Chiba Institute of Technology

Doctoral Dissertation

Methods based on data analysis for assisting selection of success factors contributing to product planning and development strategy

（製品計画・開発の戦略に寄与するデータ分析に基づく成功要因の選択支援手法の研究）

March, 2021

Yutaka Iwakami

Abstract

In product planning and development strategy in enterprises, it is required to surely create value according to the resources (budget and human resources) invested. However, since there are various success factors, it is difficult for each enterprise, organization, or individual to find the optimum solution based on empirical rules. Therefore, efforts have been made to systematize the factors in product planning and development strategy from a scientific point of view. In recent years, there are many approaches to analyze factors by dividing them into roles called KGI and KPI. KGI (Key Goal Indicator) is a target indicator, for example, product sales, market share, etc. KPI (Key Performance Indicator) is an intermediate indictor to achieve KGI, for example, customer reaction to prototypes, etc. In this approach, the main purpose is to search for KPIs that have a large impact on KGIs. For example, factor analysis and its advanced form, SEM (covariance structure analysis), are effective methods for clarifying hidden factors common to KGIs and KPIs and understanding the relationships among them.

On the other hand, in today's rapidly changing business environment, it is important not only to understand the relationships among factors, but also to select appropriate factors according to given conditions.

Therefore, in this study, the author applied Bayesian network to the performance survey data of project planning and development strategy in various enterprises. Bayesian network is a graphical probabilistic inference model that regards each factor as a random variable and draws it as a node (oval) and also expresses their relationships by edges (arrows) among them derived from conditional probability.

As a result, some new insights were obtained, such that efforts to increase the contribution of products to the natural environment will work positively from the perspective of collaboration with external organizations or human resource development. But in actual business, it is required not only to obtain knowledge as unexpected awareness, but also to properly select KPIs in internal planning and keywords in external appeal satisfying the given conditions such as, "What should retailers do to increase product sales?" or "Does the relationship between the contribution of products to the environment and the number of sales differ depending on the scale of annual sales?". However, when attempting to select such KPIs and keywords, the following issues arise.

Issue 1: business type and annual sales are not properly reflected within analysis result when they are merely added as factors into a normal Bayesian network.

Issue 2: When focusing on specific factors, the partial structure of a normal Bayesian network may not match the actual domain knowledge.

Issue 3: It is difficult to determine what keywords are best for appealing in order to enhance the effect of the factors to be focused on.

In this study, the author proposes the methods to solve these three issues, with optimized Bayesian network leveraging LDA (Latent Dirichlet Allocation), Random forest, and keyword expanding and classifying by combination of Word2Vec and Hierarchical clustering.

Finally, the effectiveness of the proposed method in this study was confirmed by systematizing these three methods as a series of processes and applying them to two model cases. In this way, this study assists selection of success factors required in rapid and precise product planning and development strategy. In addition, since each method of this study can be automated by programming, in the future it will be possible to realize a system that supports enterprises to select factors by themselves in product planning and development strategy.

Table of Contents

1. Introduction

1-1. Background

KPIs/KGIs are not only studied for product planning and development strategy in enterprises (1-1:Sandner, 2011; 1-2:Petersen, 2018), but also in more extensive and diverse areas such as environmental preservation and energy policy (1-3:Desjeux et al., 2015; 1-4:Stern, 1999). In order to understand relationships among KPIs/KGIs, various studies have been conducted. For example, a framework of energy, environment, and economy has been critically investigated and analyzed from the concept of technology learning (1-5:Kahouli-Brahmi, 2008).

Another method was based on efficiency and productivity (1-6:Kuosmanen et al., 2013). Also, a method was proposed for achieving a balance between environmental values and economic values in an enterprise based on case studies of three textile factories in Sri Lanka (1-7:Shiwanthi et al., 2018).

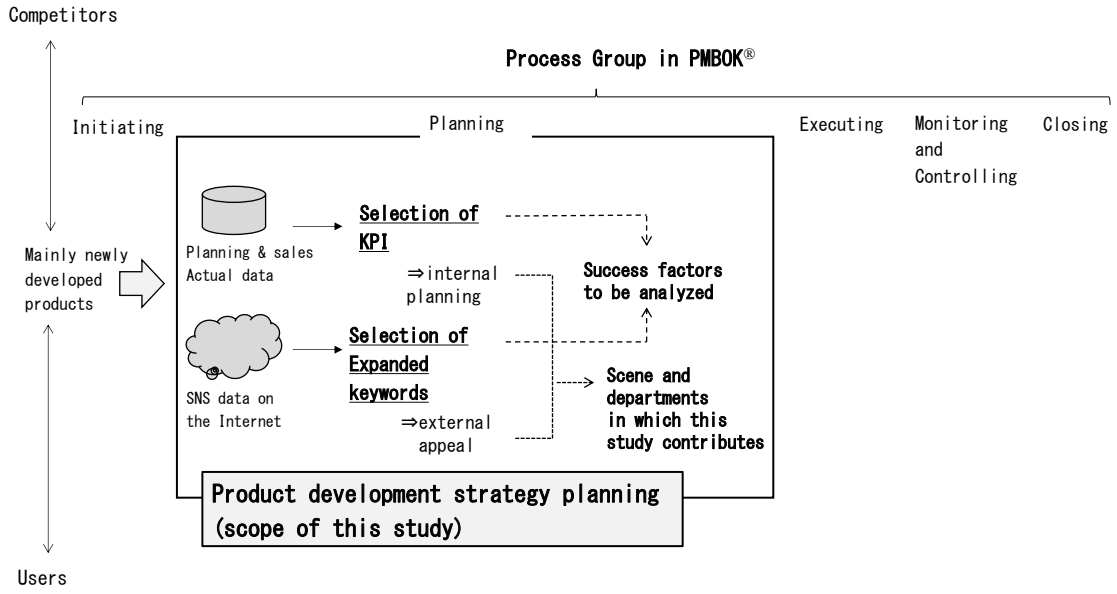
As product planning and development is greatly influenced by various factors (1-8:Chandy & Tellis, 1998), it was not easy to understand how to select KPIs for improving KGIs and few studies have dealt with these issues mathematically (1-9:Krishnan & Ulrich, 2001). Furthermore, not only KPIs in internal planning bur also keywords for external appeal should be properly selected to make product planning and development strategy success in the current business environment.

Therefore, some progressive methods have been awaited in order to assist enterprise to select success factors such as KPIs and appealing keywords in product planning and development strategy.

1-2. Purpose of This Study

Based on the background above, purpose of this study is to provide methods to select success factors such as KPIs for internal planning and keywords on SNS for external appeal in mainly new product planning and development strategy based on the data obtained from actual performance or on the Internet. Figure1-1 shows the scope of this study with the position in the process group of PMBOK®.

Figure 1-1. Scope of this study

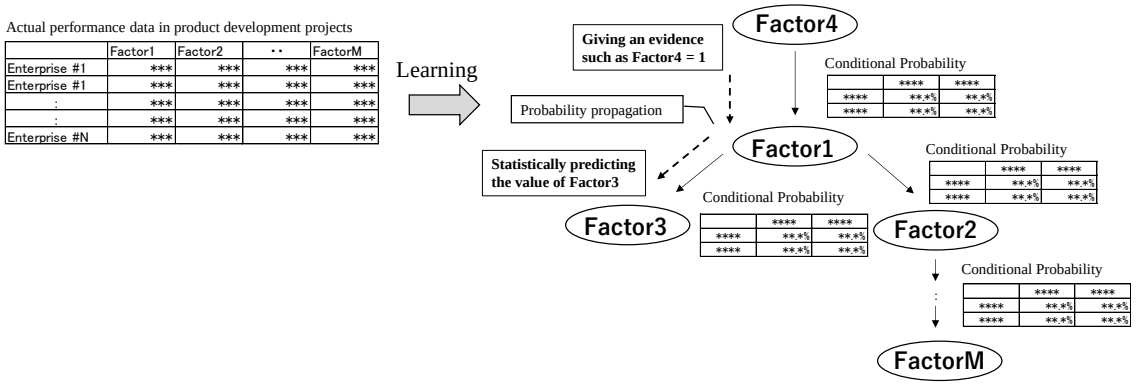


1-3. Visualization and Probabilistic Inference with Bayesian network

Various studies such as factor analysis and its advanced form SEM (Structural Equation Modeling) have been conducted as methods for understanding the relationships among factors. However, in order to accomplish the purpose of this study, it is required not only to understand the relationships among factors, but also to predict the change of them. For this reason, this study leverages Bayesian network.

Bayesian network is one of the analytical methods to which Bayesian statistics is applied. By applying Bayesian network analysis, factors observed as events within actual performance data in product planning and development projects are considered as random variables and are represented as nodes. Directed edges are also drawn among them. Each edge has a conditional probability, which is defined for the two nodes connected by it according with its direction. Such a graph structure is constructed by learning the data. (1-10:Pearl, 1995). Bayesian network also enables probabilistic inference such as how the change of one node affects other nodes. This is because the change is transmitted via edges to other nodes. This is called “probability propagation” or “belief propagation” (1-11:Yedidia et al., 2001). In this study, the description of “a node affects other nodes” means that the change of a node influences other nodes by probability propagation. There are several studies which utilize probabilistic inference between events to solve the problems involving complex factors (1-12:Anderson et al., 2004; 1-13:Ishizaka & Labib, 2013; 1-14:Nadkarni & Shenoy, 2001; 1-15:Nguyen, 2015; 1-16:Rahman & Ripon, 2014). Based on these features and advantages, also in this study, Bayesian network is used as a fundamental tool for visualizing the relationships among factors and for predicting the changes of them.

Figure 1-2. Basic Concept of Bayesian network



1-4. Proposed Methods

Following normal process of Bayesian network analysis, also in this study, actual performance data of product planning and development projects is firstly applied. And some new insights were obtained, such that efforts to increase the contribution of products to the natural environment will work positively from the perspective of collaboration with external organizations or human resource development. [Step1]

However, when trying to achieve the purpose in Section 1-2, following three issues arise.

Issue 1: business type and annual sales are not properly reflected within analysis result when they are merely added as factors into a normal Bayesian network.

Issue 2: When focusing on specific factors, the partial structure of a normal Bayesian network may not match the actual domain knowledge.

Issue 3: It is difficult to determine what keywords are best for appealing in order to enhance the effect of the factors to be focused on.

In this study, the author proposes the following solving methods to each of these three issues.

Method 1: Reflecting business type and annual sales by introducing the topic nodes obtained via LDA (Latent Dirichlet Allocation) into Bayesian network [Step2]

Method 2: Adjusting the structure of Bayesian network by initializing edges according to the result of Random forest [Step3]

Method 3: Expanding keywords according with comparison via the result of Word2Vec and classifying them with Hierarchical clustering [Step4]

Finally, by combining these three methods, it becomes possible to predict the change of factors, to handle specified factors properly according to the type of industry and annual sales, and furthermore to find the best keyword for appealing important factors. [Step5] This is exactly the purpose of this study mentioned in Section 1-2.

1-5. Composing Papers of This Study

This study consists of the following five papers that have already been published in international journals and the original part of this paper that combines their achievements.

[Paper 1]

Extraction of Fundamental KPIs in New Product Development Using Bayesian Network Analysis

Publisher:	International Association of P2M
Journal:	Journal of International Association of P2M
Authors:	Hironori Takuma, Yutaka Iwakami
DOI:	https://doi.org/10.20702/iappmjour.14.1_446

[Paper 2]

Consideration of Fundamental KPIs and Their Relationship with Environmental Protection in New Product Development Using Bayesian Network Analysis

Publisher:	IEEE
Conference:	2019 International Conference on Information Management and Technology (ICIMTech)
Authors:	Hironori Takuma, Yutaka Iwakami
DOI;	https://doi.org/10.1109/ICIMTech.2019.8843762

[Paper 3]

Analyzing enterprise attribute dependent KPIs/KGIs by Bayesian network leveraging LDA

Publisher:	IGI Global
Journal:	International Journal of Project Management and Productivity Assessment
Authors:	Yutaka Iwakami, Hironori Takuma, Motoi Iwashita
DOI:	Already accepted now in press

[Paper 4]

Properly initialized Bayesian Network for decision making leveraging random forest

Publisher: Sciedu Press

Journal: Artificial Intelligence Research

Authors: Yutaka Iwakami, Hironori Takuma, Motoi Iwashita

DOI: https://doi.org/10.5430/air.v9n1p36

[Paper 5]

Improving Matching Process with Expanding and Classifying Criterial Keywords leveraging Word Embedding and Hierarchical Clustering Methods

Publisher: Springer Nature

Journal: Review of Socionetwork Strategies

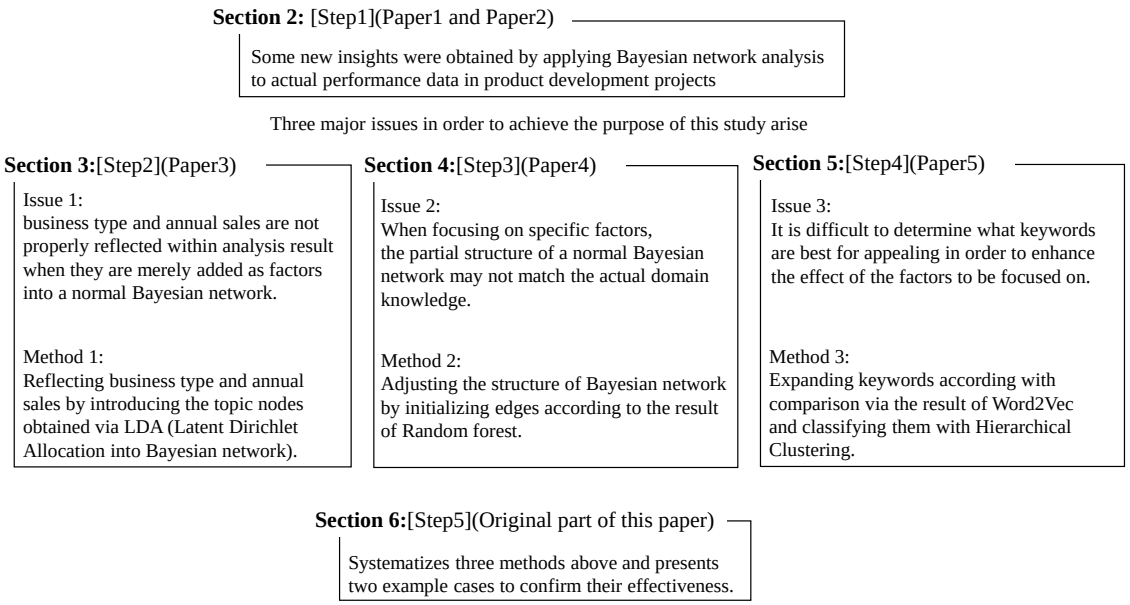
Authors: Yutaka Iwakami, Hironori Takuma, Motoi Iwashita

DOI: https://doi.org/10.1007/s12626-020-00063-4

1-6. Structure of This Paper

The structure of this paper, including the steps and methods described in Section 1-4 and the existing papers described in Section 1-5, is as follows.

Figure 1-3. Structure of this paper



As a tool for retrieving and analyzing of data, one of the representatives of statistical computing language and environment “R” is used in this study. Another well-known tool for Bayesian network analysis is BayoLinkS of NTT DATA Mathematical Systems, which is used in various studies such as estimation of students' learning states. (1-17:Kondo et al., 2019) or pattern classification of value creative consensus building process.(1-18:Hamada et al., 2019)

2. Applying Bayesian network to Product Planning and Development

2-1. Background

In order to understand relationships among factors (especially KGIs/KPIs) in product planning and development projects and to predict their changes, it is first necessary to define the factors that are candidates for KGIs/KPIs. These factors are observed as indicators such as annual sales of the product, impression of stakeholders and contribution degree of the external organization, etc. in actual product planning and development projects. These are also survey items when collection performance data of these projects. There are several approaches to how to define the factors. (2-1:Astebro& Michela, 2005; 2-2:Arai et al., 2014; 2-3:Olsen, 2003; 2-4:Nakamura et al., 2011; 2-5:Nakayama et al., 2015) This study adopts the results of an already proven study using the following methods. (2-6:Takuma et al., 2015)

1. List the candidates of factors from various existing research results. Then organize them based on the four viewpoints of the Balanced scorecard (Financial, Customer, Internal business process, Learning and growth), as it is a proven framework in project management using KGIs/KPIs.
2. Interview the frequency of use of the listed candidates with three experts, confirm the validity and excess / deficiency of the candidates, and organize a proper set of candidates as the factor list.

Table 2-1 lists the factors of product planning and development projects obtained in this way. Each series of A, C, D, E corresponds to the four viewpoints of the Balanced scorecard in order.

Table 2-1. List of factors in product planning and development projects

A series: Sales and revenue	
A1. Budget amount procured	Numerical input question. Values are log-transformed and discretized into five ranks.
A2. 3-year sales amount	Numerical input question. Values are log-transformed and discretized into five ranks.
A3. 3-year profit amount	Numerical input question. Values are log-transformed and discretized into five ranks.

A4. 3-year profit rate	Numerical input question. Values are log-transformed and discretized into five ranks.
C series: Customer and partner corporation	
C1. Customer response to prototype	Six ranks（1: no prototype, 2 to 3: disreputation, 4: no opinion, 5 to 6: good reputation）
C21. Average number of customers in product appealing scene	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No customers.
C22. Maximum number of customers in product appealing scene	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No customers.
C3. Impression for stakeholders	Six ranks (1: no response, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C4. Media's response to products	Six ranks (1: no response, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C5. Product's Contribution to natural environment and society	Six ranks (1: no contribution, 2 to 3: adverse effect, 4: no opinion, 5 to 6: favorable effect)
C61. Number of external companies cooperating the project	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No cooperation.
C62. Contribution degree of the external organizations	Six ranks (1: no cooperation, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C71. Number of venture companies cooperating the project	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No cooperation.
C72. Contribution of the venture companies	Six ranks (1: no cooperation, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C81. Number of outsourced companies cooperating the project	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No cooperation.

C82. Contribution degree of the outsourced companies	Six ranks (1: no cooperation, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
D series: Business process	
D1. Clearness of target and organization definition	Five ranks 1: (ambiguity) < 5 (clear)
D2. Procurement status of resources	Five ranks: 1 (deficiency) < 3 (no excess or deficiency) < 5 (abundance)
D3. With or without approval from relevant organizations	1: With approval 2: Without approval
D4. Quality inspection result of manufacturing process	Six ranks (1: no management, 2 to 3: permissiveness, 4: standard, 5 to 6: strict)
D5. Period of time required for product development	Numerical input question(months). Values are discretized into five ranks.
D61. 3-year sales volume	Numerical input question. Values are log-transformed and discretized into five ranks.
D62. 3-year contract completion rate	Numerical input question (%). Values are discretized into five ranks.
D7. 3 years sales number market share	Numerical input question (%). Values are discretized into five ranks (2 to 6) and 1: Unknown
D8. Number of troubles occurring in 3 years	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: Unknown
D91. Number of academic papers published	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No publication.
D92. Number of presentations in prominent academic journals	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No presentation.
D101. Number of in-house papers published	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No publication.

D102. Number of commendations for internal papers	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No commendation.
D111. Number of patent registrations	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No application.
D112. Success rate of patent registration	Numerical input question (%). Values are discretized into five ranks (2 to 6) and 1: No application.
D113. Number of registrations of design rights, trademark rights, and copyright	Numerical input question. Values are discretized into five ranks (2 to 6) and 1: No application.
E series: Learning and growth of human resources	
E1. Capability satisfaction of human resources	Six ranks: 1: unknown, 2 (row) < 6 (high)
E2. Status of human resource development	Six ranks: 1 no development, 2 (no development) < 6 (well-developed)
E3. Accumulation status of knowledge	Six ranks: 1: unknown, 2 (row) < 6 (high)
E4. Activation status of internal community	Six ranks: 1 no activation, 2 (completely inactive) < 6 (very active)
E5. Evaluation on the timing of market introduction	Six ranks: 1: unknown, 2 (row) < 6 (high)

Figure 2-1 summarizes the items in Table 2-1 by applying them to the scheme of PMBOK®. The horizontal axis represents the process groups, and the vertical axis represents the knowledge areas in PMBOK®. The rightmost column where items related to performance evaluation are arranged does not exist in the PMBOK® scheme. In this way, it can be confirmed from the PMBOK® scheme that this study analyzes the relationship between the factors related to product planning and development strategy and actual results.

Figure 2-1. Comparison of survey items and PMBOK® scheme

Knowledge areas / Process groups	Initiating Process Group	Planning Process Group	Executing Process Group	Monitoring and Controlling Process Group	Closing Process Group	Achievement evaluation
Integration Management		D3. With or without approval from relevant organizations				
Scope Management		D1. Clearness of target and organization definition				A2. 3-year sales amount D61. 3-year sales volume D62. 3-year contract completion rate D62. 3-year contract completion rate
Time Management		D5. Period of time required for product development				E5. Evaluation on the timing of market introduction
Cost Management		A1. Budget amount procured				A3. 3-year profit amount A4. 3-year profit rate
Quality Management						D4. Quality inspection result of manufacturing process D8. Number of troubles occurring in 3 years
Human Resource Management		C61. Number of external companies cooperating the project C62. Contribution degree of the external organizations C71. Number of venture companies cooperating the project C72. Contribution of the venture companies C81. Number of outsourced companies cooperating the project C82. Contribution degree of the outsourced companies				E1. Capability satisfaction of human resources E2. Status of human resource development E3. Accumulation status of knowledge E4. Activation status of internal community
Communication Management						D101. Number of in-house papers published D102. Number of commendations for internal papers
Risk Management						
Procurement Management		D2. Procurement status of resources				
Stakeholder Management		C1. Customer response to prototype C21. Average number of customers in product appealing scene C22. Maximum number of customers in product appealing scene				C3. Impression for stakeholders C4. Media's response to products C5. Product's Contribution to natural environment and society D91. Number of academic papers published D92. Number of presentations in prominent academic journals D111. Number of patent registrations D112. Success rate of patent registration D113. Number of registrations of design rights, trademark rights, and copyright

In the following part of Section 2, a questionnaire on actual product planning and development projects is conducted based on the factor list defined in this way, and the results are analyzed using Bayesian network

2-2. Example Case

In this study, as an example case, an online questionnaire on actual performance data of product planning and development projects for 250 companies. This survey was conducted by Nork Research, the Japanese domestic research firm which the author is affiliated with. The questionnaire consists of the following three steps.

Step1: Screening
Step2: Questioning
Step3: Cleaning

Details of each step are described below.

Step1: Screening

The research firm has registered candidates' population consisting of the personal data of more than 200,000 working people who agreed to be invited to this kind of survey in advance. At first, invitation e-mails are sent to about 10,000 candidates who were randomly selected from the population. After candidate's approval to take part in the questionnaire, they proceed to complete

the pre-questions provided by the web-based questionnaire system. The role of the pre-questions is to select proper respondents according to the purpose of the questionnaire. The condition for screening given as the pre-questions are described in Table 2-2.

Table 2-2. Condition of screening

Job responsibility	for screening survey respondents, only managers or higher responsibilities are selected, not included in analysis.
Business type of respondent's enterprise	Any of the following: agriculture, forestry and fisheries / mining / discrete manufacturing / process manufacturing / construction / wholesale / retailing / transportation / IT-related service / general service (except IT-related) /others
Annual sales of respondent's enterprise	Any of the following: less than 500 million yen / 500 million -3 billion yen /3 billion yen - 5 billion yen /5 billion - 10 billion yen /10 billion -30 billion yen / 30 billion - 50 billion yen / more than 50 billion yen
The amount of capital of respondent's enterprise	The amount of capital of survey respondent's enterprise exceeds 100 million Japanese yen
The experience of product development of respondent's enterprise	Survey respondent's enterprise has experience of product development more than two years

With these pre-questions, respondents whose job titles are executive or managerial were selected and their enterprises have recent experience (within two years) of its own product planning and development projects. If the amount of capital is relatively low, the enterprise tends to be a subcontractor of other larger ones and do not have their own projects. That is not suitable for this survey. Therefore, capital conditions are also set. Through this screening step 250 + α appropriate respondents are extracted.

Step2: Questioning

This step is called the main survey. The extracted 250+ α respondents answer the questionnaire

about the achievements of their own product planning and development projects. The question items are the factor list defined in Section 2-1.

