A Data Modelling and Visual Analysis of Science and Technology Policy Keyword

Seung Su Chun
Division of Information & Knowledge
Korea Institute of Science and Technology Evaluation and Planning
Seoul, South Korea
dabins@kistep.re.kr

***Abstract*— Recently, enterprises and major countries are increasing their R&D investment for economic growth through science and technology upbringing. As the size of investment increases, it becomes more difficult to understand and analyze the status and performance of detailed projects, programs and technologies. This paper deals with a method of analyzing the relationship between major technological keywords by text mining large scale policy documents for R&D investment. This methodology helps us better understand the relationship between the complicated policy/technology and the investment flow, and support the analysis with a visualization of the relationship.**

Keywords-technology policy; data mining; network model; visual analysis

I. DATA PROPERTY OF SCIENCE AND TECHNOLOGY POLICY ANAYSIS

In today's R&D investment, it is essential to understand the relationship between investment structure and technology. Analysis of investment policy in research and development is essential for understanding the relationship between detailed programs and technology. In addition, to coordinate investment in R&D, it is necessary to understand the relationship between technologies as well as investment flows. Recently, a network analysis methodology that can grasp the relationships between various entities through intuitive visualization and apply sophisticated analytical methods has been suggested as a useful approach. Network analysis is one of the Business Intelligence analytical methods for strategy establishment, rapid decision making and risk management. It can model meaningful data by networking complicated data. The development in technology and the maturity of an information-knowledge based society accelerate technological competition among enterprises and countries [1]. These sophisticated systems are designed to find useful information from the large data repositories and developed to be used quickly in decision making. The availability of data, such as big data systems, has become widespread, and researches have been actively conducted on visualization techniques for computational thinking and rapid decision support [2]. The progress of intelligent systems through relational meaning and interpretation by structuring and modeling analysis of data and information is being promoted in various fields. Here, we have demonstrated the application of data structuring methods and visualization techniques for the analysis of unstructured data statistics on R&D policies.

II. VISUAL ANALYSIS OF NETWORK MODELING

In the past days, investment analysis used to be based on business classification. But it is important to analyze the linkage between business and technology in convergence-oriented R&D, and the visual analysis becomes very useful. Network modeling is used as a key process for network analysis visualization by conceptually defining the entities that are components of a network map and the relationships between them. The action of policy production, refinement, evidence, and diffusion can be seen as a composition of content types such as knowledge, people, and activity. Complex policy documents can be classified into people, services, processes and resources, and reconstructed into hierarchical networks [3]. Firstly, we have organized the policy document as in Table 1 and extracted the keywords from the contents of the document by the text mining technique. We have visualized the relationship between keywords as shown in Figure 1 by analyzing the proximity (distance) and relevance (frequency) between the keywords in the context and sentences in the document.

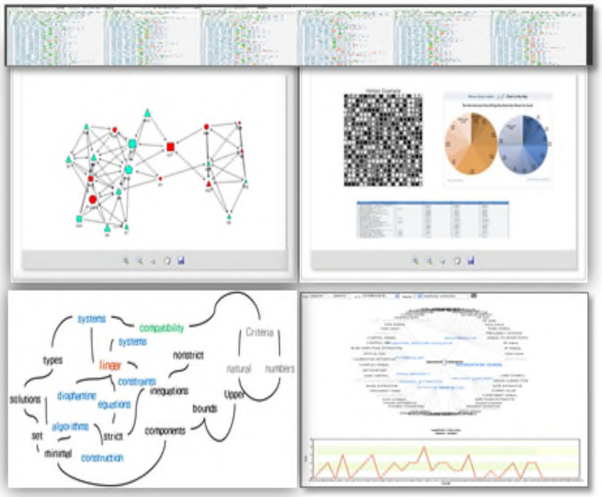
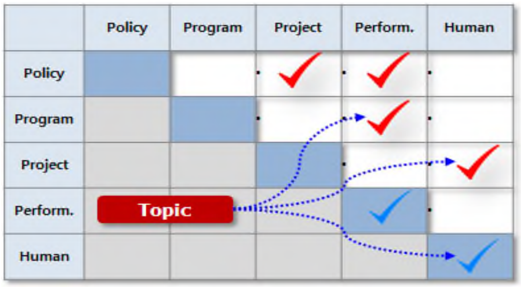


