefore, in this book chapter, we introduce the forecasting method for sustainable technology. Our method consists of various techniques based on statistical analysis and machine learning. Also we use patent documents as objective data for sustainable technology forecasting. In our case study, we consider artificial intelligence (AI) technology to illustrate how our methodology could be applied to real domain. This book chapter is organized as follows. We propose a forecasting method for sustainable technology in section 2. In section 3, we show a case study and its application in real domain. Lastly, we illustrate our conclusions in section 4.

2. STATISTICAL MODELING FOR SUSTAINABLE TECHNOLOGY MINING

We introduce a methodology of sustainable technology mining. The methodology is composed of three methods based on statistical modeling. First, we select the candidate keywords from the structured patent data using association rule mining (ARM). The ARM is made up of support, confidence, and lift measures for extracting meaningful rules. The support for two keywords A and B is defined as follows (Han, et al., 2012).

$$\text{Support}(A \rightarrow B) = P(A \cap B)$$

This means that the probability of intersection of A and B with all outcomes from both in A and B (Ross, 2017). The larger this probability value, the greater the chance that A and B will occur at the same time. Next we find the association between antecedent (left hand side, A) and consequent (right hand side, B) using the confidence as follow (Han, et al., 2012).

$$\text{Confidence}(A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)}$$

This is the conditional probability of B given A representing the influence of A on B. To develop the technology based on keyword A, we need to develop the technology of keyword A first. Lift is a measure of correlation between keywords A and B as follow (Han, et al., 2012).

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

$$\text{Lift}(A \rightarrow B) = \frac{P(A \cap B)}{P(A)P(B)}$$

This is not a probability. If the lift value is one, the technologies based on keywords A and B are independent each other. In addition, if the value is less than one, the keywords A and B are negatively correlated. The other way, when the lift value is larger than one, the keyword A is positively correlated with the keyword B.

Second, we use the social network analysis (SNA) to find the core keywords from all technology keywords in given technology field. The SNA consists of node and edge in the graph data structure (Scott, 2012; Main and Savitch, 2010). Each node represents a keyword with sub technology in target technology domain. From the visualization of SNA, we find the central nodes (keywords) representing the association of other keywords. Finally, we construct the multiple regression model for technology analysis. The central keywords become the dependent variables in our regression model, and other keywords connecting the central keywords are independent variables as follow.

$$\text{dependent keyword} = \text{f}(\text{independent keywords}) + e$$

where e is error with mean=0 and variance=σ^2. We combine the results of ARM, SNA, and regression models for understand and forecast sustainable technology of given technology filed. Next, we perform case study on artificial intelligence (AI) technology to illustrate how our proposed methodology could be applied to real fields.

3. CASE STUDY

We selected AI technology as one of practical domains for our methodology. AI technology is one of the areas where sustainable technology is essential for humanity in the future. We retrieved the patent documents and journal papers related to AI technology from 2014 to 2016. The keyword search expression used to collect patents and papers related to AI technology is as follows.

(((Artificial OR machine OR reinforce OR supervised OR deep OR unsupervised) AND (learn OR intelligence) OR ((neural OR neuron) AND (network OR feedback))) AND AD=2014:2016

We searched the patents from patent databases of WIPSON and the papers from Scopus database of Elsevier (Elsevier, 2018; WIPSON, 2018). We collected 5,358 patents and 6,000 papers, pre-processed them, and finally used 10,343 patent and paper documents. In this case study, we extracted 104 keywords from the data as follows; acting, agent, algorithm, analysis, appearance, application, approximate, architecture, art, artificial, awareness, Bayesian, behavior, belief, brain,