Step3: Cleaning
Finally omit improper data, such as response time was too short, the sales revenue is too large, and etc. As a result, a CSV formatted row database of 250 records is obtained.

2-3. Application of Bayesian network

In this section, Bayesian network analysis is applied to the data collected in Sec2-2. Bayesian network analysis is a kind of statistical modelling method based on Bayesian statistics. A Bayesian network is a graphical representation of relationships among statistical events in the form of events drawn as nodes and relationships drawn as edges. It also satisfies the two conditions below.

Condition 1:

> Each edge is directed, and any chains of directed edges are not cycled (Directed Acyclic Graph: DAG).

Condition 2:

> The probability of a node is calculated with conditional probability of its parent nodes, of which directed edges are connecting towards the node.

Under the two conditions above, it is mathematically proven that the joint probability of events within the network can be calculated as a product of conditional probabilities between parents and children (2-7:Lee, 2011). That is to say, Bayesian network is an efficient way to represent relationships among events using DAG and conditional probabilities for calculating probabilities of multiple events. With such a network structure, Bayesian network has also these two advantages.

Advantage 1: Relationships among multiple events can be visually recognized.

Advantage 2: The influence on other events caused by the change of an event can be calculated, by the change being propagated via nodes and edges.

The latter process is called “probability propagation” or “belief propagation” (2-8:Yedidia et al., 2001). In probability propagation, setting a value for a node is called “providing evidence”. In this way, Bayesian network is useful in analyzing relationships among many factors in product

planning and development projects. In addition, Bayesian network is also used in various management fields such as quality management (2-9:Nguyen, 2015), ecological risk management. (2-10:Pollino et al., 2007) and others (2-11:Li et al., 2016; 2-12: Cui et al., 2006; 2-13:Ezawa, 1998)

Following three steps below, Bayesian network is constructed by learning the actual performance data of product planning and development projects obtained in Section 2-2. This procedure itself is common in Bayesian network analysis.

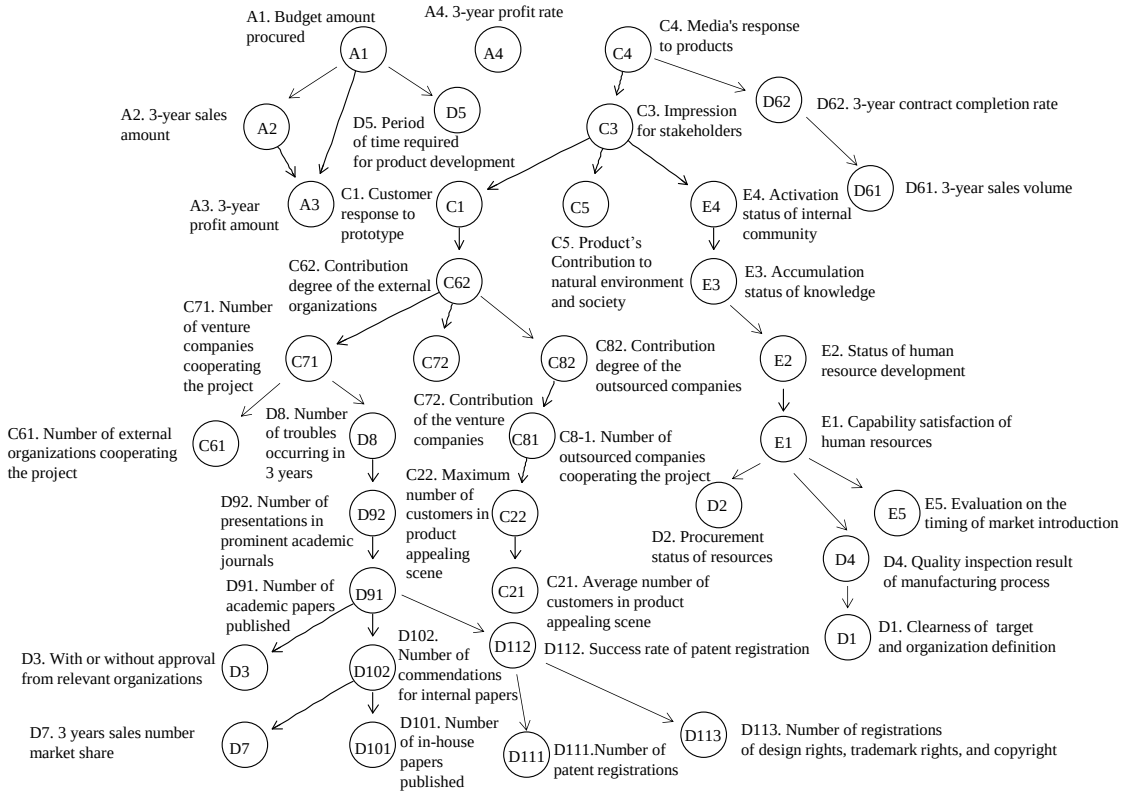
Step1: Data discretization
In Bayesian network, the value of node needs to be discrete, in order to set an evidence to it. Therefore, numeric entry questions are converted into multiple-choice questions in advance. In Table 2-1, it is already described, which question is to be discretized.

Step2: Selection of edge candidates
In order to avoid the influence of data specific bias as much as possible, iterative extraction (100 times, the same number of records) from the data is performed using the bootstrap method. Edges with an accuracy greater than $x = 0.7$ ($0 < x < 1$) in the result of the iteration are configured in initial state of the construction of Bayesian network as candidates of actual edges.

Step3: Network learning:
Construction of Bayesian network is performed. There are several algorithms for this process such as “Hill-climbing algorithm” or “K2 algorithm”. In this study, “Tabu search” is used (2-14:Pan et al., 2019; 2-15:Zhang et al., 2019). Tabu Search is an approach that marks recent operation as prohibited and tries to avoid repetition of the same operation, which might cause local minimum trap. Tabu Search itself is a kind of generalized algorithm used not only in Bayesian network but also in various applications.

The Bayesian network obtained in this way is shown in Figure 2-2.

Figure 2-2. Bayesian network obtained from data of 250 records



Looking at the obtained Bayesian network, there are some parts where nodes are not connected to others or where nodes with similar meanings are adjacent to each other. Therefore, before conducting a detailed analysis, less important nodes need to be omitted. Based on the network structure, domain knowledge and preliminary probabilistic propagation analysis, the following nodes will be deleted.

A1. Budget amount procured

A1 has the same meaning of "D2. Procurement status of resources" from a business point of view and creates redundancy, and shows similar trend in the preliminary analysis. It is also expected to A2 and A3 are connected with other nodes by elimination of A1.

A3. 3-year profit amount

A3 is not connected with other nodes. By summarization of the data, it is also confirmed that A3 has little correlation with other nodes. Due to external factors such as debt, it is assumed not closely related to other factors.

C22. Maximum number of customers in product appealing scene

"C21. Average number of customers in product appealing scene" can be used as a substitute of C22 from a business point of view and from the preliminary analysis.

Furthermore, C21 and C22 are adjacent.

C3. Impression for stakeholders

C3 has the similar trend with “C4. Media's response to products” from a business point of view and from the preliminary analysis. Furthermore, C3 and C4 are adjacent.

C81. Number of outsourced companies cooperating the project

C81 has the similar trend with “C61. Number of external companies cooperating the project” from a business point of view and from the preliminary analysis.

C82. Contribution degree of the outsourced companies

C82 has the similar trend with “C62. Contribution degree of the external organizations” from a business point of view and from the preliminary analysis. Furthermore, C62 and C82 are adjacent.

D3. With or without approval from relevant organizations

D3 is only connected with D9 at the end of the network. Form a business point of view, it does not strongly affect the achievement of product planning and development. Furthermore, by summarization of the data, it is also confirmed that D3 has little correlation with other nodes.

D5. Period of time required for product development

D5 is only connected with A1, which is to be deleted, at the end of the network. Form a business point of view, it does not strongly affect the achievement of product planning and development. Furthermore, by summarization of the data, it is also confirmed that D5 has little correlation with other nodes.

D62. 3-year contract completion rate

D62 has the similar trend with “D61. 3-year sales volume” from a business point of view and from the preliminary analysis. Furthermore, D61 and D62 are adjacent.

D92. Number of presentations in prominent academic journals

D92 has the similar trend with “D91. Number of academic papers published” from a business point of view and from the preliminary analysis. Furthermore, D91 and D92 are adjacent.

D101. Number of in-house papers published

D101 has the similar trend with “D102. Number of commendations for internal papers” from a business point of view and from the preliminary analysis. Furthermore, D101 and D102 are adjacent.

D111. Number of patent registrations

D111 has the similar trend with “D112. Success rate of patent registration” from a business point of view and from the preliminary analysis. Furthermore, D111 and D112 are adjacent.

D113. Number of registrations of design rights, trademark rights, and copyright

D113 is only connected with D112 at the end of the network. Form a business point of view, it does not much affect the achievement of product planning and development. Furthermore, by summarization of the data, it is also confirmed that D113 has little correlation with other nodes.

E1. Capability satisfaction of human resources

From a business point of view and from the result of preliminary analysis, it is assumed that it would be difficult for respondent to give quantitative evaluation to this factor at present.

E3. Accumulation status of knowledge

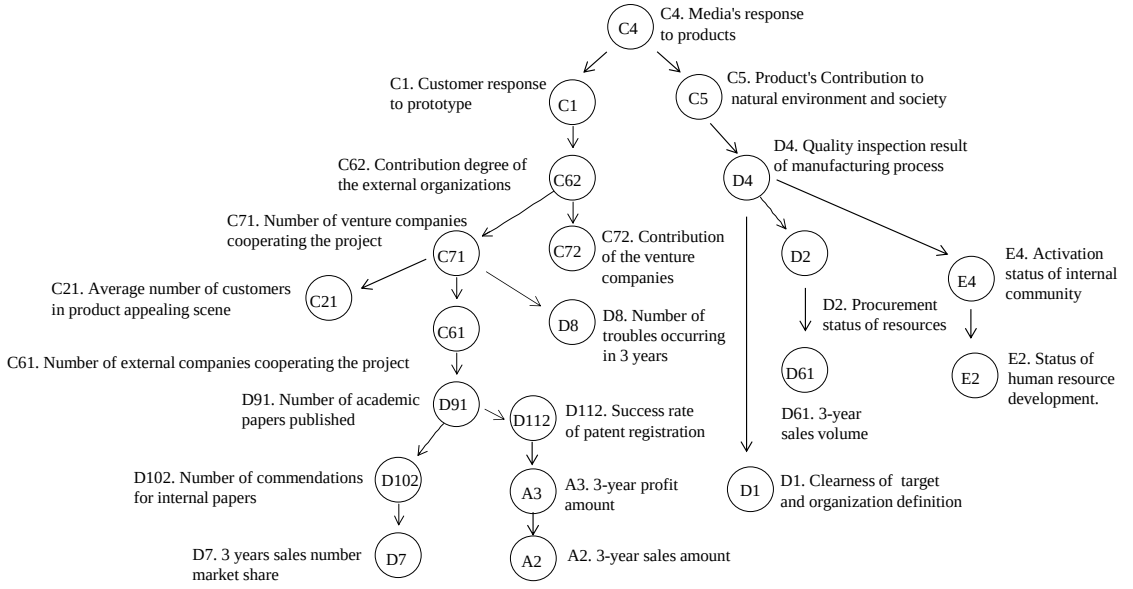
From a business point of view and from the result of preliminary analysis, it is assumed that it would be difficult for respondent to give quantitative evaluation to this factor at present.

E5. Evaluation on the timing of market introduction

E5 is only connected with E1 at the end of the network. Form a business point of view, it does not much affect the achievement of product planning and development. Furthermore, by summarization of the data, it is also confirmed that E5 has little correlation with other nodes.

Figure. 2-3 shows the result of Bayesian network construction again after selecting the factors in this way.

Figure 2-3. Bayesian network obtained from data of 250 records after eliminating redundant nodes



2-4. Acquisition of New Insights

In this section, some new insight will be found by probabilistic inference using the Bayesian network obtained in Section 2-3. Before getting into the detail, notation and terminology required for performing the analysis are organized again below.

Node and Edge:

A random variable represented by a circle on the Bayesian network is referred to as a node, whereas an arrow representing a conditional probability connecting variables is referred to as an edge.

Rank:

A discrete value that can be taken by a random variable subject to a Bayesian network analysis, which is generated from a questionnaire for collecting observed data, is referred to as a rank.

Evidence:

When inference is performed via a Bayesian network, specifying a value for a node is expressed as "providing evidence." Evidence includes hard evidence, which explicitly provides a specific value, and soft evidence, which specifies a possible value with a probability value.

Score:

Providing that the r-th rank value of the node X, which is the target of score calculation, is Rank (X(r)), and the probability that the node X has the value of Rank (X(r)) in the Bayesian network is BN (X(r), evd) when evidence = evd is given, the score of the node X under given evidence can be calculated as follows:

$$\mathrm{score}(\mathrm{X}, \mathrm{evd}) = \sum_{r} \mathrm{Rank}(\mathrm{X}(\mathrm{r}) \times \mathrm{BN}(\mathrm{X}(\mathrm{r}), \mathrm{evd})) \qquad (2-1)$$

As an initial stage of analysis, first take a bird's eye view of the obtained Bayesian network. From business point of view, the following four nodes correspond to KGIs.

"A2. 3-year sales amount"

"A3. 3-year profit amount"

"D61. 3-year sales volume"

"D7. 3 years sales number market share"

These four nodes are located at the bottom of the network as child nodes. The directions of edges do not necessarily represent strict causal relationships, but these locations of KGIs coincide well

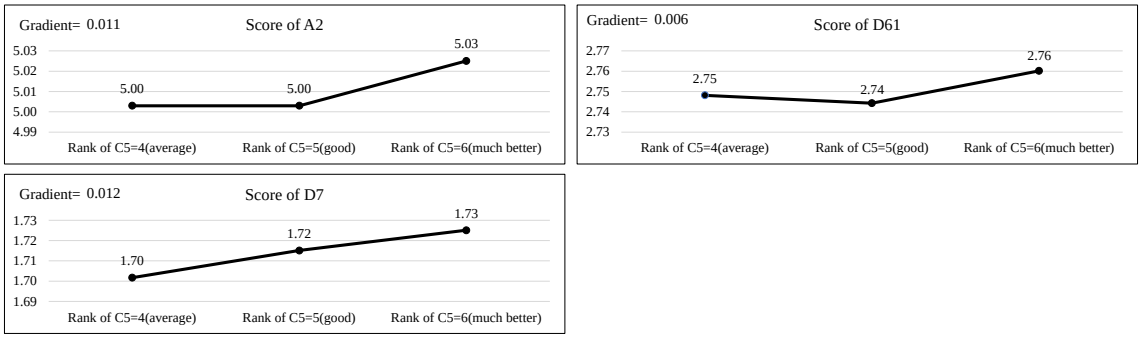
with actual situation, such as the value of KGIs are outcomes from KPIs.

The top parent node is “C4. Media's response to products”, which is the most important node affecting all other nodes. Actually, increasing the reputation of the media contributes to the improvement of product recognition and appeal, and ultimately has a positive effect on the number of sales and market share. Therefore, the location of C4 also matches the experience in actual business, too. here, probabilistic propagation will be performed for searching some new insights.

Next to C4, there is a node “C5. Product's Contribution to natural environment and society”. Recently, Promoting the contribution of the product to the environment is also effective in raising the reputation of the media. Therefore, predict how increasing the value of C5 will improve KGIs by probabilistic propagation. Since A2 and A3 are adjacent, and profit and number of sales are almost linked, KGIs to calculate the score are A2, D61 and D7.

Figure 2-4 shows the score of nodes A2, D61, and D7 respectively, when increasing the rank of C5 from 4 to 6.

Figure 2-4. Score change of A2, D61 and D7 according to C5

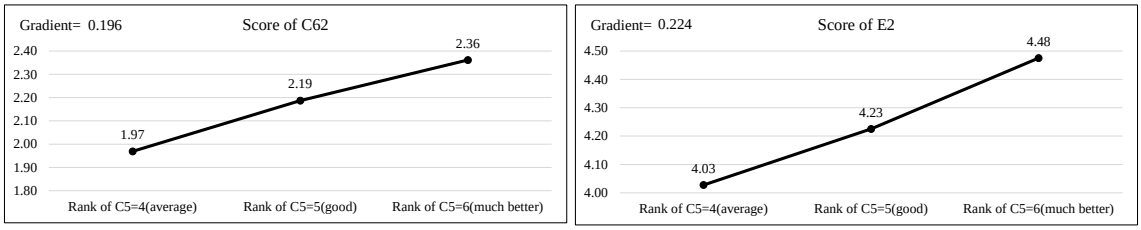


The gradient in the graph is the slope of the approximate straight line of the graph. As seen from the value of the gradient, these three graphs are almost flat, so C5 will not directly affect the KGIs, such as A2, D61 and D7. However, C5 may have a positive impact on other KPIs. When checking the score changes of various KPIs when the rank of C5 was changed, the following two nodes showed relatively high values.

“C62. Contribution degree of the external organizations”
“E2. Status of human resource development”

Figure 2-5 shows the score of nodes C62 and E2 respectively, when increasing the rank of C5 from 4 to 6.

Figure 2-5. Score change of C62 and E2 according to C5



Since the value of the gradient indicates, these two KPIs are influenced by the change of C5. Therefore, an effort to increase the contribution of products to the natural environment will work positively from the perspective of collaboration with external organizations or human resource development. In the future, by comparing benefit and burden of C5 by further detailed survey, it will be possible to measure the effect of C5 from multiple points of view. This approach can be also applied to the theme such as SDGs or security.

2-5. Three Major Issues Faced in Actual Businesses

In Section 2-4, after constructing Bayesian network from the actual performance data of product planning/development projects and selecting factors based on domain knowledge and preliminary analysis, some new insights on product planning/development and environmental protection could be obtained by probability inference.

But in actual business, it is required not only to obtain new unexpected insights, but also to predict the changes of KGIs/KPIs under the specified conditions, and to clarify which factor should be focused on, such as, “What should retailers do to increase product sales?” or “Does the relationship between the contribution of products to the environment and the number of sales differ depending on the scale of annual sales?”. However, when attempting to perform Bayesian network analysis in consideration of various conditions required in actual business as described above, the following problems arise.

Issue 1: business type and annual sales are not properly reflected within analysis result when they are merely added as factors into a normal Bayesian network.

Issue 2: When focusing on specific factors, the partial structure of a normal Bayesian network may not match the actual domain knowledge.

Furthermore, in efforts such that increasing the contribution of products to the natural environment, external appeal on the Internet is also important. In that case, the following issues will be added.

Issue 3: It is difficult to determine what keywords are best for appealing in order to enhance the effect of the factors to be focused on.

In the following Section 3, 4 and 5, the author proposes solutions for these three issues respectively.

3. Analysis Considering Business Type and Annual Sales

3-1. Background

In this section, of the three issues raised in Section 2, the proposed method to solve the following Issue 1 is described.

Issue 1: business type and annual sales are not properly reflected within analysis result when they are merely added as factors into a normal Bayesian network.

The author proposes the method for Reflecting business type and annual sales by introducing the topic nodes obtained via LDA (Latent Dirichlet Allocation) into Bayesian network. LDA (Latent Dirichlet Allocation) is a statistical model that is often used for document classification (corpus classification), which is based on the appearance frequencies of words. Firstly, words through all documents are classified into groups according to their frequency of appearances. This group is called “topic.” Then two types of probability distributions are generated. The first are word distributions, which represent the frequency of appearance of each word in each topic. Namely, there are as many word distributions as there are topics. The others are topic distributions, indicating which topic is likely to appear for each document, and each document has its own topic distribution. This model is learned according to the given data of documents and words. This model shows what topics each document is categorized into and what words each topic is characterized by.

LDA can be applied not only to document classification, but also to customer classification, based on the number of each item purchased, if “word” is replaced by “item,” and “document” by “purchaser” (3-1:Iwata & Sagawa, 2012). LDA is also applicable to the classification of images (3-2:Fei-Fei & Perona, 2005).

In this way, if some features from data can be replaced by frequency of appearance, classification by LDA can be done. In this study, topics reflecting business types and annual sales are generated and incorporated into the Bayesian network, which makes it possible to perform probability propagation among KPIs/KGIs, considering business type and annual sales.

3-2. Example Case

In order to describe the proposed method concretely, actual performance data of product planning and development projects are used as an example case in this section. Though the data screening conditions and questionnaire implementation method are the same as in Section 2, the number of

records has been increased to 992, and the question items are further improved based on other existing studies. (3-3:Porter & Kramer, 2011; 3-4: Tsochantaridis, Joachims, Hofmann, & Altun, 2005) The question items, which are also factors in Bayesian network analysis, are shown in Table 3-1.

Table 3-1. Question items

A series: Sales and revenue	
A1. 3-year sales	Numerical input question. Values are log-transformed and discretized into seven ranks.
A2. 3-year profits	Numerical input question. Values are log-transformed and discretized into seven ranks.
C series: Customer and partner corporation	
C1. Customer response to prototype	Six ranks（1: no prototype, 2 to 3: disreputation, 4: no opinion, 5 to 6: good reputation）
C2. Average number of customers in product appealing scene	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No customers.
C3. Media's response to products	Six ranks (1: no response, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C4. Product's contribution to natural environment and society	Six ranks (1: no contribution, 2 to 3: adverse effect, 4: no opinion, 5 to 6: favorable effect)
C51. Number of external organizations cooperating in the project	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No cooperation.
C52. Degree of contribution of the external organizations	Six ranks (1: no cooperation, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
C61. Number of venture companies cooperating in the project	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No cooperation.
C62. Contribution of the venture companies	Six ranks (1: no cooperation, 2 to 3: bad impression, 4: no opinion, 5 to 6: favorable impression)
D series: Business process	

D1. Clarity of requirements for products	Five ranks 1: (ambiguity) < 5 (clear)
D2. Procurement status of resources	Five ranks: 1 (deficiency) < 3 (no excess or deficiency) < 5 (abundance)
D3. Strength of management force	Six ranks (1: no management, 2 to 3: permissiveness, 4: standard, 5 to 6: strict)
D4. 3-year sales volume	Numerical input question. Values are log-transformed and discretized into five ranks.
D5. 3-year market share	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No information.
D6. Number of failures occurring in 3 years	Numerical input question. Values are log-transformed and discretized into five ranks (3 to 6) and 1: No information 2: zero case.
D7. Number of academic papers published	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No publication.
D8. Number of commendations for internal papers	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No commendation.
D9. Success rate of patent registration	Numerical input question. Values are log-transformed and discretized into five ranks (2 to 6) and 1: No application.
E series: Learning and growth of human resources	
E1. Status of human resource development	Six ranks: 1 no development, 2 (no development) < 6 (well-developed)
E2. Activation status of internal community	Six ranks: 1 no activation, 2 (completely inactive) < 6 (very active)

In addition, as with Section 2, the items related to enterprise attributes are shown in Table 3-2. (S1 is job responsibility, but not used in analysis, as it is for screening survey respondents.)

Table 3-2. Enterprise attributes

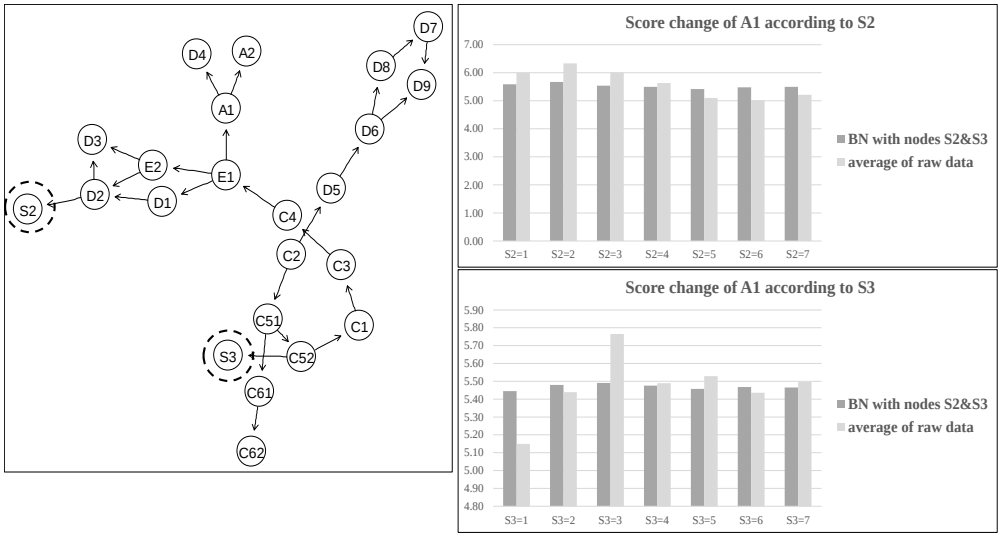
S series: Attribute of enteprises	
S2. Business type	agriculture, forestry and fisheries / mining / discrete manufacturing / process manufacturing / construction / wholesale / retailing / transportation / IT-related service / general service (except IT-related) /others
S3. Annual sales	less than 500 million yen / 500 million -3 billion yen /3 billion yen - 5 billion yen /5 billion - 10 billion yen /10 billion -30 billion yen / 30 billion - 50 billion yen / more than 50 billion yen

Since the theme of this section is analysis including enterprise attributes such as S2 (Business types) and S3 (Annual sales), the number of nodes subject to Bayesian network analysis is S series = 2, A series = 2, C series = 8, D series = 9, E series = 2, will be a total of 23.