Figure 1. Example of keyword based network analysis

TABLE I. META MATRIX OF POLICY KEYWORD



	Policy	Program	Project	Perform.	Human
Policy			✓	✓	
Program				✓	
Project					✓
Perform.	Topic			✓	
Human					✓

Table 1 shows the document classification system for R&D investment, which is used as the attribute value of keyword and relation. The document classification in Table 1, such as the policy decision process, is used to analyze the frequency changes and topic relationships of specific keyword topics extracted through text mining. The data of keywords identified as entities (nodes, points) that can generate a network map can be classified into five types: policy, program, project, performance, and human. The extracted keywords can represent the relationship between the categories with the document classification of the meta-metrics as an attribute. The same keywords extracted from the meta matrix documents express the relationship between different categories and visualize them by constructing main topics between the categories [4][5]. The keyword based network model is important to understand the decision flow for a specific technology because decision-making in R&D investment, establishment of strategies, and evaluation and adjustment of R&D programs are all sequentially processed.

III. POLICY TOPIC DETECTION AND MODEL ANALYSIS

The aim of this paper is to analyze and visualize the analytical relationship between keywords using the existing business classification method, which employs text mining. This section describes how to calculate the relationship between policy documents and keyword topics. In order to understand the implied meaning of knowledge, the interaction and effects between topics should be analyzed in-depth as they are the meaning unit of knowledge [6]. In the analysis of the relationship between keywords, we use the Celton index method for properties such as connectivity between nodes and proximity. In this study, we define the process (*P*) document of classification as Table 1 and analyze the relevance of the extracted keywords for the Selton index. The keyword-based network model was used for efficiency analysis of the technology topic. At first, the intensity of relations may be quantified based on the connection intensity between certain topics. P_i is the number of i in a certain topic, and P_j is the number of j in a certain topic, and P_{ij} is the number of i and j in a certain topic. If the intensity of relations is r, and r is calculated as the distance between the context and the frequency of appearing at the same time between different topics, then:

$$r = \frac{P_{ij}}{\sqrt{P_i P_j}}$$

The connection relations are divided into degree, closeness and proximity, and have absolute value and relative value. The Degree Centrality (C) of having the most nodes between topics has a higher connection intensity when certain topics are revealed in other major areas of knowledge, and it has a central role in this area. Distance refers to the inter-context distance between any keyword topics. The Closeness Centrality has a central role when certain topics are close to other topics, and distance (D) D_{ij} is the shortest course which connects between topics i and j, and g is the number of the overall node. In the network, the extracted keyword topic is a node. When the relation value r between the topics is high, the edge is constituted. The Between Centrality, where different networks are connected, plays a mediating role in the relation and is necessary for interpreting relationships of different topics. g_{jk} is the shortest course number between two topics (j and k), $g_{jk}(i)$ is the frequency of passing topic i between two topics j and k($j \neq k$), and g is the number of the overall node. Through centrality analysis between topics, we can interpret meaningful characteristics of the knowledge model, and expand the range with difference and connection between knowledge models.

IV. CONCLUSIONS

In this paper, we have described the method of visualizing the relationship of keywords extracted through text mining in the policy document for technology investment and analyzing them by network. This topic network visualization helps to understand the relationships between specific keywords easily and helps interpret the structure of relations with document classification according to the policy process. We have also demonstrated a method for interpreting the meaning according to the network structure and the distance between keyword topics. In future work, we will apply this technique to a large amount of data and complex documents, and also develop a tool to visually analyze the relational path between specific keyword topics.

REFERENCES

[1] G. Yezersky, “General Theory of Innovation, International Federation for Information Processing,” Vol.250, Springer, 2007, pp. 45-55.

[2] N. L. Rovira, “the future of computer aided innovation, IFIP International Federation for Information Processing,” Springer, Vol.277, 2008, pp. 3-4.

[3] C. Heitmeyer, “Using abstraction and model checking to detect safety violations in requirements specifications,” IEEE Transactions on Software Engineering, vol. 24, no. 11, 1998.

[4] W. Chan, “Temporal Logic Queries,” Proceedings of CAV 2000, Springer, LNCS 1855, 2000.

[5] X. Wang and T. Vitvar, “WSMO-PA: Formal Specification of Public administration service model on Semantic Web Service Ontology,” IEEE ICSS’07, 2007.

[6] W3C, “Web Service Description Language (WSDL),” ver.2.0 recommendation, 2007.