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

cognitive, collaborative, communication, component, computer, computing, context, conversation, corpus, data, decision, deep, design, dialogue, ensemble, expert, extraction, feedback, figure, framework, game, generalization, grammar, graph, ground, hierarchical, human, image, independent, inference, information, intelligence, interface, interpretation, knowledge, language, learning, life, logic, machine, mental, mind, mining, modeling, morphological, moving, multilingual, natural, nature, network, neural, neuro, object, ontology, operation, optimal, pattern, perceiving, perception, planning, policy, probability, processing, programming, reasoning, recognition, reinforcement, representation, restoration, retrieval, semantic, sentence, spatial, speaker, speech, state, statistics, stochastic, synthesis, system, text, time, tracking, translation, uncertain, understanding, video, vision, and voice. In this process of keyword extraction, we used the R data language and its 'tm' package (Feinerer et al., 2008; Feinerer and Hornik, 2018; R Development Core Team, 2018). To select meaningful keywords, we carried out ARM using the AI keyword data. In this case study, we used 'arules' R package to compute support, confidence, and lift in ARM (Hahsler et al., 2018). Table 1 shows the association rules with confidence of 0.9 or higher.

Table 1: Association rule mining result (confidence >= 0.9)

Antecedent	Consequent	Support	Confidence	Lift
{neural,speech}	{network}	0.0102	0.9725	3.2593
{algorithm,art}	{state}	0.0239	0.9724	8.9245
{image,neural}	{network}	0.0144	0.9675	3.2428
{algorithm,art,data}	{state}	0.0113	0.9669	8.8741
{art,data}	{state}	0.0243	0.9544	8.7587
{art}	{state}	0.0513	0.9499	8.7178
{network,speech}	{neural}	0.0102	0.9464	4.4821
{neural,system,time}	{network}	0.0145	0.9434	3.1619
{deep,neural}	{network}	0.0204	0.9336	3.1291
{art,system}	{state}	0.0127	0.9291	8.5266
{neural}	{network}	0.1957	0.9267	3.1061
{neural,time}	{network}	0.0336	0.9206	3.0856
{neural,system}	{network}	0.0782	0.9204	3.0847

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

{algorithm,neural,system}	{network}	0.0160	0.9171	3.0738
{design,neural}	{network}	0.0170	0.9167	3.0723
{neural,object}	{network}	0.0165	0.9144	3.0648
{data,neural}	{network}	0.0785	0.9124	3.0579
{algorithm,neural}	{network}	0.0366	0.9111	3.0535
{image,network}	{neural}	0.0144	0.9085	4.3027
{data,neural,system}	{network}	0.0324	0.9054	3.0345

The rule with the greatest confidence (0.9725) is as follows.

$$\{\text{neural}, \text{speech}\} \rightarrow \{\text{network}\}$$

The support and lift are 0.0102 and 3.2593 respectively. To develop the technology related to network, we must develop the technologies related to neural and speech. So, the technology based on this keyword (network) is important to AI technology. Also, we can consider the keyword of network as sustainable area for AI technology. Table 2 illustrates the association rules with confidence between 0.6 and 0.9.

Table 2: Association rule mining result (0.6 <= confidence < 0.9)

Antecedent	Consequent	Support	Confidence	Lift
{data,neural,time}	{network}	0.0153	0.8927	2.9918
{algorithm,data,neural}	{network}	0.0176	0.8835	2.9611
{neural,state}	{network}	0.0170	0.8800	2.9494
{neural,pattern}	{network}	0.0122	0.8690	2.9124
{deep,network}	{neural}	0.0204	0.8147	3.8581
{algorithm,data,state}	{art}	0.0113	0.6802	12.5861
{network,object}	{neural}	0.0165	0.6759	3.2009
{network,pattern}	{neural}	0.0122	0.6667	3.1572
{network}	{neural}	0.1957	0.6559	3.1061

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

{network,state}	{neural}	0.0170	0.6494	3.0757
{deep}	{network}	0.0250	0.6475	2.1702
{algorithm,state}	{art}	0.0239	0.6399	11.8398
{network,system}	{neural}	0.0782	0.6370	3.0167
{network,time}	{neural}	0.0336	0.6237	2.9535
{data,network}	{neural}	0.0785	0.6222	2.9467
{expert}	{system}	0.0170	0.6219	1.5392
{analysis,system}	{data}	0.0122	0.6176	1.4629
{data,network,time}	{neural}	0.0153	0.6172	2.9229
{algorithm,network,system}	{neural}	0.0160	0.6081	2.8796
{data,network,system}	{neural}	0.0324	0.6080	2.8793
{network,system,time}	{neural}	0.0145	0.6073	2.8760

In this table, we can find many consequents are related to the keyword of neural. Of course, there are many network keywords in the consequent. Table 3 shows the association rules whose confidence values are between 0.5 and 0.6.