3-3. Issue to be Solved

Before describing the proposed method in detail, reconfirm the issue to be solved in this section using the example case. For reflecting the effects of business type and annual sales on Bayesian network, it might be simplest to include “S2. Business type” and “S3. Annual sales” as ordinal nodes. The left diagram in Figure 3-1 shows the results of Bayesian network analysis, where S2 and S3 are added as ordinal nodes. The right graphs show the scores of “A1. 3-year sales” (one of the KGIs), when the value of S2 (upper graph) and S3 (lower graph) are changed. For comparison, the average A1 scores aggregated from the raw data are also plotted in the right graphs.

Figure 3-1. Bayesian network with S2 and S3, and inference results compared with raw data



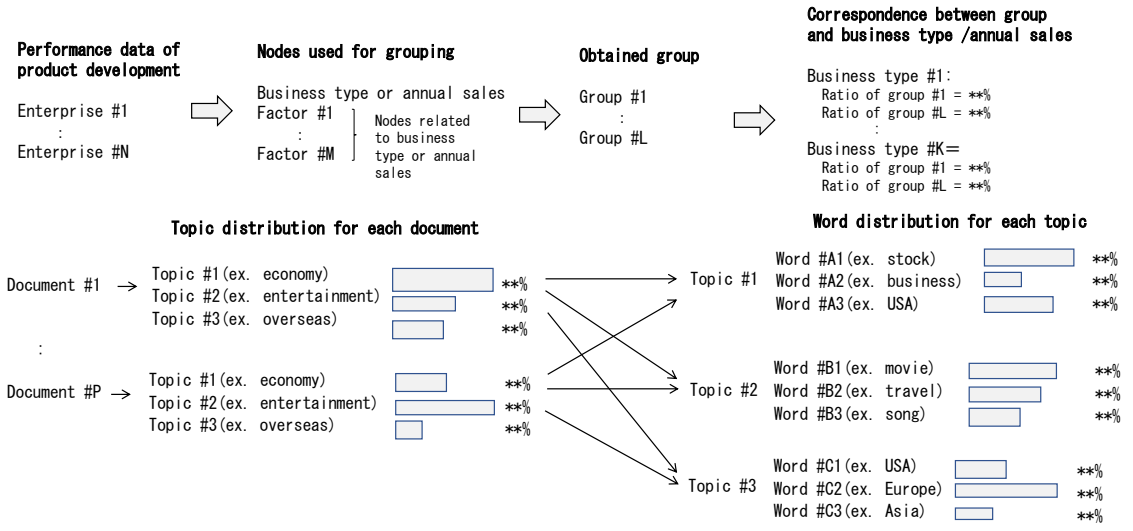
As the graphs in Figure 3-1 show, whereas the A1 scores fluctuate according to business type or annual sales for the raw data, the results are almost uniform in the result of Bayesian network. This is because S2 and S3 are located at the edge of the network, and their influences on other nodes are limited. However, S2 and S3 are basic attributes of enterprises and should have some effects on the A1 score as seen in the raw data. Therefore, the improvement by leveraging an LDA, as described in Concept and outline of proposed methods section, is required. This is an example of the issue mentioned in Section 3-1.

3-4. Proposed Method

One way to solve the issue in Section 3-3 is to construct Bayesian network for each business type or annual sales. However, there are enterprises such as SPAs (Specialty store retailer of Private label Apparel) that combine the characteristics of both manufacturing and retail. Furthermore, new product planning/development may often span multiple business types. As for annual sales, it might be difficult to analyze the relationship between product planning/development and annual sales growth, if Bayesian networks are divided into several categories of annual sales.

Another way is to add some group of nodes associated with business type or annual sales. At that time, as shown in the upper part of Figure 3-2, if the group of nodes can include information on factors related to business type or annual sales, it will be possible to properly reflect the difference on them. When enterprises correspond to documents, factors including business type/annual sales correspond to words, and groups correspond to topics, it has the same scheme as LDA (Latent Dirichlet Allocation) used in document classification, as shown in the lower part of Figure 3-2.

Figure 3-2. Comparison between the proposed method and LDA



However, since business type and annual sales correspond to words in LDA, it is necessary to calculate the association between business type/annual sales and the group from the results of LDA and the actual data.

To solve the issue in Section 3-3, the author proposes to apply LDA to several items closely related with S2 and S3, respectively. Then the topics in terms of S2 and S3 are extracted and incorporated into the Bayesian network as nodes. These nodes called “topic nodes.” Topic nodes are random variables reflecting business type and annual sales of an enterprise with topic distribution. The differences due to business type and annual sales can be suggested to Bayesian network by providing soft evidence on topic nodes.

There are two reasons for choosing LDA here, in addition to the commonality of the schemes shown in Figure 3-2.

The first is that the attributes of the company can be expressed flexibly. In particular, regarding business type, it is possible to express something like "Enterprise #1 has a manufacturing industry element of 70% and a retail industry element of 30%", and it will be possible to handle cross-industry forms.

The second is compatibility of probability propagation in Bayesian network. If S2 and S3 are directly added as ordinal nodes, the differences in business type and annual sales are given as hard evidence on Bayesian network. In this case, KGIs/KPIs might become conditionally independent by separation with S2 and S3. This makes it difficult to understand the relationships among KGIs/KPIs. On the other hand, the proposed method reflects the differences in business type and annual sales as soft evidence. Thus, probability propagation can be performed just like ordinal Bayesian network. Similar studies about the structure of Bayesian network are active in the fields of machine learning research (3-5:Linderman, Adams, & Pillow, 2016). And, contrary to this study, Bayesian approach is becoming applied to existing classification methods like clustering (3-6:Heller & Ghahramani, 2005).

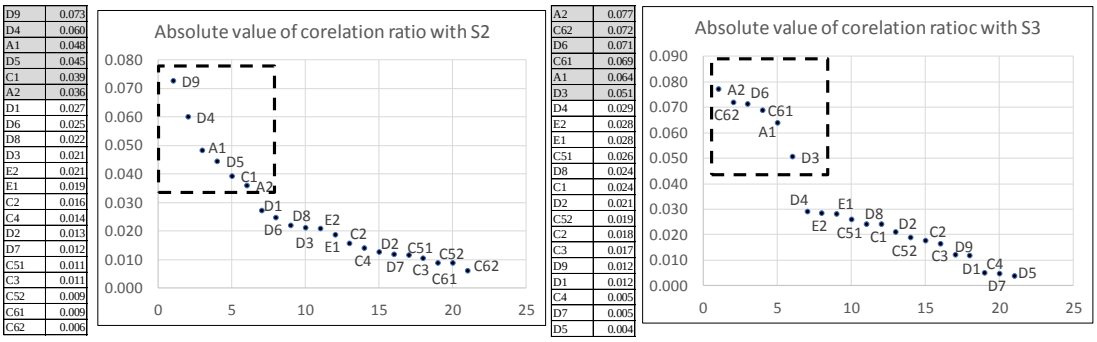
The proposed methods consist of two steps. The first is “generation of topic nodes with LDA”. The second is “construction and configuration of the Bayesian network with topic nodes.” The details of each step are described in the following, respectively.

3-4-1. Generation of Topic Nodes via LDA

For generating topic nodes instead of S2 and S3, LDA is required to be applied to 992 records,

with items listed in Table 3-1 and 3-2. As it is not efficient to include all the items into LDA, items that are closely related with S2 and S3 should be selected. Figure 3-3 shows items that have higher absolute values of correlation coefficient with S2 and S3, respectively.

Figure 3-3. Items highly correlated with S2 and S3



In each graph in Figure 3-3, items are sorted in descending order of absolute value of correlation coefficient with S2 and S3. As shown with dotted line, D9, D4, A1, D5, C1, and A2 are closely related with S2. On the other hand, A2, C62, D6, A1, and D3 are closely related with S3. Therefore, as shown in Table 3-3, LDA is applied to the two item groups, respectively.

Table 3-3. Items subject to the LDA in terms of S2 and S3

LDA in terms of S2	**LDA in terms of S3**
A1. 3-year sales	A1. 3-year sales amount
A2. 3-year profits	A2. 3-year profit amount
C1. Customer response to prototype	C61. Number of venture companies cooperating in the project
D4. 3-year sales volume	C62. Contribution of the venture companies
D5. 3-year market share	D3. Strength of management force
D9. Success rate of patent registration	D6. Number of troubles occurring in 3 years

In LDA, it is necessary to convert each item value into countable numerical data. In this study, multiple choice questions are designed to have equal intervals among the choices. That means rank2 can be regarded as a double rating of rank1. As for C1, D3, D61, and C62, there is a bidirectional choice of evaluation in a question. In those cases, a question is divided into two countable, unidirectional questions. For numerical questions such as A1, A2, D4, D5, D6, and D9,

values are log transformed in order to prevent a few large values from forming a single topic. These conversions are summarized in Table 3-4.

Table 3-4. Conversion prior to LDA analysis

Items subject to LDA in terms of S2	Quantification for LDA analysis
A1. 3-year sales	0⇒0, 1⇒1, 2 or more ⇒log transformation
A2. 3-year profits	0⇒0, 1⇒1, 2 or more ⇒log transformation
C1. Customer response to prototype	C1 is divided into two items: C1A representing the degree of "Unpopular" and C1B representing the degree of "Popular". "No trial" as 0 and "Equivocal" as 1. Others are divided into "Unpopular" and "Popular" from 2 to 6, respectively.
D4. 3-year sales volume	0⇒0, 1⇒1, 2 or more ⇒log transformation
D5. 3-year market share	0⇒0, 1⇒1, 2 or more ⇒log transformation
D9. Success rate of patent registration	0⇒0, 1⇒1, 2 or more ⇒log transformation

Items subject to LDA in terms of S3	Quantification for LDA analysis
A1. 3-year sales amount	0⇒0, 1⇒1, 2 or more ⇒log transformation
A2. 3-year profit amount	0⇒0, 1⇒1, 2 or more ⇒log transformation
C61. Number of venture companies cooperating in the project	0⇒0, 1⇒1, 2 or more ⇒log transformation
C62. Contribution of the venture companies	C62 is divided into two items: C62A representing the degree of "Bad impression" and C62B representing the degree of "Good impression". "No cooperation" as 0 and "Equivocal" as 1. Others are divided into "Bad impression" and "Good impression" from 2 to 6, respectively.
D3. Strength of management force	D3 is divided into two items: D3A representing the degree of "Tolerant" and D3B representing the degree of "Strict". "No management" as 0 and "Standard" as 1. Others are divided into "Tolerant" and "Strict" from 2 to 6, respectively.

D6. Number of troubles occurring in 3 years	0⇒0, 1⇒1, 2 or more ⇒log transformation

After the conversion in Table 3-4, LDA is applied to items related with S2 (A1, A2, C1A, C1B, D4, D5, and D9) and S3 (A1, A2, C61, C62A, C62B, D3A, D3B, and D6), respectively. At first, it is necessary to specify the number of topics. In this study, four major evaluation methods for the topics number are calculated (3-7:Arun, Suresh, Madhavan, & Murthy, 2010; 3-8:Cao, Xia, Li, Zhang, & Tang, 2009; 3-9:Deveaud, SanJuan, & Bellot, 2014; 3-10:Griffiths & Stevvers, 2004). Then the optimal topics number is decided based on the sum of the four methods score. In this way, the optimal number of topics is four for S2 and five for S3. Gibbs sampling is used in the inference of the probabilistic model of the LDA (3-11:Papanikolaou, Foulds, Rubin, & Tsoumakas, 2017). These nine topics (9 = 4+5) cannot be incorporated into the Bayesian network yet. Four topics (S21_pre, S22_pre, S23_pre, S24_pre) in terms of S2 and five topics (S31_pre, S32_pre, S33_pre, S34_pre, S35_pre) in terms of S3 are shown in Table 3-5. These tables show the appearance rate of items for each topic and its variance on the right edge.

Table 3-5. Appearance rate of items for each topic

Appearance rate of items for each topic in terms of S2

S2	A1	A2	C1A	C1B	D4	D5	D9	Variance
S21_pre	14.9%	41.2%	1.5%	3.8%	36.5%	1.7%	0.5%	0.03
S22_pre	55.1%	22.7%	1.3%	5.9%	12.7%	1.6%	0.7%	0.04
S23_pre	26.2%	1.0%	1.5%	10.6%	28.6%	6.1%	3.6%	0.01
S24_pre	42.1%	24.9%	0.8%	10.0%	13.1%	2.1%	7.1%	0.02

Appearance rate of items for each topic in terms of S3

S3	A1	A2	C61	C62A	C62B	D3A	D3B	D6	Variance
S31_pre	37.2%	36.6%	2.1%	0.2%	5.3%	7.1%	2.3%	9.2%	0.02
S32_pre	54.7%	37.2%	0.1%	0.1%	0.1%	1.5%	5.3%	0.9%	0.05
S33_pre	19.7%	52.6%	1.6%	0.2%	3.3%	1.0%	18.7%	2.9%	0.03
S34_pre	59.4%	28.1%	0.2%	0.1%	1.3%	0.7%	9.4%	0.7%	0.05
S35_pre	56.0%	29.3%	0.8%	0.3%	2.4%	4.9%	5.5%	0.8%	0.04

Each enterprise belongs to any topic in terms of S2 and S3. Therefore, when the probability that an enterprise belongs to topic A is written as P(A), those are satisfied for each enterprise.

$$P(S21\_pre) + P(S22\_pre) + P(S23\_pre) + P(S24\_pre) = 1 \quad (3-1)$$
$$P(S31\_pre) + P(S32\_pre) + P(S33\_pre) + P(S34\_pre) + P(S35\_pre) = 1 \quad (3-2)$$

That means, if the values of three topics in terms of S2 are determined, the values of the rest are also fixed. The same applies to topics in terms of S3. Therefore, one topic should be omitted, respectively. As a topic with little change would be the appropriate candidate for omission, S23_pre and S31_pre are omitted, according to the variance shown in Table 3-5. By renaming S21_pre, S22_pre, and S24_pre to S21, S22, and S23, three topics in terms of S2 are obtained. In the same way, S32_pre, S33_pre, S34_pre, and S35_pre are renamed to S31, S32, S33, and S34 and form four topics in terms of S3.

As the final step of generating topic nodes, the values of topics are discretized prior to being incorporated into Bayesian network analysis in the same way as described in Table 3-1. Considering the distribution of values, S21, S22, and S23 are discretized into six ranks, and S31, S32, S33 and S34 are discretized into four ranks. Finally, three topic nodes (S21, S22, S23) in terms of S2 and four topic nodes (S31, S32, S33, S34) in terms of S3 are obtained. These nodes are closely related to S2 and S3, respectively, and reflect the differences in business type and annual sales among enterprises.

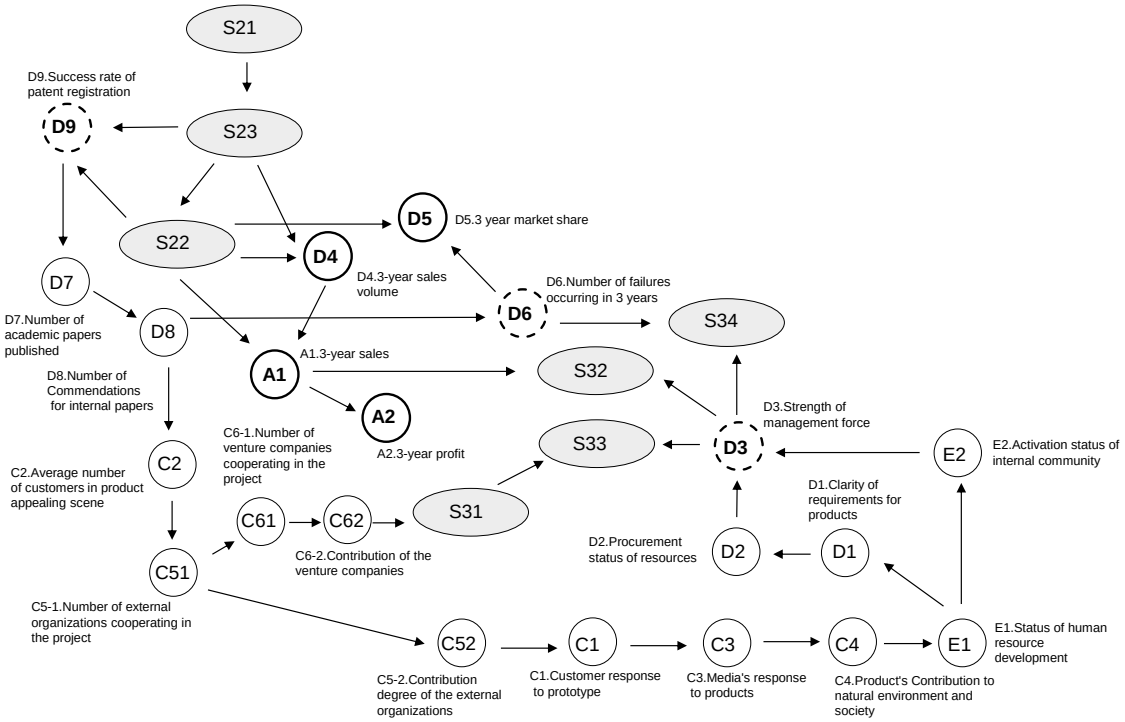
3-4-2. Construction and Configuration of Bayesian Network with Topic Nodes

The latter half of the proposed method is to incorporate topic nodes into the Bayesian network analysis. The original network in Figure 3-1 has 23 nodes, which are S2, S3, A series =2, C series =8, D series =9 and E series =2. With the proposed method, topic nodes are added instead of S2 and S3. As a result, this new network with topic nodes has 28 nodes, which are topic nodes in terms of S2 (S21, S22, S23), topic nodes in terms of S3 (S31, S32, S33, S34), A series =2, C series = 8, D series = 9 and E series = 2.

The construction process of the new network is performed in the same way as the original network. In this study, prior to the actual construction of Bayesian network, candidate directed edges are selected. At first, boot strap sampling is applied by randomly and iteratively extracting samples from 992 records (100 times, the same size of the original data for each). Then Bayesian network analysis is applied to those 100 data sets, separately. If a directed edge has ratio of appearance larger than x = 0.8 ($0 < x < 1$) out of 100 results, this edge is designated as a candidate of the actual construction of the Bayesian network. (Since higher accuracy is required than the analysis in Section 2, the condition of x is set to 0.8, which is stricter than 0.7)

After designating candidates of directed nodes, the actual construction of Bayesian network is performed in the same way as Section 2. In this way, the new network with topic nodes is obtained, as shown in Figure 3-4.

Figure 3-4. Bayesian network with topic nodes



The remarkable points of Figure 3-4 are as follows:

Topic nodes (represented as gray ovals in Figure 4):

S21, S22, S23 are topic nodes in terms of S2. They are close to each other. This result is natural, as these nodes have the same role of reflecting S2. In the same way, topic nodes in terms of S3 (S31, S32, S33, S34) are also close to each other.

KGIs (represented by the thick-line circle in Figure 4):

The following nodes are KGIs showing achievements in product planning and development projects:

"A1. 3-year sales"

"A2. 3-year profit"

"D4. 3-year sales volume"

"D5. 3-year market share"

KGIs are close to topic nodes in terms of S2 and S3. This result coincides well with common insight in real business, such that achievements of product planning and development projects differ in business type and annual sales. This is one of advantages of the new network with topic nodes compared with the original network.

Important KPIs (represented by the dotted-line circle in Figure 4):

The following nodes are directly connected with KGIs or topic nodes. As nodes that are close to each other have generally strong influences on each other in Bayesian network, these nodes would have strong influences on KGIs in terms of differences in business type and annual sales. In this meaning, the following are important KPIs in this study:

"D3. Strength of management force"

"D6. Number of failures occurring in 3 years"

"D9. Success rate of patent registration"

The next step is to configure the new network by giving soft evidence to topic nodes. By summarizing 992 records, the distribution of topic nodes for each attribute of S2 and S3 are obtained. For example, Table 3-6 shows the distribution of S21 according to attributes of S2.

Table 3-6. Distribution of S21 according to attributes of S2

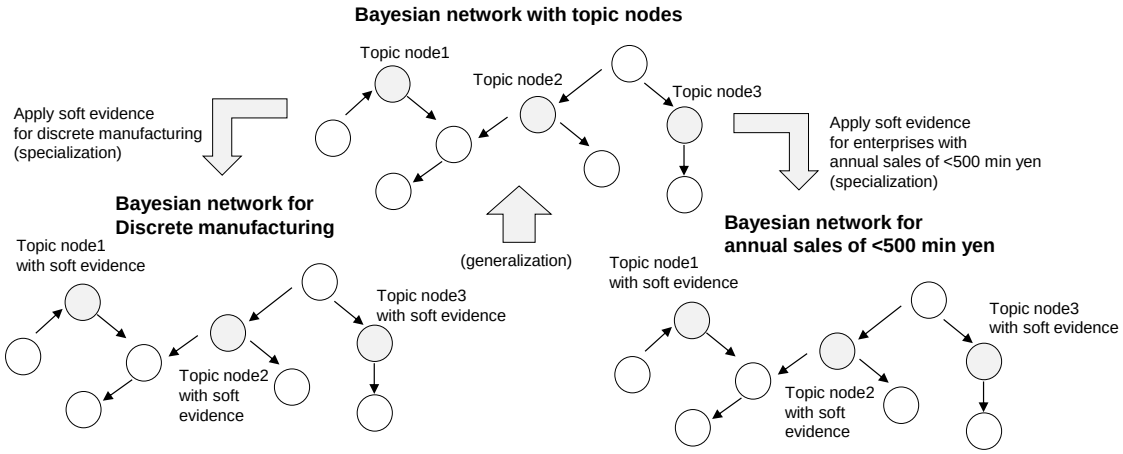
S21	rank1	rank2	rank3	rank4	rank5	rank6
agriculture, forestry, and fisheries	0.0%	0.0%	25.0%	12.5%	25.0%	37.5%
mining	0.0%	0.0%	0.0%	33.3%	33.3%	33.3%
discrete manufacturing	2.8%	3.4%	17.0%	17.0%	25.6%	34.1%
process manufacturing	0.7%	2.2%	14.7%	22.1%	24.3%	36.0%
construction	0.0%	1.8%	9.7%	19.5%	44.2%	24.8%
wholesale	0.0%	1.4%	6.9%	15.3%	33.3%	43.1%
retailing	0.0%	3.5%	1.8%	8.8%	36.8%	49.1%
transportation	0.0%	0.0%	8.5%	10.6%	34.0%	46.8%
IT-related service	0.7%	3.4%	10.2%	21.1%	38.1%	26.5%
General service (except IT-related)	0.0%	1.8%	10.0%	18.2%	30.0%	40.0%
others	0.8%	0.0%	2.4%	15.4%	43.1%	38.2%

The distributions of other topic nodes are obtained in the same way. These distributions play the role of combining attributes of business type and annual sales with the value of topic nodes. As

they represent the value of topic nodes with probability, they can be configured into the Bayesian network as soft evidence.

In this way, the construction and configuration of the Bayesian network with topic nodes is completed. In real business, it is not realistic to perform Bayesian network analysis individually for each attribute of S2 (11 attributes of business type) and S3 (7 attributes of annual sales), because 77 different networks are required to be generated. On the other hand, the proposed methods specialize one Bayesian network to fit 77 different patterns of S2 and S3 by giving soft evidence on topic nodes obtained from the LDA. Conversely, Bayesian networks with topic nodes are considered to be generalizations of each network for specific attributes in terms of S2 and S3. This structure is useful to explore relationships between KGIs/KPIs, while considering differences in business type and annual sales. The concept of this specialization and generalization is shown in Figure 3-5.

Figure 3-5. Specialization and generalization of the Bayesian network with topic nodes



There are already several studies that classify data into several latent classes by LDA and apply Bayesian network analysis for each class. The difference of this method is that topic nodes are generated with LDA and introduced into one Bayesian network that can represent multiple state of classes.

In the future, business type topic nodes may be able to analyze the relationship between business type conversion and product planning/development strategy. In addition, annual sales topic nodes may be able to analyze the relationship between the expansion of the enterprise scale and the results of product planning and development, and give the answer to "whether the integration of SMEs will lead to the improvement of product planning and development capabilities".

3-5. Result and Conclusion

In this section, the validity and effectiveness of the obtained Bayesian network with topic nodes is confirmed. At first, in "3-5-1. Values of Interest", how influences of important KPIs on KGIs are affected by annual sales and industry is shown. Next, "3-5-2. Average Value Range of KGIs According to Important KPIs" summarizes the differences shown in 3-5-1 according to various KPIs/KGIs. Furthermore, "3-5-3. Relevance Considerations" makes sure that the obtained Bayesian network shows reasonable results when comparing the originals and the data. Finally, in "3-5-4. Conclusion", the author states the conclusions of this section.

3-5-1. Values of Interest

In this section, by leveraging the new network with topic nodes, actual probability propagation is performed to see how influences of important KPIs on KGIs differ in business type and annual sales. By giving evidence on a KPI, a score of a KGI is calculated. For example, Figure 3-6 shows how A1 score changes according to the ranks of D3, and Table 3-7 is the actual value of the graph.

Figure 3-6. Score change of A1 according to D3

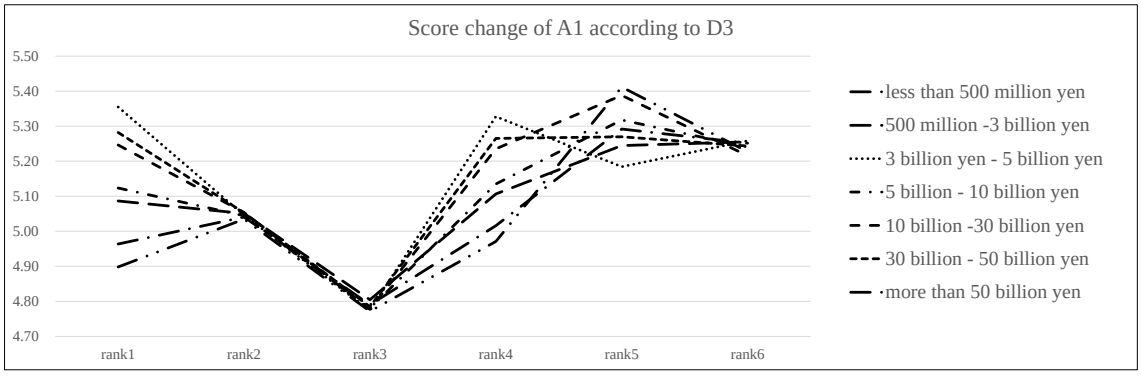


Table 3-7. Score change of A1 according to D3 and its average value range

	rank1	rank2	rank3	rank4	rank5	rank6
less than 500 million yen	4.90	5.03	4.77	4.97	5.41	5.23
500 million–3 billion yen	5.09	5.05	4.80	5.11	5.24	5.26
3 billion yen–5 billion yen	5.35	5.05	4.77	5.33	5.18	5.26
5 billion–10 billion yen	5.12	5.04	4.78	5.14	5.32	5.24
10 billion–30 billion yen	5.25	5.05	4.78	5.24	5.39	5.22
30 billion–50 billion yen	5.28	5.05	4.78	5.26	5.27	5.24

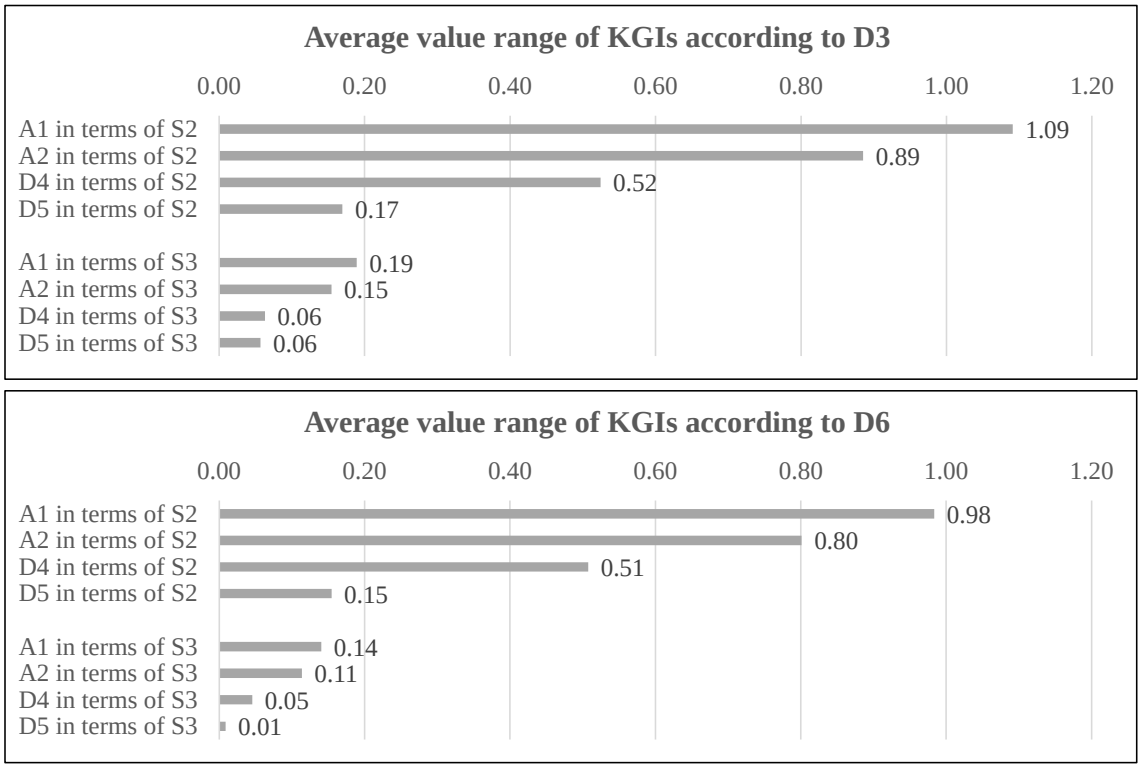
more than 50 billion yen	4.96	5.04	4.79	5.02	5.29	5.25
Average of value range of A1 through annual sales (*)	0.46	0.02	0.03	0.36	0.23	0.04
Average of (*) through ranks of D3	0.19					

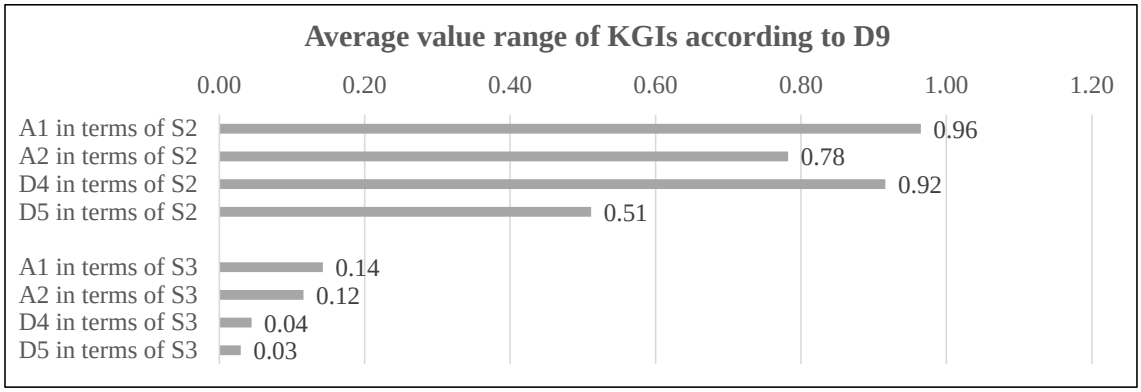
The focus here is to find how influences of important KPIs (D3, D6, D9) on KGIs (A1, A2, D4, D5) differ in S2 (business type) and S3 (annual sales). It would become complicated to enumerate all the scores of KGIs according to KPIs, as shown in Table 3-7. Therefore, for each pair of concerning KPI and KGI, the average value range of KGI, as for S2 or S3, through the ranks of KPI is calculated. (In the example in Table 3-7, 0.19 in the bottom of the table is the value of interest.)

3-5-2. Average Value Range of KGIs According to Important KPIs

Figure 3-7 shows the average value range of KGIs according to important KPIs (D3, D6, and D9).

Figure 3-7. Average value range of KGIs according to important KPIs (D3, D6, and D9)





The average value range of KGIs (A1, A2, D4, and D5), according to important KPIs (D3, D6, and D9), are more influenced by differences in business type than in annual sales. On the other hand, a closer look at Figure 3-4 shows that there are the following three patterns of locations of nodes:

Pattern1: There are topic nodes in terms of S3 in front of KGIs. (D3 and its neighbors)
Pattern2: Connected with one of the KGIs, without topic nodes between them. (D6)
Pattern3: There are topic nodes in terms of S2 in front of KGIs. (D9 and its neighbors)

Considering positional relationships with KGIs and topic nodes, it is assumed that the following results are obtained for each pattern:

Pattern1: The scores of KGIs differ in S3 but not in S2, according to D3 and its neighbors.
Pattern2: The scores of KGIs are not affected by S2 and S3, according to D6.
Pattern3: The scores of KGIs differ in S2 but not in S3, according to D3 and its neighbors.

But a more precise look at Figure 3-4 shows that three out of four KGIs (A1, D4, D5) are directly connected with topic nodes in terms of S2, and A2 is only connected with A1. Therefore, Figure 3-4 shows that all four KGIs are more strongly affected by the difference in business type. This visual observation coincides with the result shown in Figure 3-7. In this way, by leveraging the Bayesian network with topic nodes, the influences of KPIs on KGIs, with consideration for business type and annual sales, can be inferred not only with probability propagation but also with visual observation of the network.

It is also important to perform the detailed probability propagation in order to see how one KPI affects a particular KGI. For example, “C4. Product's Contribution to natural environment and society” is a KPI, which is related to sustainability. It is one of the recent points of interest in

product planning and development projects to understand in which business type sustainability is also effective in increasing sales of the product. The answer is obtained from the result of average value range of A1 according to C4 in terms of business type, as shown in Figure 3-8.

Figure 3-8. Average value range of A1 according to C4 in terms of business type

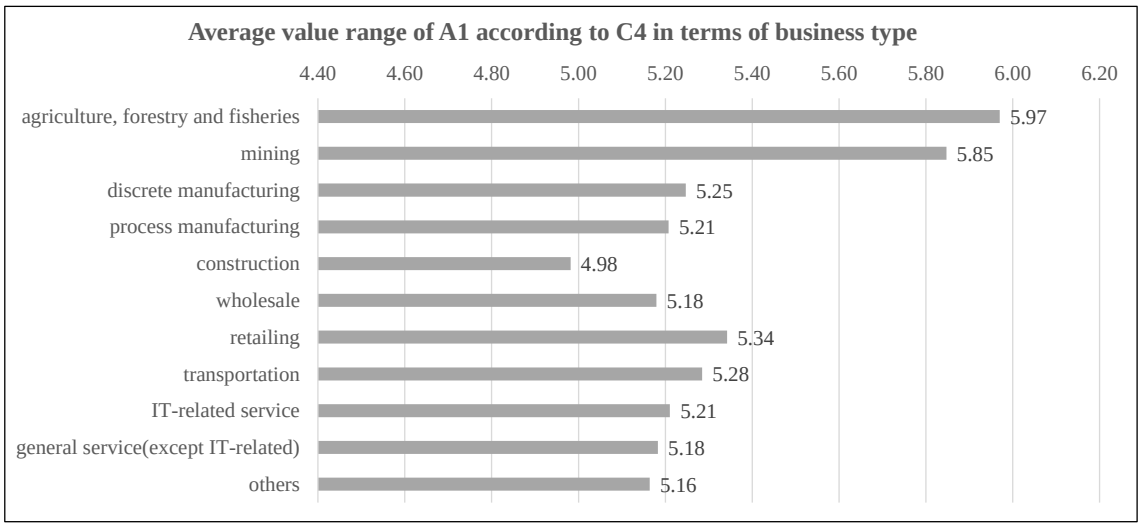


Figure 3-8 shows that sustainability initiatives also contribute to sales growth of the product, especially in business types such as “agriculture, forestry, and fisheries” and “mining,” but have relatively little contribution to sales growth in other types such as “construction.” This is a new insight that could not be obtained by Bayesian network analysis of Section 2. In this way, the new network with topic nodes is expected to support sustainability efforts based on business type.

3-5-3. Relevance Considerations

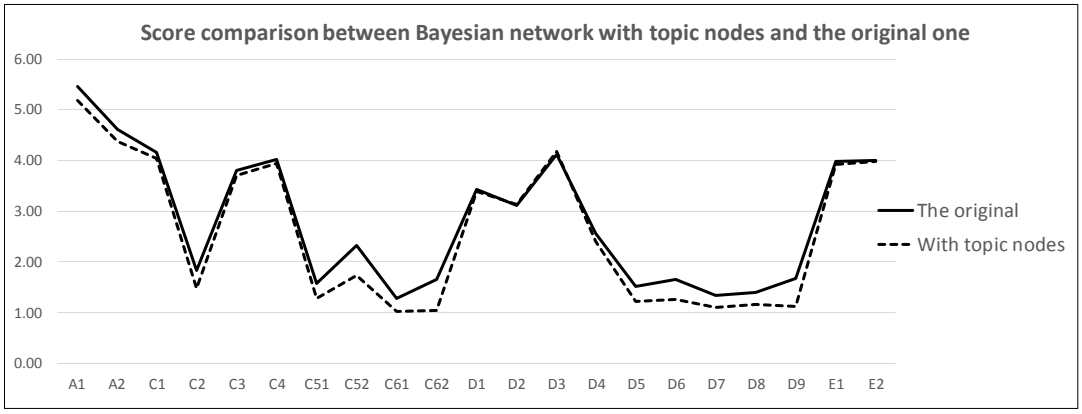
In the previous section, the effectiveness of the Bayesian network with topic nodes is demonstrated through actual probability propagation. In this section, the validity of the Bayesian network with topic nodes is confirmed from the following three viewpoints:

3-5-3-1. Consistency with the original network

At first, it is necessary to confirm that the Bayesian network with topic nodes is not totally different from the original one in terms of existing nodes. Figure 3-9 shows the score comparison between the Bayesian network with topic nodes and the original one.

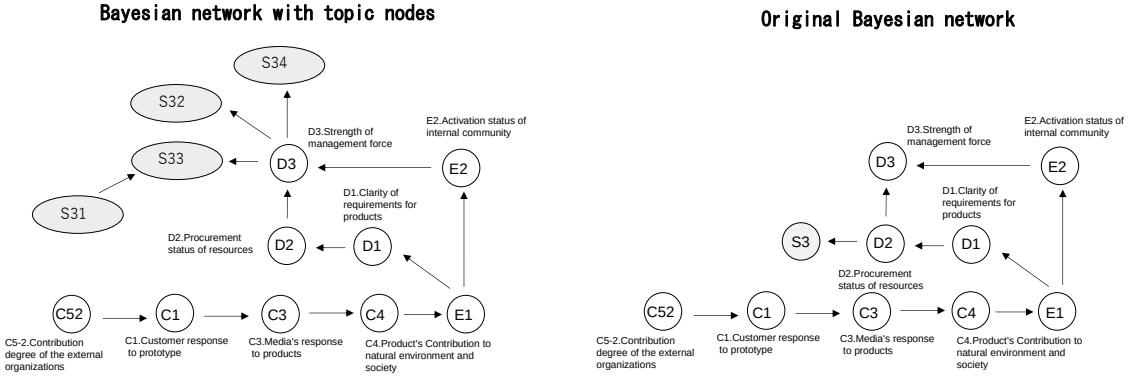
Figure 3-9. Score comparison between the Bayesian network with topic nodes and the original one

	A1	A2	C1	C2	C3	C4	C51	C52	C61	C62	D1	D2	D3	D4	D5	D6	D7	D8	D9	E1	E2
The original	5.47	4.60	4.15	1.82	3.80	4.01	1.57	2.31	1.27	1.64	3.42	3.12	4.13	2.56	1.52	1.65	1.34	1.39	1.67	3.97	4.00
With topic nodes	5.19	4.38	4.03	1.48	3.70	3.94	1.28	1.74	1.02	1.04	3.39	3.14	4.18	2.40	1.22	1.25	1.10	1.16	1.12	3.92	3.97



Also, Figure 3-10 shows a comparison of the existing nodes between Bayesian network with topic nodes and the original Bayesian network. It can be seen that the structure has not changed due to the insertion of the topic node. The same applies to other parts of the network.

Figure 3-10. Comparison of the existing nodes between Bayesian network with topic nodes and the original Bayesian network



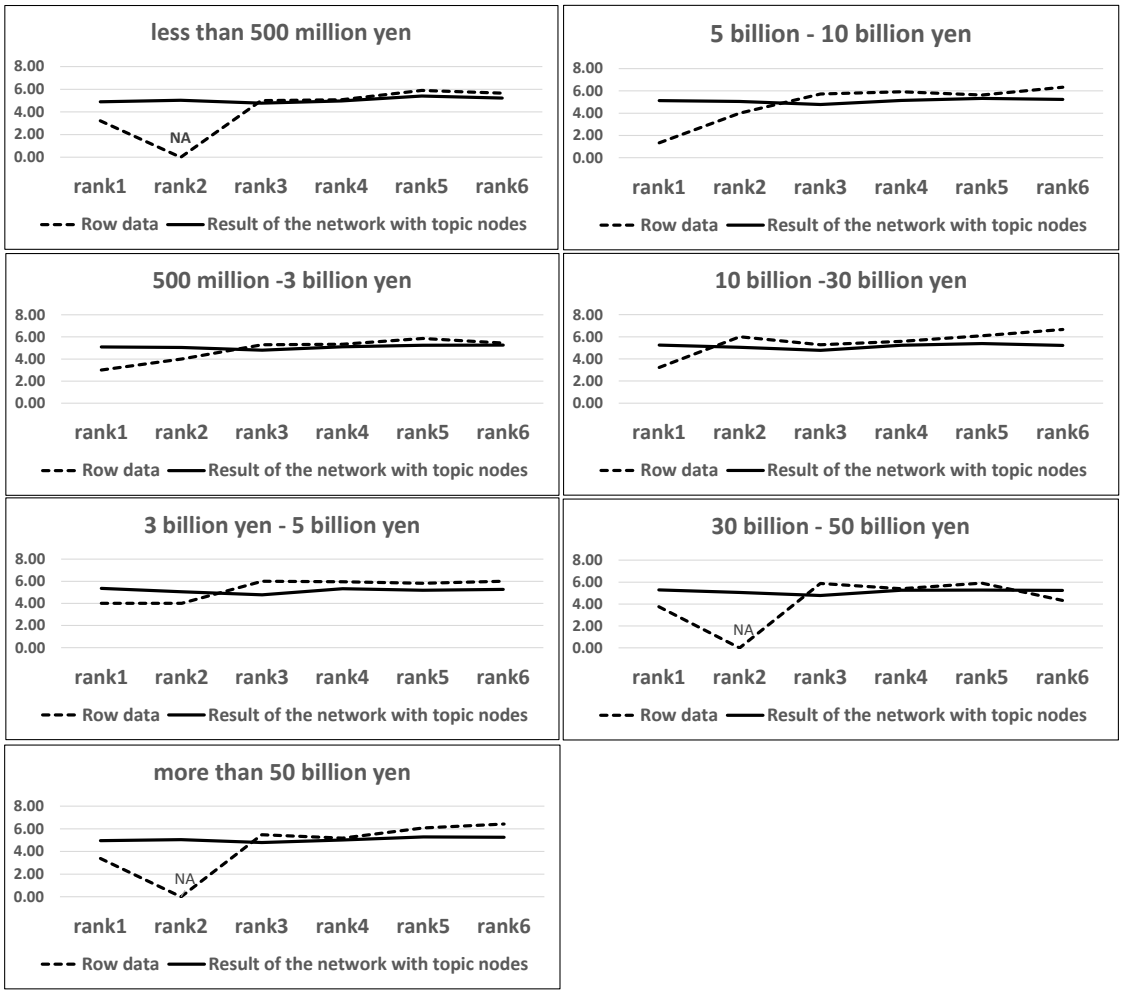
In this way, it is confirmed that the Bayesian network with topic nodes is essentially the same as the original one, as the score of all existing nodes are almost unchanged, as shown in Figure 3-9 and Figure 3-10.

3-5-3-2. Comparative check with the raw data

It is also important to do a comparative check with the raw data. Figure 3-11 shows the

comparison of the A1 score according to D3 in terms of S3. One is obtained by summarizing the raw data, and the other is the result of probability propagation with the network with topic nodes. "NA" indicates there is no corresponding data in the raw data.

Figure 3-11. Comparison of A1 score according to D3 in terms of S3



The result of the network with topic nodes is almost the same as the raw data. There are some points where the difference between them is relatively large. But these differences show that the Bayesian network complemented the missing or singular values in the raw data. As shown in each graph of Figure 3-11, all of the NAs are found to be properly complemented, considering the neighboring values. Furthermore, in the graph of 5 billion–10 billion yen, the value for rank 1 in the raw data is singular. This value is specifically small, because the score of A1 is especially biased towards small values, compared with other cases. Therefore, it is also confirmed that the Bayesian network with topic nodes coincides well with the raw data.

3-5-3-3. Verification by other research results

Finally, consistency with the results obtained in existing, related studies is confirmed. The second graph of Figure 3-7 shows that the average value range of D5 according to D6 is not influenced by differences in business type and in annual sales. That means the relationship between "number of failures occurring in 3 years" (D6) and "3-year market share" has almost the same tendency, regardless of business type or annual sales. On the other hand, there are several academic papers mentioning issues like the relationship between D5 and D6. For example, excessive quality control is known to be an interference in product planning and development (3-12:Cooper, 2019). In another example, it is pointed out that rationality of thinking is sometimes required, which means failure should not be overly avoided (3-13:Eliëns, Eling, Gelper, & Langerak, 2018). Furthermore, sustainable supply operation, distribution, product recovery, and return practices are claimed to be important (3-14:Gupta, Dangayach, Singh, Meena, & Rao, 2018). These examples are all mentioned as common insight of product planning and development, regardless of business type or annual sales.

Furthermore, the third graph of Figure 3-7 shows that the average value range of KGIs (A1, A2, D4 and D5) according to D9 is more influenced by business type than annual sales. The similar tendency was found in Japanese large-scale research and development promotion services (3-15:Arai, Takuma, & Kameyama, 2015). Therefore, the result of the proposed methods also coincides with other related studies.

3-5-4. Conclusion

Inference using Bayesian network is an effective method for clarifying relationships between KPIs and KGIs in product planning and development projects. At the same time, it is necessary to consider the differences in business type and annual sales. However, incorporating business type and annual sales into the Bayesian network as nodes does not provide expected results. In this study, LDA, which is widely used for document classification, is applied to generate topic nodes into the Bayesian network, which reflects the business type and annual sales of enterprises. The differences in business type and annual sales can be set as soft evidence on topic nodes. When soft evidence is given to this Bayesian network with topic nodes, the network is regarded to be specialized for specific attributes of enterprises. In this way, with the Bayesian network with topic nodes, the relationship among KGIs/KPIs can be effectively inferred, according to attributes of enterprises.

4. Analysis Focusing on Specific Factors

4-1. Background

In this section, of the three issues raised in Section 2, the proposed method to solve the following Issue 2 is described.

Issue 2: When focusing on specific factors, the partial structure of a normal Bayesian network may not match the actual domain knowledge.

The author proposes the method for Adjusting the structure of Bayesian network by initializing edges according to the result of Random forest. This proposed method can be applied to other than product planning and development projects. Therefore, taking into account the subsequent relationship with Section 5, the author explains the proposed method by taking a scene of selecting keywords to be appealed on SNS.