Table 3: Association rule mining result (0.5 <= confidence < 0.6)

Antecedent	Consequent	Support	Confidence	Lift
{analysis}	{data}	0.0278	0.5890	1.3949
{spatial}	{data}	0.0174	0.5863	1.3887
{algorithm,data,network}	{neural}	0.0176	0.5833	2.7626
{decision}	{system}	0.0145	0.5792	1.4334
{pattern,system}	{data}	0.0165	0.5758	1.3637
{design,network}	{neural}	0.0170	0.5733	2.7150
{algorithm,learning}	{data}	0.0152	0.5688	1.3473
{algorithm,network}	{neural}	0.0366	0.5674	2.6869

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

{deep}	{neural}	0.0219	0.5650	2.6757
{algorithm,pattern}	{data}	0.0128	0.5523	1.3081
{machine}	{data}	0.0291	0.5356	1.2685
{pattern}	{data}	0.0366	0.5346	1.2661
{speech}	{network}	0.0108	0.5333	1.7875
{learning}	{data}	0.0416	0.5296	1.2542
{learning,system}	{data}	0.0135	0.5283	1.2513
{logic}	{system}	0.0167	0.5242	1.2975
{speech}	{neural}	0.0105	0.5190	2.4581
{video}	{system}	0.0128	0.5156	1.2762
{machine,system}	{data}	0.0125	0.5119	1.2124
{design,network}	{system}	0.0152	0.5114	1.2657
{processing}	{data}	0.0116	0.5085	1.2043
{algorithm,framework}	{data}	0.0123	0.5060	1.1984
{feedback}	{system}	0.0131	0.5056	1.2514

From the results of Tables 1, 2, and 3, we confirmed that the keyword network is the most important keyword for AI technology from a sustainable technology perspective. We also selected 25 keywords from the results of the association rule as follows; algorithm, analysis, art, data, decision, deep, design, expert, feedback, framework, image, learning, logic, machine, network, neural, object, pattern, processing, spatial, speech, state, system, time, and video. These keywords are keywords that appear more than once in association rules with a confidence level greater than 0.5. Next, we performed the SNA using the 25 keywords from the previous results. We used R 'sna' package for carrying out our SNA (Butts, 2008; Butts, 2018).