As described in Section 4-2, the author picks up an existent example of a beer maker and its representative product, which are noted “the maker” and “the product” respectively. If a consumer posts a keyword related with beer such as “sharpness” in SNS, the maker needs to know how often this keyword occurs with the product, in order to determine whether they should include “sharpness” into their marketing messages in SNS. If the keyword more often occurs in SNS with the product, it is considered to be engaged with and closely related with the product. It is called “engaged keyword” in this study. On the other hand, the keyword with less often occurrence with the product, it is called “non-engaged keyword”. The business success for the maker depends on how to select engaged keywords effectively.

As described in Section 4-2, in Bayesian network analysis of this study, two sets of tweet records are retrieved from Twitter. One is “engaged tweets”, which contain the product. The other is “non-engaged tweets”, which only contain an ordinal word “beer”. These two sets are combined into a dataset. Each record of the dataset has columns indicating whether contains each engaged keyword or not, and also has engaged/non-engaged flag as the last column. By applying Bayesian network analysis, the inference can be performed such as, tweets with keyword A are likely to be engaged tweets, but with the combination of keyword B, it does not. However, as mentioned in Section 4-3, the situation occurs such as, even when tweets with the combination of keyword A and keyword B are likely to be engaged tweets based upon the inference, the actual search result on Twitter shows the opposite result.

The cause of such disagreement would reside in the characteristic of Bayesian network. In the example above, the engaged/non-engaged flag is a kind of target node (explained variable), which has different role from nodes of engaged-keywords (explanatory variables). But Bayesian network usually handles all these nodes equally. Therefore, some adjustment will be required to apply Bayesian network for decision making task like this case.

On the other hand, Random forest is a proven method for analyzing the influence of explanatory variables upon explained variable (4-1:Harris, 2013; 4-2:Khalilia et al., 2011). As described in Section 4-4, the author proposes to configure initial state of Bayesian network leveraging the result of Random forest analysis. The initial state consists of a few nodes around the target node and several edges between these nodes and the target. The former is called “initial nodes” and the latter is called “initial edges” in this study. Initial nodes are extracted by measuring mean decrease of Gini coefficient calculated with decision trees of Random forest, because explanatory variables with much influence on explained variable show significant decrease of the coefficient. Directions of edges correspond to conditional probability among nodes connected with those edges. Therefore, directions of initial edges are designated based on likelihood measured by similarity of conditional probability distribution between actual data and predicted result of Random forest. The similarity is calculated with Wasserstein metric. Initial nodes and initial edges are given as an initial state for the construction of Bayesian network.

As confirmed in Section 4-5, the inference result of the Bayesian network with initial state coincides well with the actual search result on Twitter. Configuring initial state leveraging the result of Random forest analysis is considered to be a kind of adjustment of Bayesian network to perform decision making with explained/explanatory variable as nodes.

4-2. Example Case

As already mentioned in Section 4-1, the author picks up an existent example of a beer maker and its representative product. Prior to this study, the author extracted 18 engaged keywords for the product leveraging Word2Vec in the same way described in the following Section 5. At first two sets of tweets are searched and retrieved on Twitter. One is the set of tweets which include the product. The other consists of tweets including ordinal keyword in terms of the business domain such as “beer” in this case. Then Word2Vec analysis is applied for the mixture of the two sets. If a keyword shows closer direction to the product than to the ordinal keyword in vector space obtained by Word2Vec, the keyword is considered to be more closely related with the product than other keywords. With this procedure, engaged keywords for the product are obtained as shown in Table 4-1.

Table 4-1. Engaged keywords of the product

Word1	Rich	Word6	alcohol percent	Word11	cheers	Word16	solid
Word2	Guzzle	Word7	craft beer	Word12	bitterness	Word17	thick
Word3	Chilled	Word8	dry	Word13	refreshing	Word18	taste
Word4	drinkable	Word9	Belgium	Word14	brisk		
Word5	Lager	Word10	fruity	Word15	strongest		

The purpose is to pick up engaged keywords more related with the product from Table 4-1 with Bayesian network analysis. For performing Bayesian network analysis, a dataset is retrieved on Twitter. The dataset consists of two types of tweets. One is “engaged tweets”, which contain the product. The other is “non-engaged tweets”, which only contain an ordinal word “beer”. Each record of the dataset has columns indicating whether contains each engaged keyword (contain=1/ not contain=0) and also has engaged/non-engaged flag as the last column by the name of “engaged” (engaged=1/not engaged=0). The total number of tweets is 1046 (engaged: 357, non-engaged: 689). The structure of the dataset is shown in Figure 4-1.

Figure 4-1. Structure of the dataset

Appearance of engaged keywords Engaged/Non-engaged flag

	Word1	Word2	・・・・	Word18	engaged
Tweet1	0	0		1	1
Tweet2	1	1		1	1
:					
:					
:					
Tweet357	1	0		1	1
Tweet358	0	1			0
:					
:					
Tweet1046	0	1		0	0

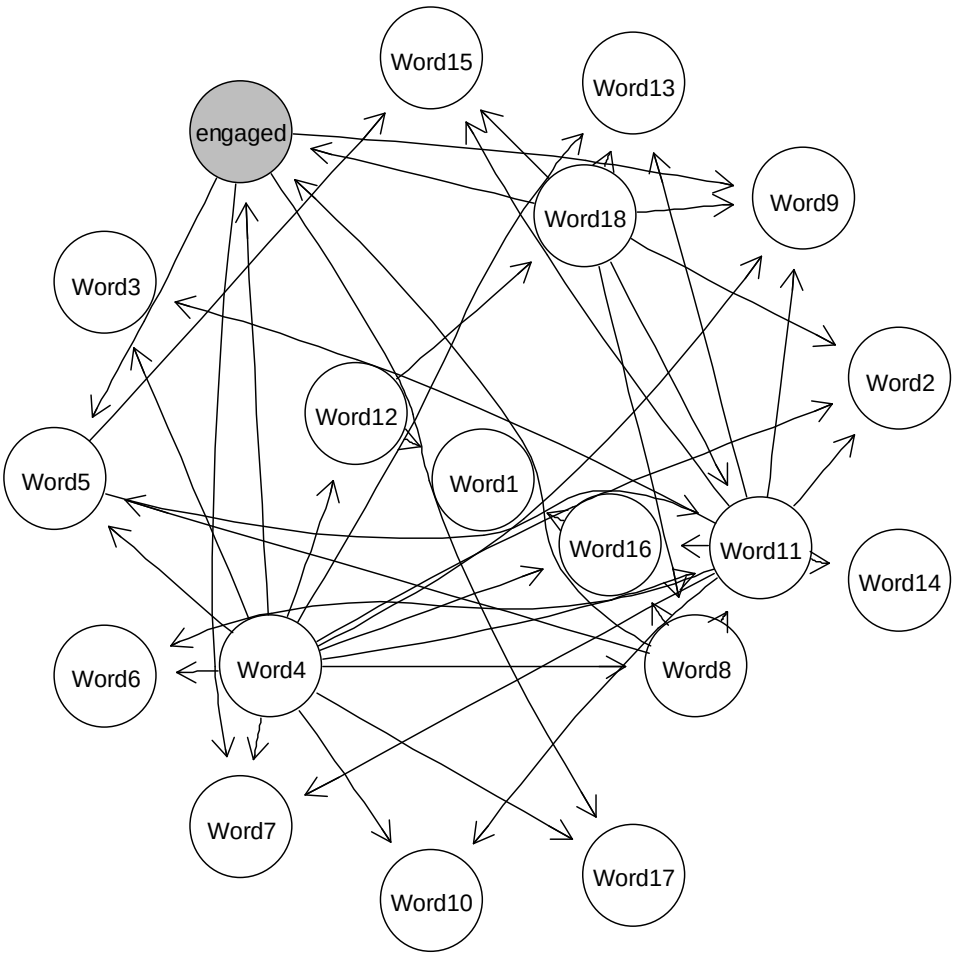
Engaged tweets (357 records)

Non-engaged tweets (689 records)

4-3. Issue to be Solved

The obtained Bayesian network is shown in Figure 4-2. The algorithm used for learning the structure of Bayesian Network is the same as described in Section 2. The target node (“engaged”) is marked as gray. It is directly connected with engaged keywords such as Word4, Word5, Word7, Word8, Word9, Word17 and Word18.

Figure 4-2. Result of ordinal Bayesian network



The network shows several visual insights. For example, the connection of three nodes, "engaged", "word4" and "word8"includes a tail-to-tail relationship (there are two edges from"Word4" to both "engaged" and "Word8"), which is one of three basic connections composing Bayesian Network. In a tail-to-tail, after "Word4" is determined, "engaged" is conditionally independent of "Word8". But the actual network in Figure 2 is much more complicated, such as "Word8" has an edge directly connected to "engaged". A few examples of probabilistic inference for those three nodes as shown in Table 4-2.

Table 4-2. Example of probabilistic inference

$P(engaged = 1 \mid Word8 = 0)$	31.2%
$P(engaged = 1 \mid Word8 = 1)$	78.6%
$P(engaged = 1 \mid Word8 = 0, Word4 = 0)$	33.4%
$P(engaged = 1 \mid Word8 = 1, Word4 = 0)$	82.3%
$P(engaged = 1 \mid Word8 = 0, Word4 = 1)$	22.3%

$P(engaged = 1 \mid Word8 = 1, Word4 = 1)$	24.5%

Table 2 shows that “engaged” is strongly affected by “Word8”, but not under condition “Word4” =1 is given. The result of the inference tells that “Word8” is not preferable for being used with “Word4” for making marketing messages effectively engaged with the product. As “Word4” has similar relationships among other nodes, the network suggests, the maker should pay attention in using “Word4”.

However, some of the probabilistic inference with the network in Figure 4-2 do not coincide with the actual search result on Twitter. For example, as “Word18” (“taste”) is a commonly used word, the combination with other taste-related words, such as “Word8” (“dry”) would be a candidate for appealing the product on SNS. Actually, the result of the inference is shown in Table 4-3.

Table 4-3. A case of the inference not correspondent with the actual data

Case1	$P(engaged = 1 \mid Word8 = 1, Word18 = 0)$	78.1%
Case2	$P(engaged = 1 \mid Word8 = 1, Word18 = 1)$	80.8%

Table 4-3 shows, tweets are likely to be engaged with the product when “Word8” is used with “Word18”. On the other hand, actual search on Twitter is performed under these two search words,

Conditon1: correspondent to Case1: the product and Word8
Conditon2: correspondent to Case2: the product and Word8 and Word18

The emerging ratio of tweets that match condition 1 is 58.0%. As for condition 2 is 9.0%. The actual search result on Twitter tells “Word18” should not be used with “Word8” in order to make “Word8” engaged with the product. The cause of this disagreement would reside in the characteristic of Bayesian network. In this case, the node “engaged” is a kind of explained variable, which has different role from other nodes acting as explanatory variables. But Bayesian network usually handles all these nodes equally. Therefore, some adjustment is required.

4-4. Proposed Method

To resolve the issue in Section 4-3, the author proposes to optimize Bayesian network by configuring initial state which reflects and emphasizes relationships between the target nodes and others. The author configures initial state to Bayesian network leveraging the result of Random forest, as it is a proven method for analyzing the influence of explanatory variables upon explained variable. The initial state consists of a few nodes around the target node and several

edges between these nodes and the target. The former is called “initial nodes” and the latter is called “initial edges.

The proposed methods consist of two steps, selecting of initial nodes and decision of direction of initial edges. In the former, Random forest is applied, because it can prevent bias in the data and can also perform simulation with the generated result, which is required in the second step. The detail is described in the following section “4-4-1. Applying Random forest” and “4-4-2. extracting initial nodes via decreasing of Gini coefficient”. In the latter, Wasserstein metric is calculated to compare two probability distribution required to determined the direction of initial edges. The detail is described in the following section “4-4-3. Designating initial edges according with Wasserstein metric” Finally, in the section “4-4-4. Inference with adjusted Bayesian network”, the adjusted Bayesian network is obtained. The details of each step are described in the following, respectively.

4-4-1. Applying Random forest

At first Random forest analysis is applied to the dataset, in which “engaged” is explained variable and 18 engaged keywords (“Word1” - “Word18”) are explanatory variables. Random forest is an advanced algorithm based on decision trees, (4-3:Ali et al., 2012) in which a lot of trees are generated according with randomly selected explanatory variables and the result is obtained as major vote of those trees. Therefore, these two parameters should be given properly in advance.

Parameter1: Number of explanatory variables selected while generating trees
Parameter2: Total number of trees generated

Along with the result of grid searching approach, (4-4:Jimenez et al., 2007) parameter1 is set to 4 and parameter 2 is set to 500 in this study. The dataset is split into train data (80% of 1046 records) and the remining is left for out of bag check. The estimated error ratio in out of bag check is 29.2%, which is higher
than usual task for decision making. Because it is not decisively determined whether a tweet is engaged or not in this example.

4-4-2. Extracting initial nodes via decreasing of Gini coefficient

The second step of the proposed method is to extract a few explanatory variables as initial nodes. Initial nodes should be explanatory variables which have more influence on the target (explained variable). While processing Random forest, Gini coefficient (4-5: Yizhaki, 1979) as defined in Equation 4-1 is calculated for a node in each tree.

$$Gini(i) = \sum_{k} p(i, engaged = k) \times \big(1 - p(i, engaged = k)\big) \qquad (4-1)$$

$Gini(i)$: Gini coefficient of node i

$p(i,\ engaged = k)$: frequency ratio of record within node i, of which value of "engaged" is k (=0 or 1)

Furthermore, Gini decrease according to an explanatory variable is defined as in Equation 4-2.

$$GiniDec(w) = \sum_{i \in Node(w)} N(i) \times Gini(i) - N\big(left(i)\big) \times Gini\big(left(i)\big) - N\big(right(i)\big) \times Gini\big(right(i)\big) \qquad (4-2)$$

$GiniDec(w)$: Gini decrease according to explanatory variable w

$Node(w)$: a set of nodes split by w

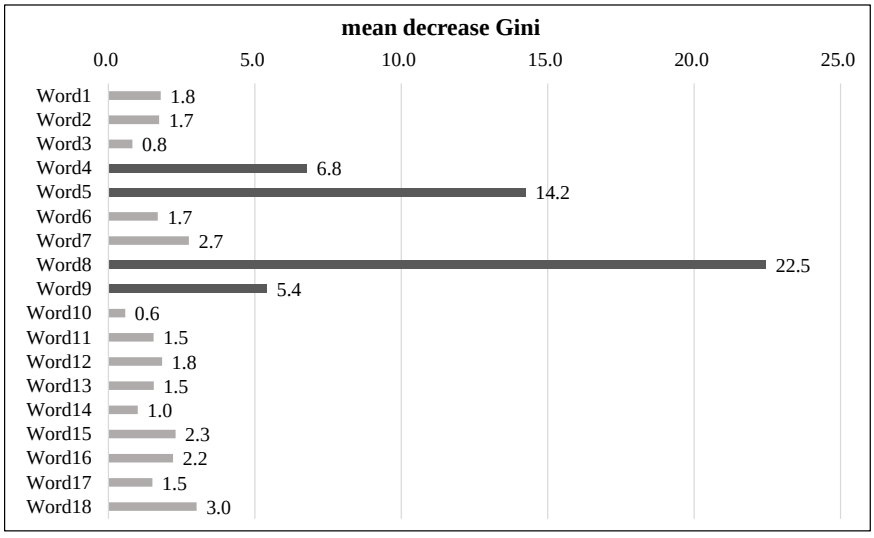
$N(i)$: number of records in node i

$left(i)$: left child node of node i split by w

$right(i)$: right child node of node I split by w

Gini coefficient represents impurity of records within a node according to classification via the target. Gini decrease represents how much impurity is improved by generating trees according with an explanatory variable. That means explanatory variables with higher Gini decrease have more influence on the target. By calculation the mean of Gini decreases across the trees generated in Random forest, importance of explanatory variables can be compared. The mean decrease Gini values of engaged keywords are shown in Figure 4-3.

Figure 4-3. Values of mean decrease Gini of engaged keywords

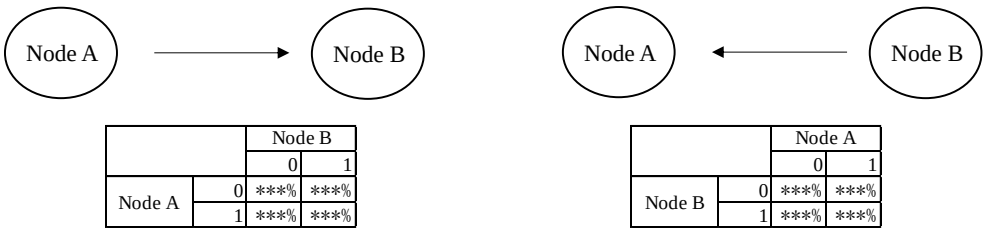


As highlighted in Figure 4-3, Word4, Word5, Word8 andWord9 have relatively more influence on the target. Therefore, these four explanatory variables are selected as initial nodes.

4-4-3. Designating initial edges according with Wasserstein metric

The latter half of configuring initial state is to designate initial edges between initial nodes and the target. As shown in Figure 4-4, in Bayesian network, an edge from Node A to Node B represents that the conditional probability of Node A given Node B is defined and vice versa.

Figure 4-4. Relationship between direction of edge and conditional probability



		Node B	
		0	1
Node A	0	***%	***%
	1	***%	***%

		Node A	
		0	1
Node B	0	***%	***%
	1	***%	***%

In order to properly reflect the influence of initial nodes on the target, initial edges also should be defined leveraging the result of Random forest analysis. If one direction is more appropriate than the other in terms of stochastic model, its conditional probability is also more similar to the distribution of the dataset than the other. As described in Section 4-4-1, 20% of the dataset is remained for out of bag check. By using this remained data, predicted result of the target node can be obtained. Then two values of the conditional probability for initial nodes are obtained. One is from the remained data (true distribution) and the other is from predicted data (learned distribution). The purpose here is to determine which is closer between the conditional probability in the left of Figure 4-4 or in the right when true distribution and learned distribution are compared.

Kullback-Leibler divergence (4-6:Shelens, 2014) is often used to compare several probability distributions, but it can't be applied in this case. Because it is a measure for bringing learned distribution closer to a specific true distribution and does not provide the distance between two different distributions. On the other hand, Wasserstein metric (4-7: Vallender, 2006) which is one of methods in the area of optimal transport, is commonly used for defining the distance between two different probability distributions. Given two probability distributions A and B, Wasserstein metric of p-th order is described as in Equation 4-3.

$$WS(A,B) = \{\inf E(\, d(A,B)^p)\, \}^{\frac{1}{p}} \qquad (4-3)$$

d: the distance defined in the domain of the two probability distributions.

E: Expected value of probability distribution

inf: Operation to calculate the infimum

As the domain of probability distributions are node values which are 0 or 1 in this case, it is sufficient to calculate the absolute value among each random variable. Therefore, the order of Wasserstein metric is configured as 1. The above definition is described by expectation of the probability distribution, but here the probability value corresponding to the combination of {0,1} and {0,1} is calculated from the data. Let $WS^*(nodeX, nodeY)$ be Wasserstein metric between true distribution and learned distribution in terms of the conditional probability according with an edge from *nodeX* to *nodeY.*

Smaller value of $WS(A, B)$ means the two probability distributions A and B are closer. Therefore, the direction of edge between the target ("engaged") and initial edges ("Word4", "Word5", "Word8", "Word9") are determined by comparing $WS^*(engaged, word*)$ and $WS^*(word*, engaged)$.The corresponding edge of smaller WS^* value should be designated as initial edges. The result is shown in Table 4-4.

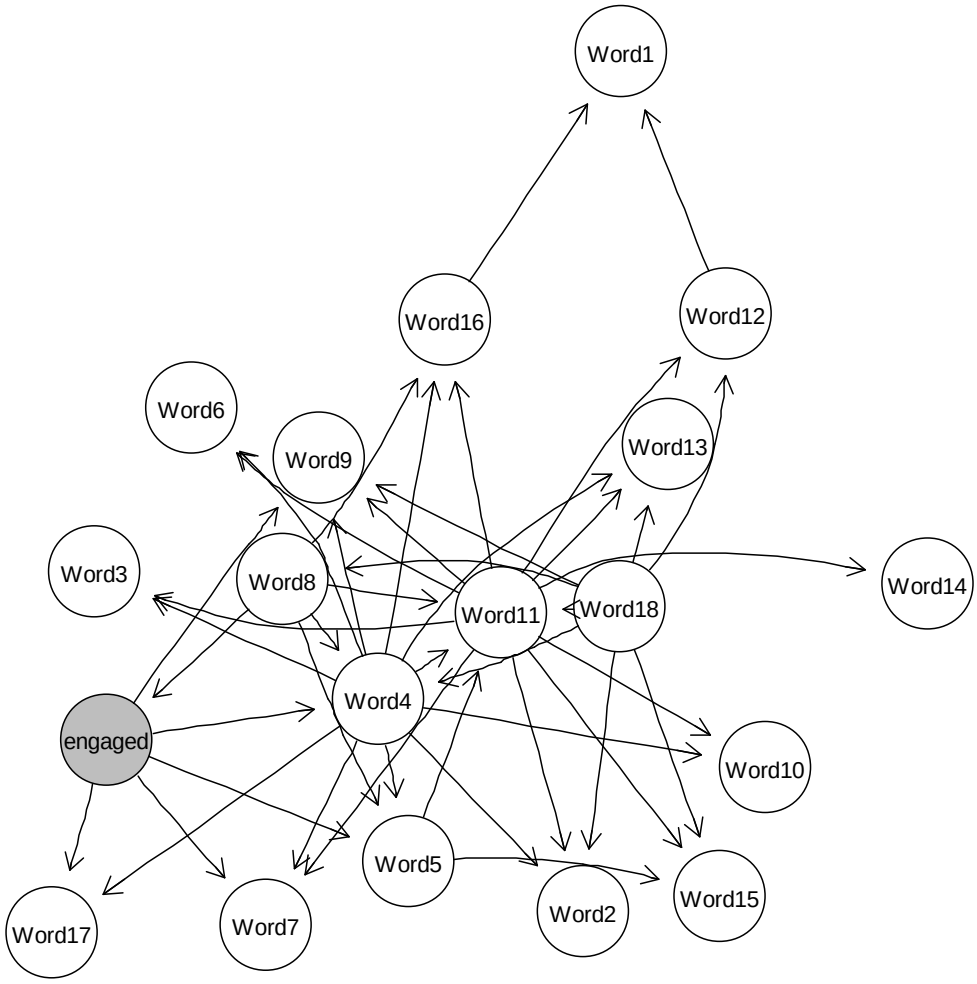
Table 4-4. Wasserstein metric and designation of initial edges

	WS^*(word*, engaged)	WS^* (engaged, word*)	designated direction of edge
Word4	0.269	0.071	engaged ⇒Word4
Word5	0.304	0.240	engaged ⇒Word5
Word8	0.217	0.260	Word8 ⇒engaged
Word9	0.010	0.006	engaged ⇒Word9

4-4-4. Inference with adjusted Bayesian network

By the procedure in Section 4-4-3, the initial nodes and initial edges are obtained. In the final step, these two conditions are set into initial state of Bayesian network and structure learning is performed in the same way in Section 4-3. The result of Bayesian network analysis with initial state is shown in Figure 4-5.

Figure 4-5. Result of Bayesian network with initial state



The target node ("engaged") is directly connected with engaged keywords such as Word4, Word5, Word7, Word8, Word9, Word17. Different form ordinal Bayesian network in Chapter 2, the edge between Word4 and the target is opposite and the edge between Word18 and the target is eliminated. The initial state leveraging the result of Random Forest analysis makes this difference.

As described in the next chapter 4-5, Bayesian network with initial state coincides well with actual search result on Twitter. Therefore, Bayesian network in Figure 4-5 is considered to be adjusted for decision making as for the target node.

4-5. Result and Conclusion

For confirming effectiveness of adjusted Bayesian network in Figure 4-5, probabilistic inference as for "Word8" and "Word18" in the same ways as in Section 4-2 is performed. The result is shown in Table 4-5.