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

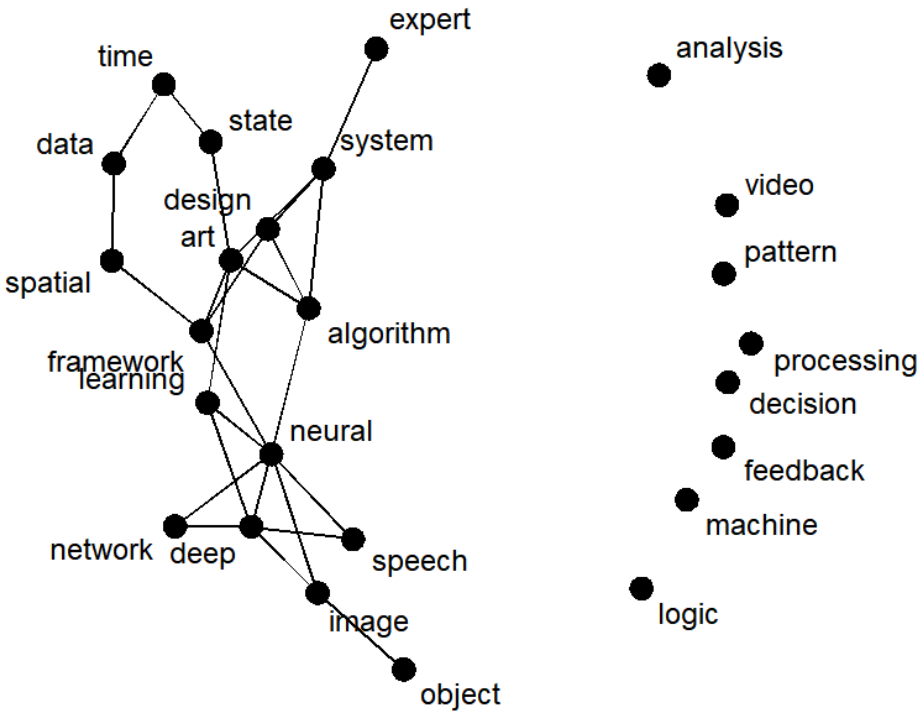

Fig. 1: SNA graph

We have found that some keywords are linked to many other keywords such as art, algorithm, neural, etc. On the other hand, some keywords are isolated alone such as analysis, video, pattern, processing, decision, feedback, machine, and logic. In the SNA graph, the more the keyword is linked to the other keywords, the more the keyword becomes the central keyword. So we investigated this graph around the keyword of network. The keyword of network is connected to deep and neural keywords. Furthermore, the keyword of neural is linked to algorithm, learning, deep, and speech. The keyword of deep is connected to image, speech, and neural. Also, the keyword of speech is connected to deep and neural at the same time. The results of ARM and SNA, we constructed the following diagram of AI technology.

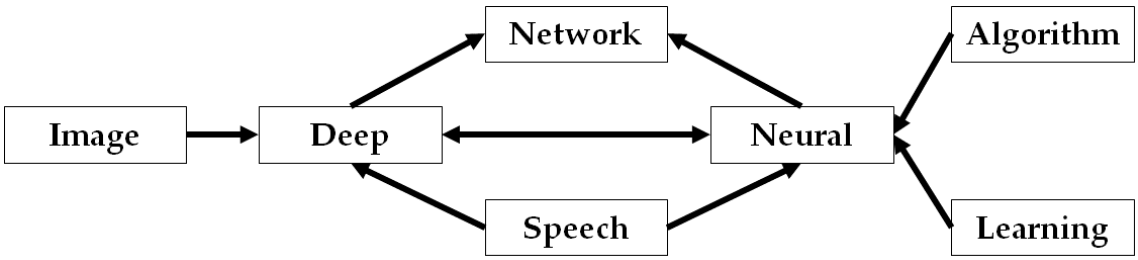

Fig. 2: AI technology diagram

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

We found that the two technologies based on deep and neural are influencing each other. Also, the technologies based on algorithm and learning affect the technology of neural, and the technology of speech influences on the technologies related to deep and neural. The technology of image affects the technology of deep, and finally the technologies based on deep and neural affect the technology of network. From the results of ARM, SNA, and AI technology diagram, we performed three regression models as follow.

Table 4: Regression result of AI keywords

Dependent	Independent	Estimate	P-value
Network	Deep	0.4341	0.0001
	Neural	0.8202	0.0001
	Algorithm	-0.0653	0.0001
Neural	Deep	0.3926	0.0001
	Learning	-0.1924	0.0001
	Speech	0.1604	0.0001
	Image	0.0904	0.0001
Deep	Neural	0.0521	0.0001
	Speech	0.0570	0.0001

First regression model consists of dependent variable of network and independent variables of deep and neural. All independent variables are statistically significant, because their probability values (p-values) are less than 0.01. Other regression models also have all significantly independent variables for dependent variables of neural and deep.

4. CONCLUSION

In this paper, we studied on a statistical modeling for sustainable technology mining. We considered ARM, SNA, and regression models to perform the proposed modeling. We found the sustainable technology rules from the associations between the technological keywords extracted from the patent documents. In our case study, we selected AI technology as target domain for the proposed method. From the result of AI case study, we found that the technologies based on network, neural, algorithm, deep, learning, speech, and image are sustainable technology for AI. Therefore, we concluded that the technologies of network, neural, and deep are the technologies resulting from the end step of AI. In addition, the technologies of other keywords are the antecedent technologies for AI.