Table 4-5. Improvement in problematic case

		Original network	Adjusted network
Case1	$P(engaged = 1 \mid Word8 = 1, Word18 = 0)$	78.1%	78.9%
Case2	$P(engaged = 1 \mid Word8 = 1, Word18 = 1)$	80.8%	78.9%

As described in Section 4-2, the actual search result on Twitter shows “Word18” should not be used with “Word8”. Original network falsely suggests that a tweet is more likely to be engaged with the product when “Word8” and “Word18” are used together. Adjusted network does not recommend this combination for improving engaged ratio. Word8 is a very common word, "taste," and is not appropriate to use as an appealing keyword. From this perspective of business knowledge, the adjusted Bayesian network can be regarded as producing more desirable results. Although. it would be best if the value of Case2 is much lower than that of Case1 in Adjusted network, the adjusted one provides better result than the original.

In the same way, as for nodes which are directly connected to the target (“engaged”) with the same directions, such as Node5, Node7, Node8, Node9 and Node17, the comparison of the inference between original and adjusted network and the emerging ratio in actual search on Twitter are shown in Table 4-6.

Table 4-6. Comparison between original and adjusted network with actual search result

	Original network			Adjusted network			Emerging ratio on Twitter search	
	Word18 =0 (*1)	Word18 =1 (*2)	Diff (*2) - (*1)	Word18 =0 (*3)	Word18 =1 (*4)	Diff (*4) – (*3)	without Word18	with Word18
Word5	62.7%	61.8%	-1.0	64.1%	59.6%	-4.4	39.0%	3.0%
Word7	15.8%	14.2%	-1.6	16.7%	12.8%	-3.9	5.0%	1.0%
Word8	78.1%	80.8%	2.7	78.9%	78.9%	0.0	58.0%	9.0%
Word9	0.4%	1.2%	0.8	0.4%	1.1%	0.7	1.0%	1.0%
Word17	1.8%	2.4%	0.6	2.0%	2.0%	0.1	0.0%	0.0%

The search conditions used for “Emerging ratio on Twitter search” column are as follows. “Word*” is “Word5”, “Word7”, “Word8”, “Word9” or “Word17”.

Conditon1:

correspondent to the item “without Word18”: the product and Word*

Conditon2:

correspondent to the item “with Word18”: the product and Word* and Word18

The actual search result shows, “Word18” is not appropriate for combined use with other words for making marketing messages engaged with the product. The Diff column in Table 4-6 of adjusted network shows less value than that of original network. That means adjusted network less recommend the use of “Word18” with other words than original one. That also means adjusted Bayesian network leveraged with the result of Random forest matches better than ordinal one in terms of decision making concerned with particular target node. In this way, by initializing nodes and edges according to the result of Random forest, it becomes possible to adjust the structure of Bayesian network and to make it match the actual domain knowledge.

5. Search for Keywords in Appealing Important Factors

5-1. Background

In this section, of the three issues raised in Section 2, the proposed method to solve the following Issue 3 is described.

Issue 3: It is difficult to determine what keywords are best for appealing in order to enhance the effect of the factors to be focused on.

If enterprises appeal their products and important factors related with the products on the Internet, it is important to understand, how much content is on the Internet that includes products and factors? or are there any content producers who voluntarily publish them? Then enterprises support such contents and content producers or use them as a reference for the appeal. In this situation, factors can be considered as criteria for searching proper contents or producers. These conditions are called “criterial keywords” in this study.

With a small number of criterial keywords, the number of obtained contents and producers will be also limited, so enterprise need to add some keywords related with criterial keywords according to their similarities. So, the author proposes the methods for expanding keywords leveraging Word2Ved and for classifying the expanded keywords with hierarchical clustering.

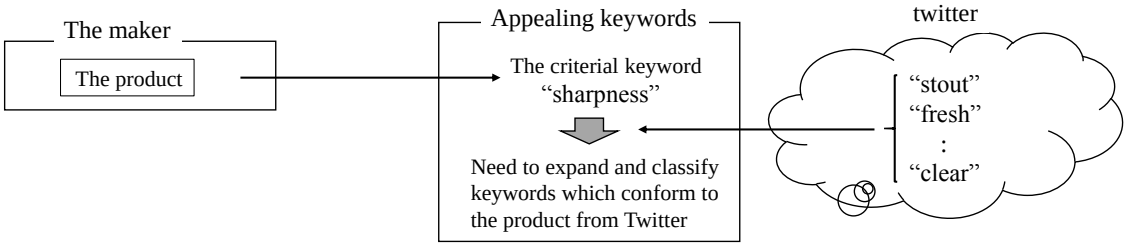
5-2. Example Case

The methods proposed in this study can be applied to any case where a product has been provided alongside a keyword that characterizes it, and the keyword needs to be expanded using keywords collected from SNS as shown in Figure5-1. Here, explanation and verification of the proposed methods is performed using the example case of a beer maker. The maker wishes to promote its primary product, which possesses the characteristic of “sharpness”. Thus, the maker wishes to expand this expression using more appealing keywords collected from SNS (Twitter in this example). However, the methods described in this section do not depend on any particular scenario. In fact, in Section 6 application for product planning and development projects is described.

In the following subsections in Section 5, the beer is denoted by “the product”, “sharpness” is denoted by “the criterial keyword”, and the proposed analytical procedures are explained using this analogy. However, all the methods described below are also applicable to a wide range of business as long as the purpose is to expand keywords characterizing a product and classify them

based on the activities of consumers on SNS.

Figure 5-1. Example case



5-3. Issue to be Solved

If the maker considers that “sharpness” well represents a characteristic of its product, this becomes one of criterial keywords. With a small number of criterial keywords, the number of obtained contents and producers will be also limited, so enterprise need to add some keywords related with criterial keywords according to their similarities. For this purpose, there are already several effective methods. Co-occurrence network analysis is one of popular methods used to measure similarities among data, based on indices such as the Jaccard coefficient. Advanced studies on this method have also combined it with other analysis techniques, such as Analytic Hierarchy Process (AHP) (5-1: Garg & Kumar, 2018; 5-2:Angelo et al., 2018)

However, keywords added by enterprises themselves might not always capture the appealing facets of the product effectively. And co-occurrence network analysis does not offer any means to expand criterial keywords based on the relevance to the product. On the other hand, general consumers may have captured the essence of the product more effectively on their social networking services (SNS). Thus, it would be useful for enterprises to expand keywords for appeal of their products properly and efficiently by extracting associated keywords from SNS.

Therefore, in this study the author proposes a method to expand keywords by leveraging word embedding. Word2Vec (5-3: Mkolov et al., 2013) is a word embedding technique that converts each keyword into a vector, and proceeds to calculate similarities between pairs of words by computing the inner product between the two corresponding vectors. By vectorizing keywords, addition among keywords becomes possible. By using this characteristic, it is possible to search for and expand new keywords that are close to both the product and the criterial keywords.

In this study, besides proposing a method to expand criterial keywords by leveraging Word2Vec, the author proposes the use of hierarchical clustering to classify the expanded keywords. A

combination of these two methods will enable enterprises to expand their criterial keywords effectively and to decide the priorities of expanded keywords.

5-4. Proposed Method

To solve the issues described in Section 5-3, the following two Task are required.

Task1: Select expanded keywords based on the product and the criterial keyword

Task2: Prioritize expanded keywords

For Task1, it is not enough to simply select keywords that are frequently tweeted with the product and the criterial keywords. It is required to measure closeness between the candidate and the product and the criterial keyword at the same time. For this purpose, Word2Vec is applied, as it can measure closeness among keywords via inner product of vectors and also can consider the product and the criterial keyword together by summation of vectors (that is called the compound vector).

For Task2, it is required to prevent a bias such as selecting expanded keywords included in the same tweet. For this purpose, hierarchical clustering is applied to classify expanded keywords based on their appearance tendency in tweets. If there are multiple expanded keywords that show the same appearance tendency, give priority to the one that is close to the compound vector and also has a high frequency of appearance.

The proposed methods consist of the four steps, “data collection”, “data preprocessing”, “keyword expansion”, and “keyword classification”. The details of each step are described in the following, respectively.

5-4-1. Data Collection

In this step, recent tweets by a sufficient number of users are retrieved. In terms of the frequency of occurrence of the product and the criterial keyword, users are classified into the two following categories.

Engaged users:	Users whose tweets include both the criterial keyword and the product
Non-engaged users:	Users whose tweets include only the criterial keyword

If users whose tweets include only the product were grouped together, their tweets would contain very diffused keywords about the product. The motivation of the maker is to create highly

appealing advertising content based on the criterial keyword. In order to prevent dilution due to keyword diffusion, tweets should be obtained from users who are adjudged to be familiar with the product based on their tweets. These are precisely the engaged users, as defined above.

However, some keywords used in the tweets of engaged users might be too close to the product, which can make it difficult for consumers to realize the features of the product in the context of the criterial keyword. Therefore, candidates for expanded keywords should be also collected from the tweets of users who only tweet about the criterial keyword. These are the non-engaged users, as defined above.

This concept of engaged and non-engaged users is also applicable in other cases, when an enterprise wants to identify keywords related to a product based on one of its specific features.

5-4-2. Data Preprocessing

In this step, text data obtained from tweets of the engaged and non-engaged users are decomposed into a set of words. As the data contain orthographic variants, morphological analysis is applied while splitting it into lemmatized words (5-4:Goel et al., 2019; 5-5: Lee & Kim, 2018). “MeCab” is a popular tool, which can handle this process in Japanese and has been used in this study.

5-4-3. Keyword Expansion

In this step, the set of words are converted into vectors using Word2Vec. Word2Vec is a sort of applications of Neural network that represents a given set of words as vectors via contextual comprehension (5-6:Mikolov et al., 2013). It embeds semantic relationships between words into the calculation of the corresponding vectors. For example, if the four words “King”, “Man”, “Woman”, and “Queen” were converted into vectors via Word2Vec, the vector obtained by using the following formula on the corresponding vectors “King” - “Man” + “Woman” would yield the vector corresponding to “Queen” (5-7:Church, 2017). Two algorithms are available to be used in Word2Vec — skip-gram and continuous bag of words also known as CBOW (5-8:Jianqian et al., 2016; 5-9:Carrasco & Sicilia, 2018). The former is supervised to predict the neighboring words of the current word, while the latter predicts the current word based on its neighbors. As the purpose in this step is to expand the criterial keyword, the skip-gram algorithm is implemented in this study.

Word2Vec can be applied to the identification and standardization of derivatives as well. For example, “Java7” and “Java9” can be identified to be different versions of the developing language, "Java" (5-10:Fukui et al., 2019). Furthermore, estimation of similarity via Word2Vec can also be applied to extract keywords representing essential aspects of products (5-11:Jing et

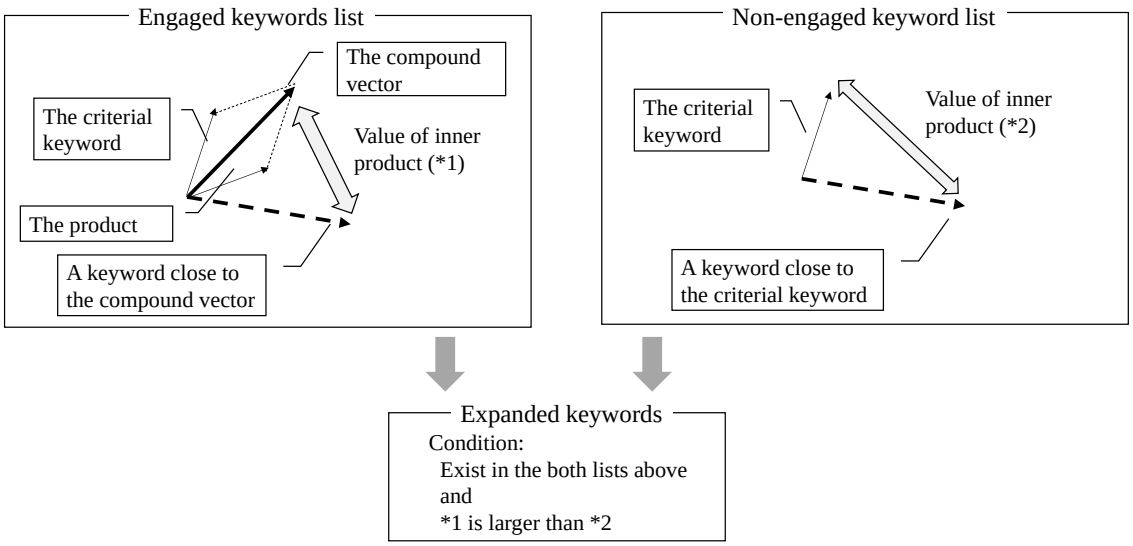
al., 2018) or detect human emotions inherent in social media content (5-12:Jan & Khan, 2020) or semantic analysis (5-13:Kim et al., 2020) and others (5-14:Jatnika et al., 2019; 5-15:Kai et al., 2019; 5-16:Wolf et al., 2014).

The similarity between any pair of vectors can be measured by computing the inner product of the corresponding vectors determined by Word2Vec. The value of the inner product is proportional to the degree of the contextual relationship between the corresponding keywords. By leveraging this advantage, a list of keywords which are close to the criterial keyword are obtained. They comprise the "list of non-engaged keywords". In addition, the vectors corresponding to the product and the criterial keyword are added together to yield the compound vector. Then a list of its neighboring keywords is obtained. They comprise the "list of engaged keywords". Finally, keywords which satisfy the following two conditions are extracted.

・It exists in both the non-engaged list and the engaged list.
・It exhibits higher inner products with the criterial keyword in the engaged keywords list than that in the non-engaged list.

The extracted keywords are close to the criterial keyword and are also related to the product. Thus, they are considered to be expansions of the criterial keyword suitable for appealing the product. An outline of this step is illustrated in Figure 5-2.

Figure 5-2. Outline of keywords expanding



5-4-4. Keyword Classification

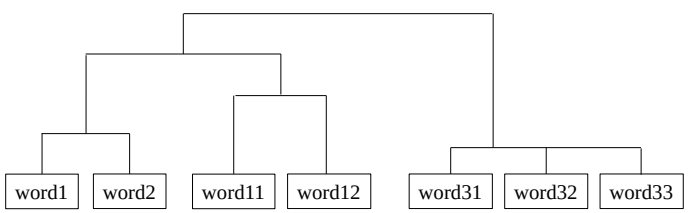
In this step, the expanded keywords obtained in the previous step are prioritized. Although the expanded keywords can be ranked based on their inner products with the criterial keyword, it is necessary to consider the tendency of occurrence as well. For example, suppose there were four expanded keywords A, B, C, and D, such that A and B are often seen in the tweets of some subgroup of users, and that C and D are often seen in the tweets of another subgroup. In order to obtain a comprehensive collection of expanded keywords, representative keywords should not only be chosen between A and B but between C and D. However, if the keywords were selected only based on their inner products with the criterial keyword, C and D might both be excluded. To prevent this improper exclusion, expanded keywords must be first classified based on their occurrence.

For the purpose above, the author applies hierarchical clustering, which is a method that classifies data based on the pair-wise distances between them after regarding them as points in an n-dimensional space (5-17:Lior & Maimon, 2005; 5-18:Chakraborty et al., 2020; 5-19:Xu et al., 2020). The k-means clustering is another method for this kind of classification, which is also attracting theme (5-20:Kim et al., 2020; 5-21:Bai et al., 2020). In k-means clustering, the number of clusters is required to be determined in advance. In real businesses, the number of expanded keywords which can be configured generally depend on related systems or marketing expenses. The number of clusters in hierarchical clustering does not need to be configured in advance. Because of this reason, hierarchical clustering is adopted in this study.

Before applying the hierarchical clustering method, the frequency of each expanded keyword corresponding to each user is calculated based on the tweet data obtained during the 5-4-2. Data Processing step. In other words, if the number of expanded keywords is N and the number of users is M, a dataset of N rows and M column is obtained, as depicted on the left side of Figure 5-3. Then, hierarchical clustering is applied to the dataset and the hierarchical classification among the expanded keywords is obtained in the form of dendrogram, as depicted on the right side of Figure 5-3.

Figure 5-3. Outline of keywords classifying

	user001	user002	…	user100
word1	1	0	…	3
word2	0	4	…	1
:	:	:	:	:
word11	2	3	…	0
word12	3	1	…	1
:	:	:	:	:
word31	1	0	…	0
word32	2	5	…	
word33	0	2	…	4



In order to obtain a comprehensive collection of expanded keywords, one keyword should be chosen from each cluster. Thus, one keyword should be chosen from all possible candidates belonging to the same cluster based on some sort of prioritization, because the number of expanded keywords that can be configured is often limited by associated systems or marketing expenses.

Therefore, the only remaining task is to identify a method to prioritize the keywords. Now, two factors indicate the importance of a keyword — the value of the inner product with the compound vector, which represents the proximity of the keyword to the product and the criterial keyword, and the frequency of occurrence of the keyword, as more frequently used keywords are generally more appealing.

Deviation from the mean is selected as a metric for the first factor. Deviations are more useful to compare similar data than the values themselves. The final values are multiplied by 100 for readability, as the absolute values of the inner products are less than 1. If the frequency of appearance of a word were taken to be the metric for the second factor, the influence of temporarily trending keywords on SNS might be inflated. For this reason, the log value (base 10) of the frequency of occurrence is adopted as the metric in this study.

Finally, the two factors are combined. Because the appeal of a keyword can be considered to be roughly proportional to its frequency of occurrence, the product of the two factors is used as the metric. Therefore, the authors propose the following metric for prioritization of the expanded words.

$$priority(word) = (inn(word) - mean) \times 100 \times log\big(freq(word)\big) \qquad (5\text{-}1)$$

inn: inner product of an expanded keywords with the compound vector
mean: mean of the inn of all expanded keywords
freq: sum of the frequencies of occurrence of each expanded keyword in the tweets of each user

Based on this metric, the priority list of expanded keywords belonging to the same cluster can be ascertained.

5-5. Result and Conclusion

In this section, the author applies the proposed method on an example case described in Section 5-3. 1020 and 2940 tweets were retrieved from engaged and non-engaged users, respectively. The number of tweets retrieved per user was 30. In the case of products like beer, appropriate expanded keywords may vary depending on the season. In this example, for simplicity, one month is taken to be the term of tweet retrieval. This parameter should be changed depending on the characteristic of each product.

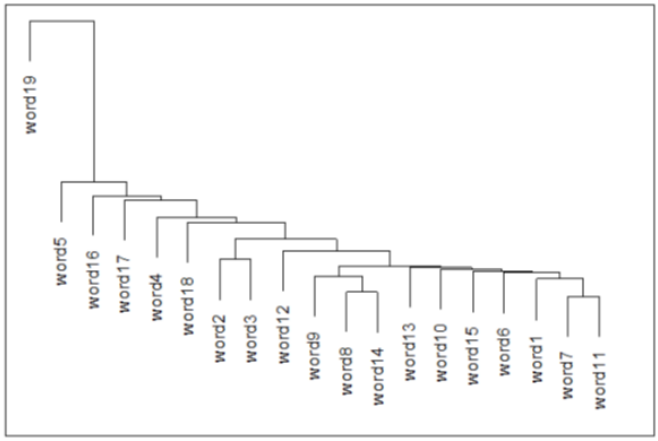
Table 5-1 presents the 19 expanded keywords and associated important factors like *inn* (word), *freq* (word) and the priority index discussed in the Sections 5-4. (In Section 4, these keywords are also used for Bayesian network analysis. But the word "wheat" is omitted, as it is too common word for the purpose of analyze in Section 4)

Table 5-1. Expanded keywords and important factors

	keyword	*inn*(word)	*freq*(word)	priority index
word1	wheat	0.165	3	0.86
word2	rich	0.165	9	1.70
word3	guzzle	0.168	3	1.02
word4	chilled	0.133	5	-1.01
word5	drinkable	0.166	18	2.41
word6	lager	0.171	4	1.46
word7	alcohol percent	0.110	3	-1.76
word8	craft beer	0.181	3	1.63
word9	dry	0.188	6	3.20
word10	Belgium	0.154	10	0.71
word11	fruity	0.165	3	0.83
word12	cheers	0.172	7	2.07
word13	bitterness	0.167	8	1.77
word14	refreshing	0.133	6	-1.09
word15	brisk	0.128	5	-1.34
word16	strongest	0.094	9	-5.06
word17	solid	0.141	13	-0.74
word18	thick	0.097	8	-4.58
word19	taste	0.098	187	-11.26

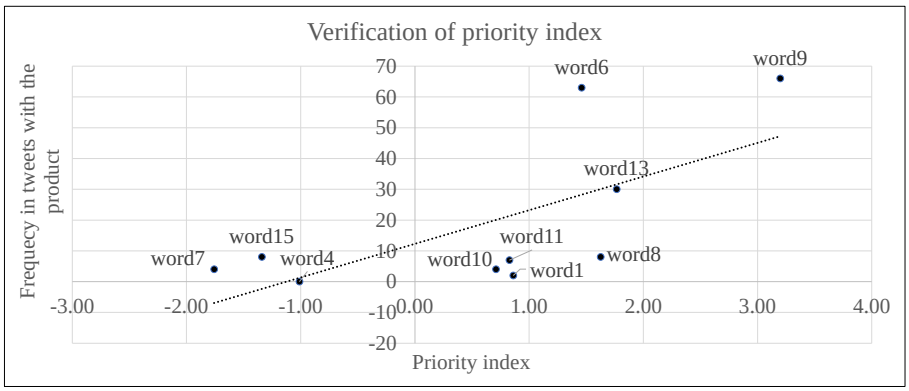
"Dry" is observed to be contextually similar to "sharpness" in the expression of beer, and it exhibited a high priority index in the analysis, too. On the other hand, the index of "taste" was observed to be low in spite of its high frequency. This can be attributed to its low *inn* (word) value, signifying that "taste" does not share a strong relationship with "sharpness" or with the product, as it is a very general word. Thus, the result of the analysis was observed to correspond with daily impressions of consumers about beer. Furthermore, the result of classification of expanded keywords above are shown in Figure 5-4.

Figure 5-4. Classification of expanded keywords



A large cluster including 10 keywords (word9, word8, word14, word13, word10, word15, word6, word1, word7, word11) is seen on the right half of the dendrogram. If one keyword is to be selected for business matching, the keyword "dry" (word9) should be selected, as it exhibits the highest priority index within the cluster, as evident from Table 5-1. However, the effectiveness of the proposed priority index should also be verified. For this purpose, the number of actual tweets including both the product and each expanded keyword in the cluster was computed and compared to its corresponding value in the priority index. The results is shown in Figure 5-5.

Figure 5-5. Verification of priority index



A trend is clearly evident keywords with higher priority index exhibit higher frequency of appearance alongside the product. That implies that the priority index proposed in this study reflects the true relationship between the product and each expanded keyword. Therefore, enterprises can use the proposed method to expand, classify, and prioritize keywords suitable for appealing the product and important factors on the Internet.

6. Systemization of Proposed Methods

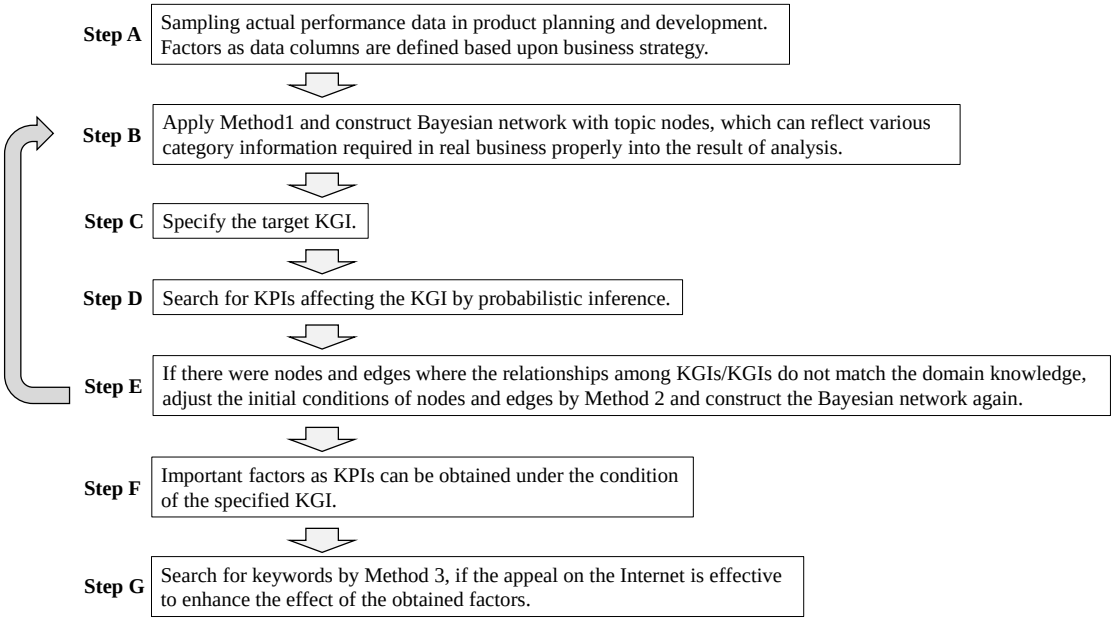
6-1. Combination of Proposed Methods

In Section 6, the purpose of this study will be achieved and confirmed by combining the following three proposed methods described in Sections 3 to 5.

Method1: Analysis Considering Business Type and Annual Sales (Section3)
Method2: Analysis Focusing on Specified Factors (Section4)
Method3: Search for Keywords in Appealing Important Factors (Section5)

Considering the issues that arise in actual business, the three methods are systematized as follows. In Section3, Method 1 is applied to business type and annual sales, but the same approach can be applied to other attributes such as user types or product types. Therefore, in Step B, application of Method 1 is noted as “reflecting various category information required in real business”.

Figure 6-1. Procedure of combination of the three proposed methods

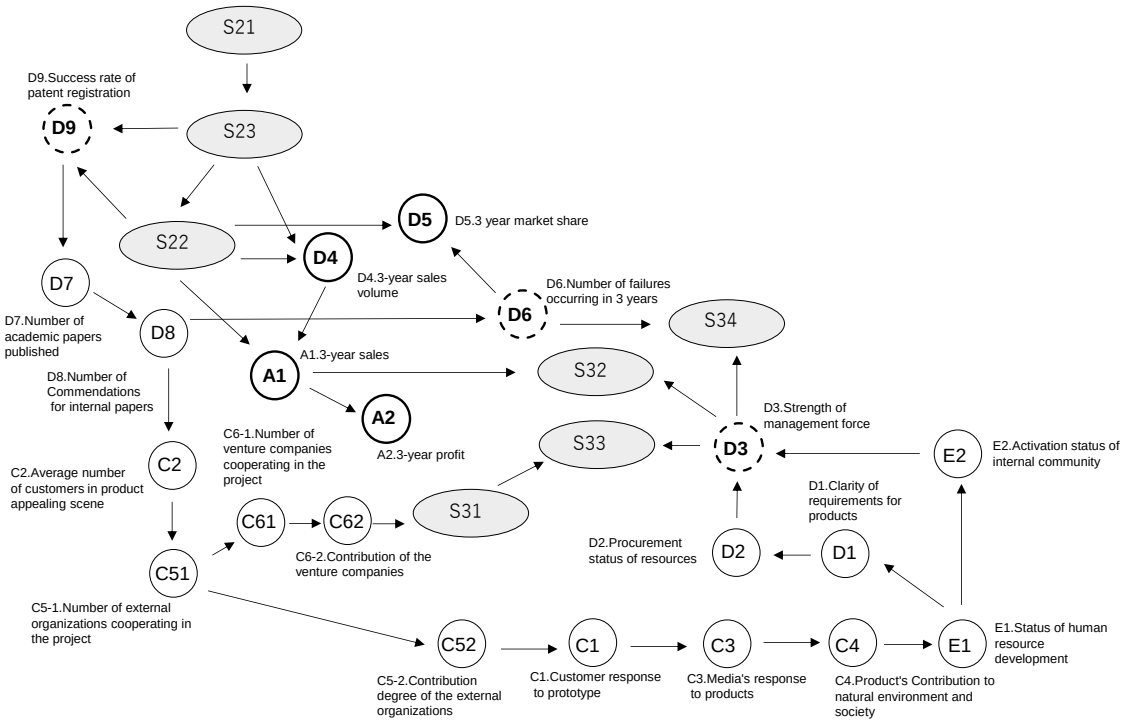


In the following Section 6-2 and 6-3, apply the procedure above to two example cases below.

Example case #1: Find out the factor that most affects annual sales of the product in retailing business.
Example case #2: Is it possible to link efforts to increase the contribution of products to the environment to sales volume, regardless of annual sales?

As for Section 3, the number of the example case records is large enough and Bayesian network that reflects business type and annual sales has already been constructed. Therefore, for two example cases above, the performance data and the network in Section 3 can be applied. That means, in the following Section 6-2 and 6-3, Step A and B in Figure 6-1 have been already completed and start from Step C using the following network obtained in Section 3.

Figure 6-2. Bayesian network that reflects business type and annual sales



6-2. Application to Example Case #1

In this section, following the procedure presented in Section 6-1, analyze the decision making in product planning and development projects shown as the example below.

Example case #1: Find out the factor that most affects annual sales of the produce in retailing business.

At first, in Step C, the KGI, business type and annual sales are designated. As annual sales of the product are represented as Node A1 in Figure 6-2, the KGI is Node A1. And in the same way done in Section 3, business type is configured as retailing business by giving soft evidence to the network. (soft evidence for annual sales is set to default)

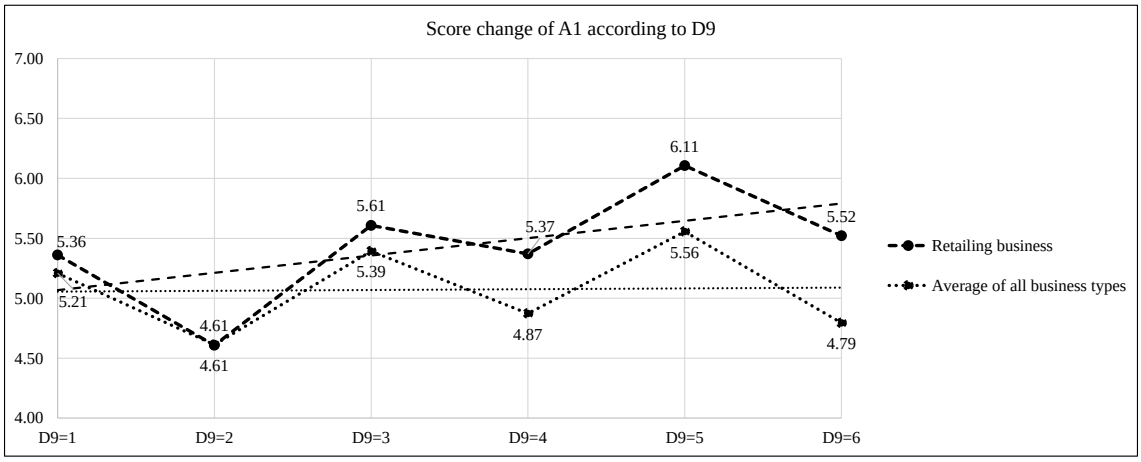
Then move on to step D. By visually checking Figure 6-2, Node D9 is the candidate which shows different trends depending on business type and has a strong influence on Node A1. To confirm this, for all nodes except the four KGIs(A1.3-year sales, A2.3-year profit, D4.3-year sales volume, D5.3 year market share), the variance of the A1 score is calculated by probabilistic inference as these nodes are changed. Table 6-1 shows the result.

Table 6-1. Variance of A1score in retailing business

C1	0.000193	D1	0.001640
C2	0.000278	D2	0.004399
C3	0.000418	D3	0.027061
C4	0.000874	D6	0.004351
C51	0.000077	D7	0.064315
C52	0.000219	D8	0.002960
C61	0.000031	D9	0.236499
C62	0.000017	E1	0.002364
		E2	0.003132

Table 6-1 shows that the variance the A1 score is the largest when D9 is changed. Furthermore, Figure 6-3 compares the A1 score when D9 changes between the retailing business and the average of all business types. (The straight line represents an approximate line) Figure 6-3 shows that in retailing business, the increase of the D9 rank contributes to the increase of the A1 score compared to the average.

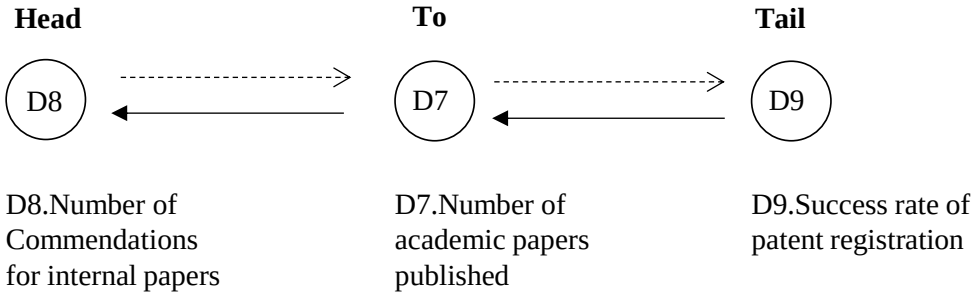
Figure 6-3. Score change of A1 according to D9



Therefore, it can be seen that improving "D9. Success rate of patent registration" is important for increasing product sales in the retailing business. The final goal is to find out what should be done to improve success rate of patent registration, but a new issue arises here.

As shown Figure 6-4, which is a part of Figure 6-2, Node D9, D7 and D8 have a Head-To-Tail relationship.

Figure 6-4. Head-To-Tail relationship among Node D9, D7 and D8



The direction of edges in Bayesian network indicate from which node to which node conditional probability is defined, and are different from causal relationships. As shown by dotted arrows in Figure 6-4, probability propagation is transmitted not only in the edge direction on the network, but also in the opposite direction. Therefore, in Figure 6-4, the change of D8 affects D9 via D7. This relationship among nodes is called Head-To-Tail, which is one of basic structure among nodes in Bayesian network. However, in Head-To-Tail, if the value of intermediate node is given, nodes at both ends are divided and become independent each other. That means, if there were two projects, whose number of academic papers published (D7) is the same, their success rates of patent registration (D9) is irrelevant from their number of commendations for internal papers (D8). However, this does not match experience in actual businesses, and usually both D7 and D8 affect D9. This is the case where Step E in Figure 6-1 is required. So, according to Method 2 in Section 4, adjust the partial structure of Figure 6-2 focusing on Node D9.

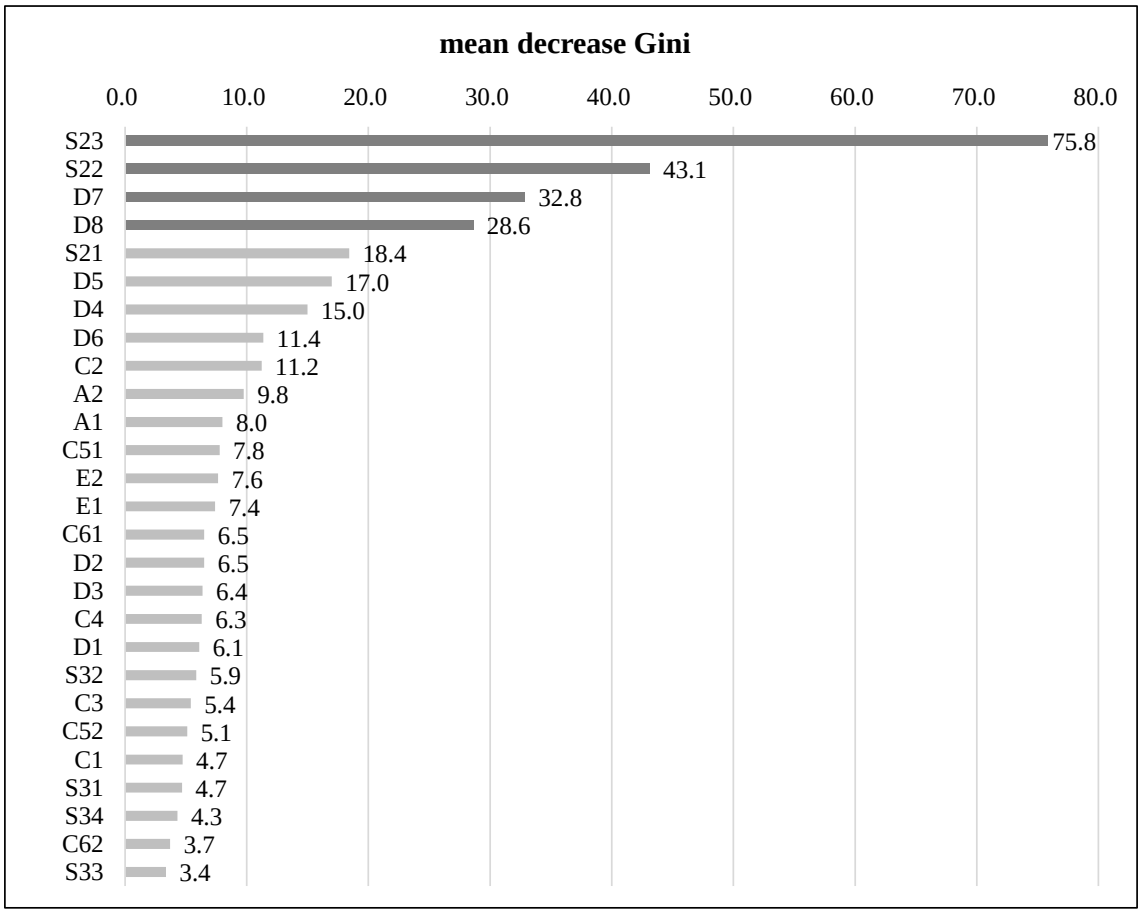
In the same way in Section 4, Random forest analysis is firstly applied, in which Node D9 is explained variable and the rest Nodes are explanatory variables. As already described in Section 4, in Random forest analysis, these two parameters should be given properly in advance.

Parameter1: Number of explanatory variables selected while generating trees
Parameter2: Total number of trees generated

Along with the result of grid searching approach, parameter1 is set to 8 and parameter 2 is set to 500 here. The dataset is split into train data (80% of 992 records) and the remaining is left for out of bag check. The estimated error ratio in out of bag check is 17.6%. Although it is higher than usual task for decision making, but it does not matter, because the purpose here is not to decisively determine the value of D9.

Then, as explained in Section 4, select nodes whose initial state should be adjusted via decreasing of Gini coefficient obtained from the result of Random forest analysis. The mean decrease Gini values of nodes according to D9 are shown in Figure 6-5.

Figure 6-5. Mean decrease Gini values of nodes according to D9



As highlighted in Figure 6-5, Node S22, S23, D7, D8 show higher value of mean decrease Gini and have relatively more influence on D9.

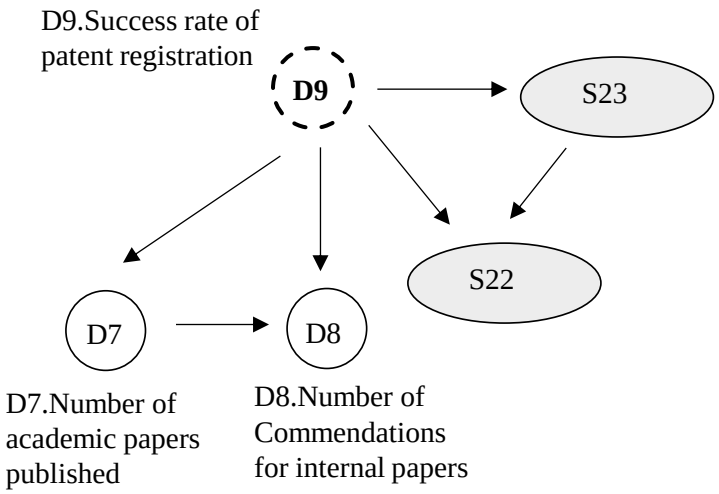
Finally, as described in Section 4, direction of these select nodes are determined by comparing Wasserstein metric between predicted data with the result of Random forest analysis and the remaining data for check. Table 6-2 shows that Wasserstein metric related with four selected nodes. WS (Node, D9) indicates Wasserstein metric of conditional probability when edge is drawn from each node to D9, and WS (D9, Nodes) is vice versa. The corresponding edge of smaller metric value should be designated as initial edges, and the result of comparison is described in "Designated direction of edge" column.

Table 6-2. Wasserstein metric and designation of initial edges

Nodes	**WS (Nodes, D9)**	**WS (D9, Nodes)**	**Designated direction of edge**
S22	0.122	0.102	D9 ⇒ S22
S23	0.571	0.068	D9 ⇒ S23
D7	2.089	0.094	D9 ⇒ D7
D8	0.112	0.100	D9 ⇒ D8

Now that step E of Figure 6-1 is complete, then perform construction of Bayesian network of step B again using the adjusted condition of nodes and edges above. Figure 6-6 is an excerpt of the related nodes from the Bayesian network obtained in this way.

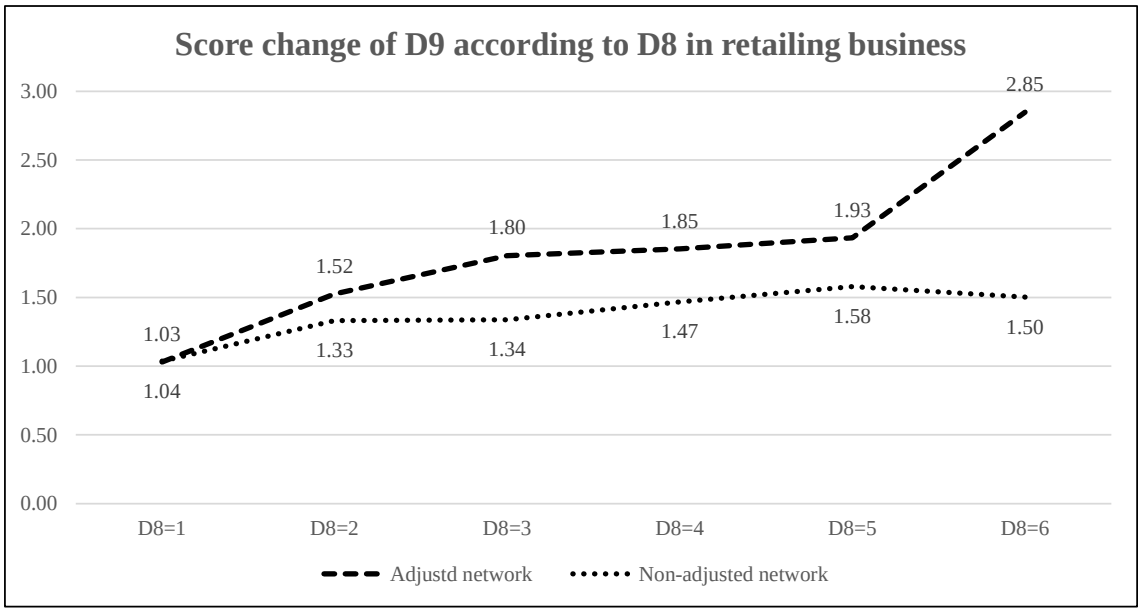
Figure 6-6. Adjusted relationship among node D7, D8 and D9



This network indicates that D9 is influenced by both D7 and D8, and that D7 and D8 are also related. This means success rate of patent registration is influenced by both number of academic papers published and number of commendations for internal papers, and

also means that activities of academic papers published and internal papers are related each other. This coincides with experience in actual businesses. In general, it is less difficult to work on in-house papers than to publish academic papers to the outside. Therefore, if D8 affects D9 positively, encouragement of internal papers would be the affordable first step in order to increase annual sales of product in retailing business. In order to confirm this point, Figure 6-7 shows the comparison between the adjusted network in Figure 6-6 and the non-adjusted one of Figure 6-2 about the score change of D9 according to D8 in retailing business.

Figure 6-7. Comparison of score change of D9 according to D8 in retailing business



In the non-adjusted network, the change of D9 score is flat, but in the adjusted network, where domain knowledge is properly reflected, D9 score increases as the rank of D8 increases. Therefore, in order to improve annual sales of products in the retailing business, it is considered effective to encourage in-house papers at first and then increase the number of patent registration.

6-3. Application to Example Case #2

In this section, following the procedure presented in Section 6-1, analyze the decision making in product planning and development projects shown as the example below.

Example case #2: Is it possible to link efforts to increase the contribution of products to the environment to sales volume, regardless of annual sales?

At first, in Step C and D in Figure 6-1, the KGI, the KPI in concern, business type and annual sales are designated. In Figure 6-2, sales volume of the product is represented as Node D4, and the contribution to the environment is as Node C4. Therefore, the KGI is Node D4 and the KPI is Node C4. As the aim is to see the change of D4 score according to C4 per annual sales category in this case, probabilistic inference is performed with setting of soft evidence for each annual sales category. (soft evidence for business type is set to default) The result of probabilistic inference based on the above settings is Table 6-3. And average rates of increase in D4 score according to C4 rank per annual sales are shown in Figure 6-8.

Table 6-3. Score change of D4 according to C4 per annual sales

	C4=1	C4=2	C4=3	C4=4	C4=5	C4=6
< 500 min yen (1)	2.34	2.37	2.37	2.38	2.40	2.42
between 500 min and 3 bin yen (2)	2.38	2.39	2.38	2.39	2.39	2.40
between 3 bin and 5 bin yen (3)	2.43	2.42	2.42	2.42	2.41	2.41
between 5 bin and 10 bin yen (4)	2.39	2.40	2.40	2.40	2.41	2.42
between 10 bin and 30 bin yen (5)	2.42	2.42	2.42	2.42	2.43	2.44
between 30 bin and 50 bin yen (6)	2.42	2.42	2.42	2.42	2.42	2.42
> 50 bin yen (7)	2.35	2.38	2.37	2.38	2.39	2.40

Figure 6-8. Average rate of increase in D4 score according to C4 rank per annual sales

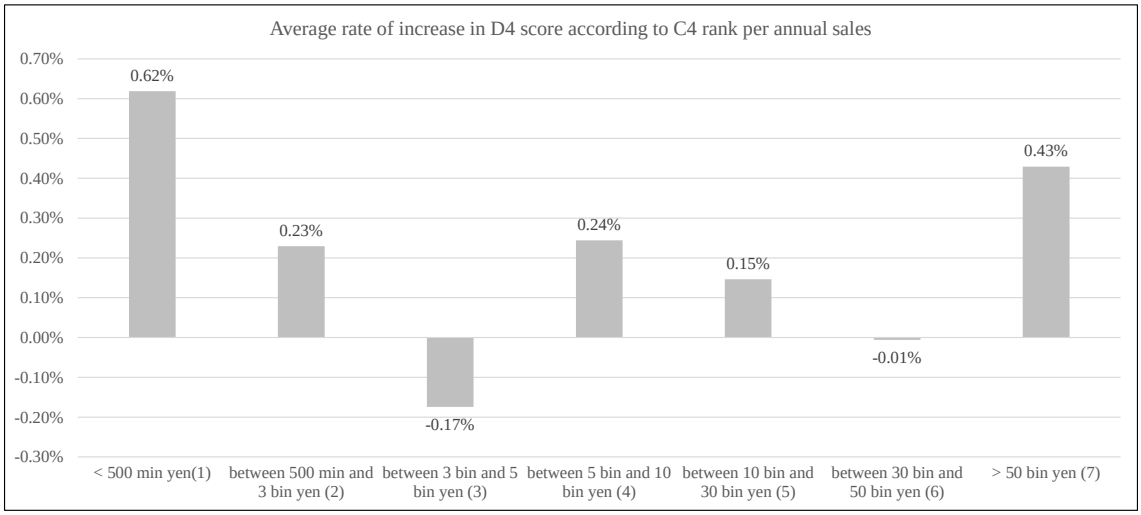
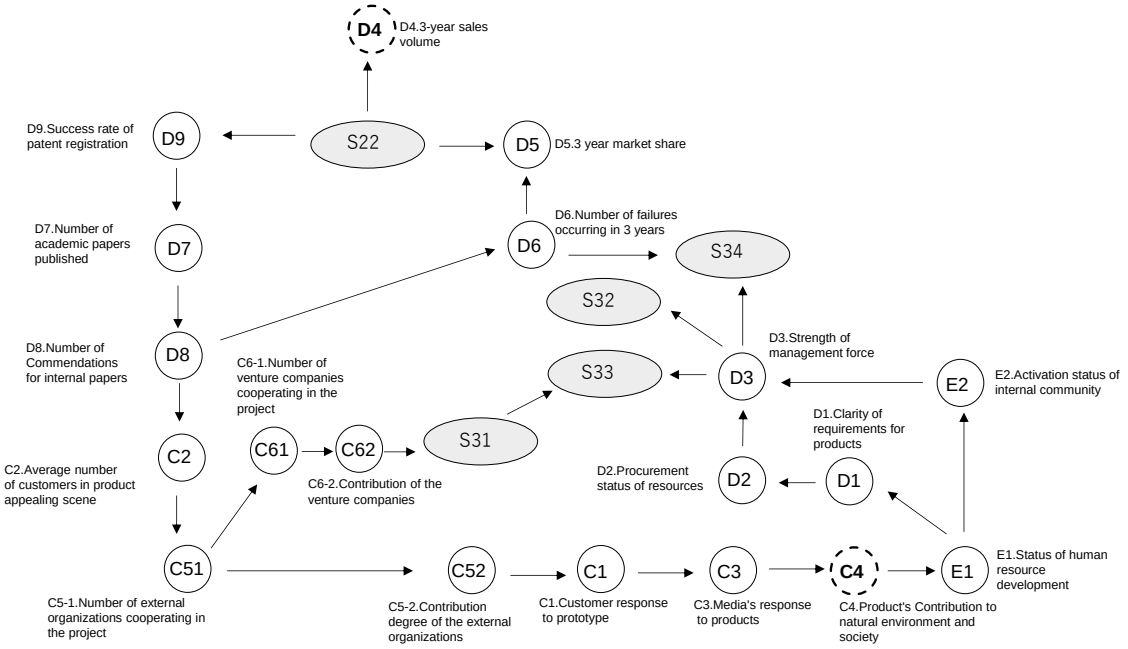


Figure 6-8 shows that average rate of increase in D5 score is relatively high in annual sales categories at the both ends as "<500 min yen (1)" and ">50 bin yen (7)", but between

them are relatively low. This is considered that the former are smaller-scale enterprises whose sales volume are likely to reflect new initiatives in the number of sales, and the latter are larger-scale enterprises which can spend enough budget on the contribution to new initiatives.