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

Our research contributes to find sustainable technologies in diverse technology fields such as bio, internet of things, big data, etc. In addition, using the proposed method, many companies are able to build research and development (R&D) strategy efficiently in technology management. In our future works, we will consider more advanced models such as convolutional neural network and recurrent neural network for sustainable technology mining.

REFERENCES

Butts, C. T. (2008) "Social Network Analysis with sna," Journal of Statistical Software, 24(6), 1-51.

Butts, C. T. (2018) Tools for Social Network Analysis-Package sna. CRAN R-project.

Choi, J., Jang, D., Jun, S., and Park, S. (2015) "A Predictive Model of Technology Transfer using Patent Analysis," Sustainability, 7(12), 16175-16195.

Choi, J., Jun, S., and Park, S. (2016) "A Patent Analysis for Sustainable Technology Management," Sustainability, 8(8), 688.

Elsevier, (2018) Scopus Journal Searching, https://www.elsevier.com/solutions/scopus

Feinerer, I., Hornik, K., and Meyer, D. (2008) "Text mining infrastructure in R," Journal of Statistical Software, 25(5), 1-54.

Feinerer, I., and Hornik, K. (2018) Package 'tm' Ver. 0.7-4, Text Mining Package, CRAN of R project,

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., and Borgelt, C. (2018) Package 'arules', Mining Association Rules and Frequent Itemsets, CRAN R-project.

Han, J., Kamber, M., and Pei, J. (2012) Data Mining: Concepts and Techniques, Third Edition, Waltham, MA, Morgan Kaufmann.

Hung, C., Lee, W., and Wang, D. (2012) "Strategic foresight using a modified Delphi with end-user participation: A case study of the iPad's impact on Taiwan's PC ecosystem", Technological Forecasting & Social Change, 80, 485-497.

Jun, S. (2018) "Bayesian Count Data Modeling for Finding Technological Sustainability," Sustainability, 10(9), 3220.

Kim, S., Jang, D., Jun, S., and Park, S. (2015) "A Novel Forecasting Methodology for Sustainable Management of Defense Technology," Sustainability, 7(12), 16720-16736.

ISSN: 2455-8834

Volume:03, Issue:12 "December 2018"

Kim, J., Jun, S., Jang, D., and Park, S. (2018) "Sustainable Technology Analysis of Artificial Intelligence Using Bayesian and Social Network Models," Sustainability, 10(1), 115.

Lee, J., Kang, J., Jun, S., Lim, H., Jang, D., Park, S. (2018) "Ensemble Modeling for Sustainable Technology Transfer," Sustainability, 10(7), 2278.

Main, M., and Savitch, W. (2010) Data Structures and Other Objects Using C++, 4th Edition, Prentice Hall.

Park, S., Lee, S., and Jun, S. (2015) "A Network Analysis Model for Selecting Sustainable Technology," Sustainability, 7(10), 13126-13141.

Park, S., and Jun, S. (2017a) "Technology Analysis of Global Smart Light Emitting Diode (LED) Development Using Patent Data," Sustainability, 9, 1363.

Park, S., and Jun, S. (2017b) "Statistical Technology Analysis for Competitive Sustainability of Three Dimensional Printing," Sustainability, 9, 1142.

R Development Core Team (2018) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Roper, A. T., Cunningham, S. W., Porter, A. L., Mason, T. W., Rossini F. A., and Banks J. (2011) Forecasting and Management of Technology, Hoboken, NJ, John Wiley & Sons.

Ross, S. M. (2017) Introductory Statistics, Fourth Edition, London, UK, Academic Press Elsevier.

Scott, J. G. (2012) Social Network Analysis, third edition, London, UK, SAGE.

Uhm, D., Ryu, J., and Jun, S. (2017) "An Interval Estimation Method of Patent Keyword Data for Sustainable Technology Forecasting," Sustainability, 9(11), 2025.

WIPSON. (2018) WIPS Corporation, http://www.wipson.com.