However, this result will change when different efforts take place at the same time. Figure 6-9 below is the excerpt of Figure 6-2, which draws path from C4 to D4.

Figure 6-9. Multiple paths from C4 to D4



There are multiple paths from C4 to D4. For example, “C3. Media's response to products” is located on the path which does not include S31, D32, S33 and S34 (these are topic nodes reflecting annual sales of enterprises). And media’s response to product will be expected also effective to the effort for contribution to the environment. Therefore, improving media’s response will be a candidate of the important factors for increasing sales volume regardless of annual sales of enterprises.

In order to confirm this expectation, perform probabilistic inference of D4 score according to C4 under condition C3=5, which means media’s response to products is good. The result is shown in Table 6-4 and its average rate of increase in Figure 6-10 respectively.

Table 6-4. Score change of D4 according to C3 per annual sales under condition C4=5

	C3=1	C3=2	C3=3	C3=4	C3=5	C3=6
< 500 min yen (1)	2.40	2.40	2.40	2.40	2.40	2.40
between 500 min and 3 bin yen (2)	2.39	2.39	2.39	2.39	2.40	2.40
between 3 bin and 5 bin yen (3)	2.41	2.41	2.41	2.41	2.41	2.41
between 5 bin and 10 bin yen (4)	2.41	2.41	2.41	2.41	2.41	2.41
between 10 bin and 30 bin yen (5)	2.43	2.43	2.43	2.43	2.43	2.43
between 30 bin and 50 bin yen (6)	2.42	2.42	2.42	2.42	2.42	2.42
> 50 bin yen (7)	2.39	2.39	2.39	2.39	2.39	2.39

Figure 6-10. Average rate of increase in D4 score according to C3 rank per annual sales under condition C4=5

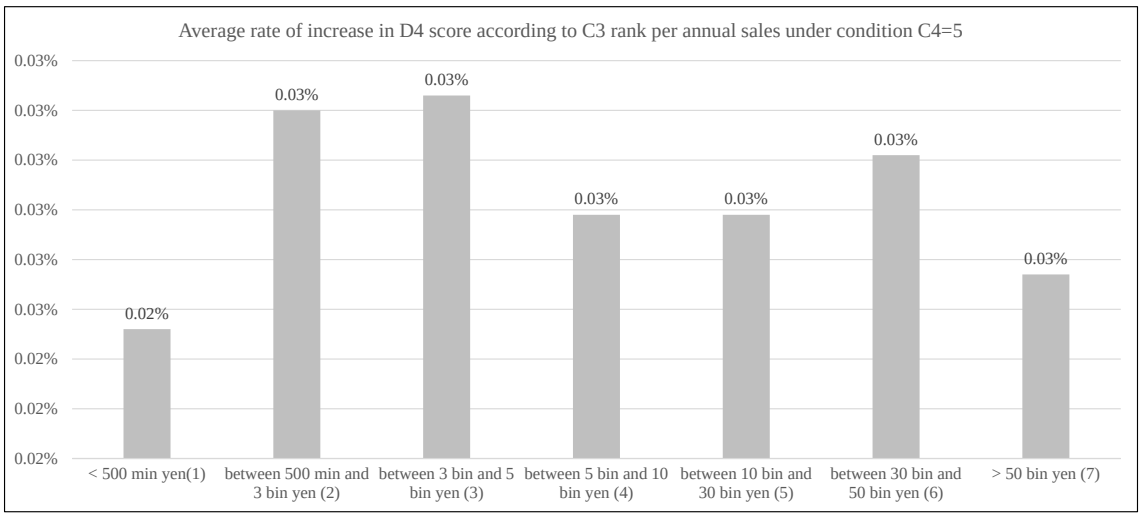


Table 6-4 and Figure 6-10 shows that increase of C4 rank positively improve D4 score under condition C4=5, regard less of annual sales category. Therefore, in order to increase sales volume by contributing to the environment of the product, appeal to media, for example via the Internet, is an important factor common to all annual sales categories.

Now that Step F in Figure 6-1 is complete, proceed to Step G. (In this case, Step E is not required, as there is no conflict against domain knowledge)

In Step G, following Method 3 in Section 5, search for keywords that are effective in appealing the contribution of the product to the environmental on the Internet.

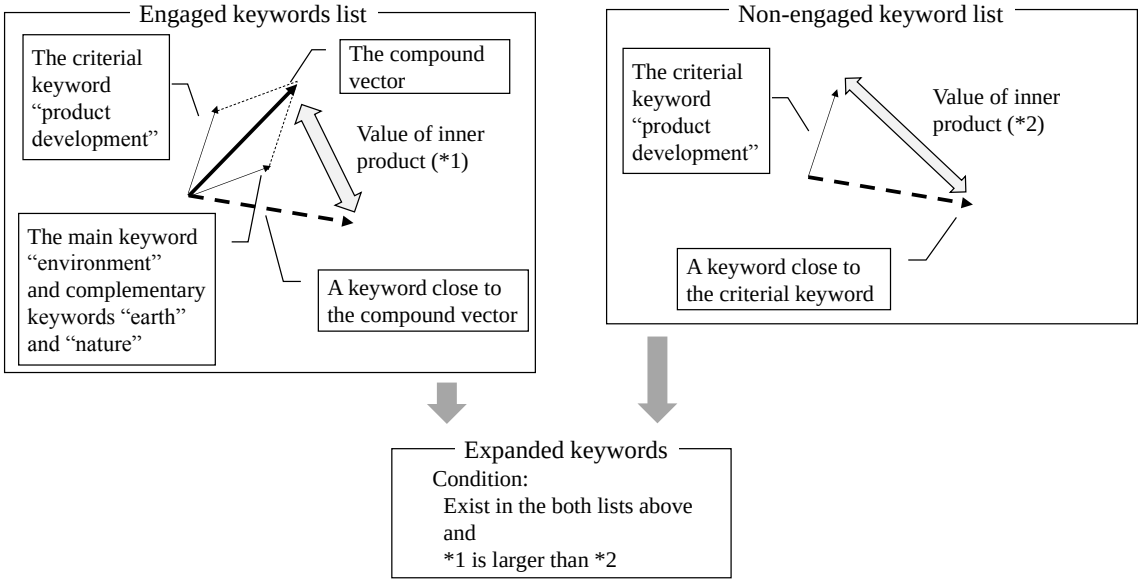
As explained in Section 5, in Method 3, the main keyword and the criterial keyword are

firstly defined. After that, the engaged list containing both the main and the criterial keyword and the non-engaged list containing only the criterial keyword are obtained from Twitter. And the keywords satisfying the following two conditions are extracted as expanded keywords for appealing.

Condition 1: Exist both in the engaged and in the non-engaged list
Condition 2: Closer to the main keyword in terms of inner product of keyword vector calculated by the result of Word2Vec analysis

In this case, the theme is contribution to the environment in product planning and development projects, so the main keyword is defined as “environment” and the criterial keyword is defined as “product development”. Furthermore, since keywords related to the program development environment etc. are mixed in with a simple "environment", “earth” and “nature” are added as complementary keywords to the main keyword. The concept of extracting expanded keywords is shown in Figure 6-11.

Figure 6-11. Concept of extracting expanded keywords



Furthermore, in the same way as described in Section 5, the parameters set in the keyword expansion process and the number of acquired data are summarized in the following Figure 6-12.

Figure 6-12. Process for extracting expanded keywords

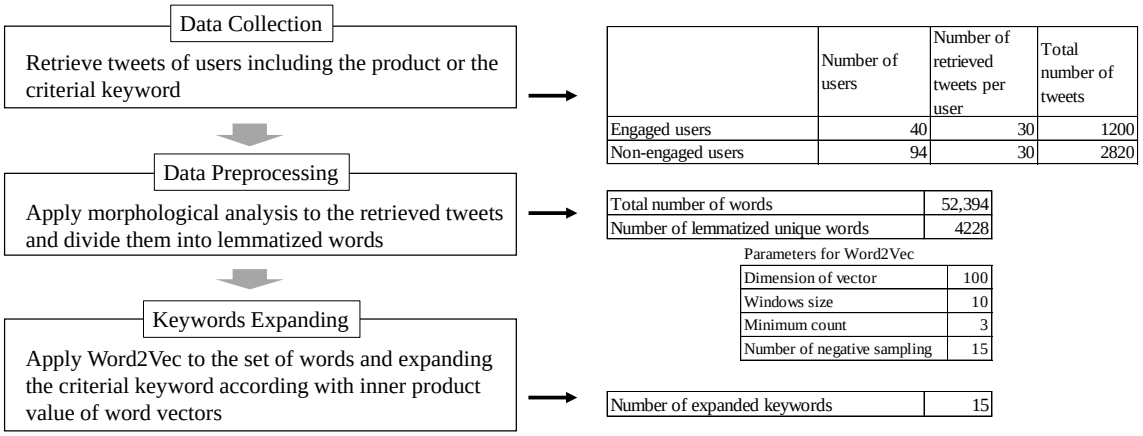


Table 6-5 lists the extended keywords obtained through the process shown in Figure 6-12. Explanation of each column in Table 6-5 are as follows, which are already described in Section 5.

inn(word): Difference defined as A-B. Where A is the inner product value between the keyword and the compound vector, and B is the inner product value between the keyword and the criterial keyword. The larger this value, the closer the relationship with the main keyword under the conditional keyword.

freq (word): The frequency of appearance of keywords in the obtained Tweet data. Frequent keywords will be more appealing.

Priority index: Priority index obtained by multiplying the difference from the average value of inn(word) by 100 times (for proper scaling) and the log value of freq (word).

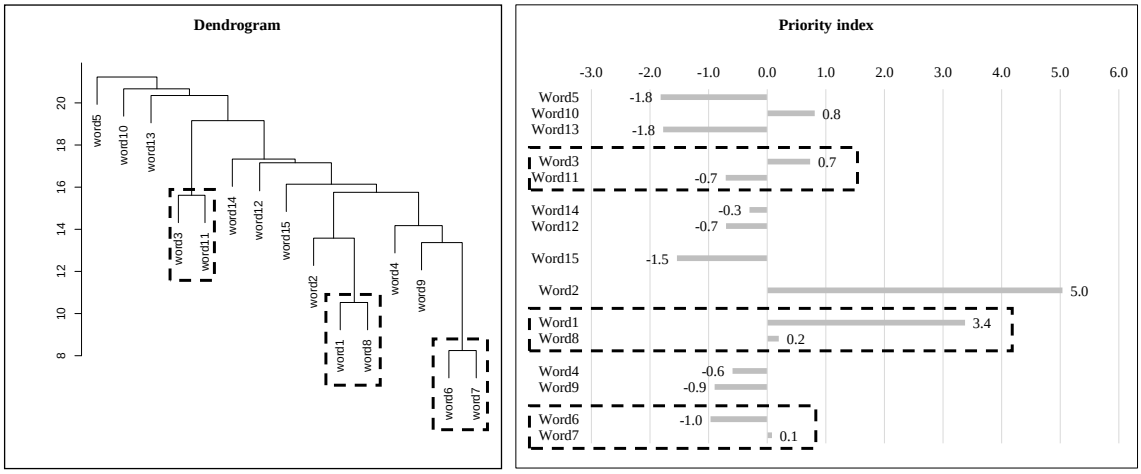
Table 6-5. Expanded keywords

	Keyword	inn(word)	freq (word)	Priority index
Word1	gentle	0.114	3	3.37
Word2	judiciary	0.116	5	5.04
Word3	hole	0.059	3	0.73
Word4	expense	0.031	3	-0.59
Word5	communicate	0.018	5	-1.82
Word6	elucidation	0.023	3	-0.97

Word7	others	0.045	4	0.08
Word8	open	0.048	3	0.20
Word9	make use of	0.025	3	-0.90
Word10	someone	0.054	6	0.81
Word11	feedback	0.029	3	-0.71
Word12	waste	0.029	3	-0.70
Word13	toilet	0.014	4	-1.78
Word14	conversion	0.039	4	-0.30
Word15	announce	0.011	3	-1.54

As already described in Section 5, the final step of Method 3 is to compare the result of hierarchical clustering and the priority index of keywords in Table 6-5. With this comparison, select the keyword with the highest priority from the keywords in the same cluster (keywords that indicate the similar appearance tendency). Because in actual business the number of keywords should be reduced in terms of budget for advertising or marketing. Figure 6-13 shows the dendrogram of Hierarchical clustering and the visualized priority index in Table 6-5.

Figure 6-13. Dendrogram and priority index of expanded keywords



As Word2 “judiciary” shows the highest index, it is the most important expanded keyword. Word3 “hole”, Word1 “gentle” and Word7 “others” has relatively higher value in its cluster designated in dotted line. But the meaning of Word7 is a little vague, so Word1 and Word3 are the secondary important expanded keywords. The value of Word10 “someone” is a little higher too, but its meaning is also vague. Therefore, Word2

“judiciary”, Word3 “hole”, Word1 “gentle” are important expanded keywords for appealing the contribution of the product on the Internet. For example, a phrase would be considered such as “This product clears judiciary restriction for preventing expansion of ozone hole in many foreign countries and is gentle to natural environment”.

6-4. Conclusion and Future Work

As confirmed in Section 6-3, by systematizing the methods of the existing papers that compose this study, it is possible to obtain answers based on probabilistic inference for various decision making in product planning and development projects, taking into consideration business type and annual sales.

In Section 6-3, the Bayesian network was constructed based on actual performance data of product planning and development project in 992 enterprises, but by increasing the number of records, it would be possible to analyze wider range of business type and annual sales.

The following three directions can be considered for further improvement of this study.

1. Analysis considering categories of the product or attributes of the user

In this study topic nodes are introduced to reflect business type and annual sales in Bayesian network analysis. By applying the same approach on categories of the product or attributes of the user, the scope of analysis can be further expanded.

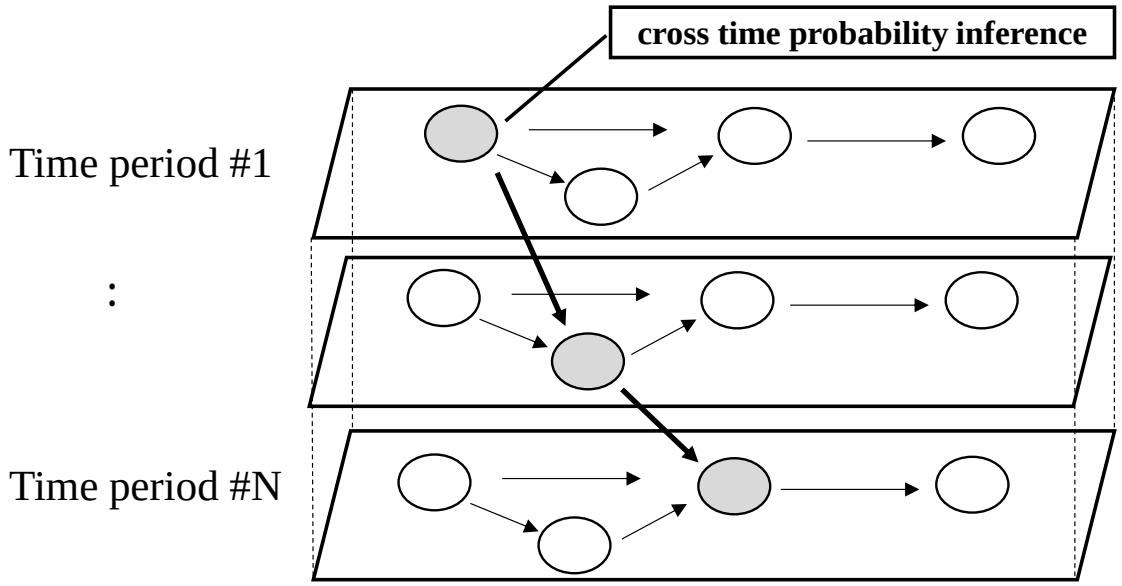
2. Searching and incorporating new factor candidates

It may be possible to search for unknown factors to be analyzed in product planning and development strategy by applying the keyword expansion method in this study. For example, in Section 6 the relationship between the contribution of products to the environment and sales volume was discussed. However, there may be other important factors that intervene between them, such as efforts to make the contribution known in the market. It is expected that the search for these factors will be realized by using the keyword expansion method in this study for case data on the Internet.

3. Clarification of causal relationship with time series

Strictly speaking, edges of Bayesian network are different from causality. On the other hand, in order to select important factors more precisely, it will be necessary to consider time series such that some effect is obtained after a certain period of time after some effort is made. In order to take this situation into consideration, a method such as generating several Bayesian Networks for each time period and performing probability inference across them can be considered, outlined in Figure 6-14.

Figure 6-14. Bayesian network considering time series



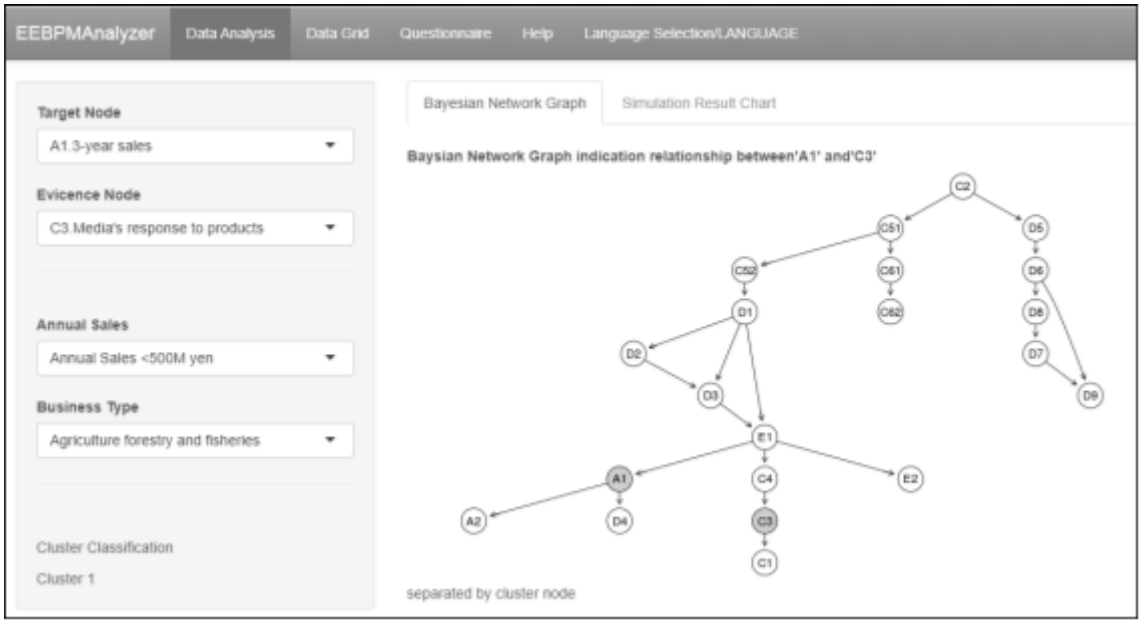
4. Construction of a system using the method proposed in this study

Once the Bayesian network is constructed based on the performance data, the subsequent processes such as configuration of business type and annual sales, probability inference, node adjustment, and search for extended keywords can be automated by programming. Therefore, in the future, it may also be possible to develop a system that assist enterprises themselves to interactively perform decision making leveraging the proposed methods in this study.

Although it is still a prototype, the author has actually developed a decision support system using Bayesian network with the support of the Takuma Laboratory of Chiba

Institute of Technology. As shown in the screen shot in Figure-6-15, with pull down menus, users of this system can select “Target Node” (KGI), “Evidence Node” (the node to which evidence is set, usually KPI), “Annual Sales” and “Business Type”, and then calculate the score of “Target Node” when the rank of “Evidence Node” changes by probabilistic inference. In this prototype system, the process of reflecting business type and annual sales is not enough unlike the method proposed in this study. In addition, it does not have the function to adjust the network structure based on domain knowledge and search for extended keywords to appeal important factors. However, by improving this system, it would be possible to construct an enhanced decision support system leveraging the proposed methods in this study.

Figure 6-15. Prototype of decision making support system with Bayesian network



Furthermore, the method proposed in this study can be applied not only to product planning and development projects but also to various decision making situations in enterprises such as investment, procurement, and recruitment, etc.

In the future, in parallel with developing an enhanced system mentioned above, the author will aim to improve the proposed methods by applying them to other fields other than product planning and development.

References

[1-1] Philip G. Sandner (2011). The market value of R&D, patents, and trademarks. In The Valuation of Intangible Assets, pages 35–72
https://doi.org/10.1007/978-3-8349-8393-0_3

[1-2] Niels C. Petersen (2018). Directional distance functions in DEA with optimal endogenous directions. *Operations Research*, 66(4), 1068-1085
https://doi.org/10.1287/opre.2017.1711

[1-3] Desjeux Y, Dupraz P, Kuhlman T, Paracchini ML, Michels R, Maigne E, Reinhard S. (2015). Evaluating the impact of rural development measures on nature value indicators at different spatial levels,: Application to France and the Netherlands. Ecological Indicators, 59, 41-61
https://doi.org/10.1016/j.ecolind.2014.12.014

[1-4] David I. Stern (1999). Use value, exchange value, and resource scarcity. Energy Policy, 27(8), 469-476
https://doi.org/10.1016/S0301-4215(99)00043-9

[1-5] Sondes Kahouli-Brahmi (2008). Technological learning in energy-environment-economy modelling: A survey. Energy Policy, 36(1), 138-162
https://doi.org/10.1016/j.enpol.2007.09.001

[1-6] Natalia Kuosmanen, Timo Kuosmanen, Timo Sipiläinen (2013). Consistent aggregation of generalized sustainable values from the firm level to sectoral, regional or industry levels. Sustainability, 5(4), 1568-1576
https://dx.doi.org/10.3390/su5041568

[1-7] Shyamika Shiwanthi, Erandathie Lokupitiya, Sena Peiris (2018). Evaluation of the environmental and economic performances of three selected textile factories in Biyagama Export Processing Zone Sri Lanka. Environmental Development, 27, 70–82
https://doi.org/10.1016/j.envdev.2018.07.006

[1-8] Rajesh K. Chandy, Gerard J. Tellis (1998). Organizing for radical product innovation: The overlooked role of willingness to cannibalize. Journal of Marketing

Research, 35(4), 474-487
https://doi.org/10.2307/3152166

[1-9] Viswanathan Krishnan, Karl T. Ulrich (2001). Product Development Decisions: A Review of the Literature. Management Science, 47(1), 1-21
https://doi.org/10.1287/mnsc.47.1.1.10668

[1-10] Judea Pearl (1995). From Bayesian Networks to Causal Networks, Mathematical Models for Handling Partial Knowledge in Artificial Intelligence. Springer, Boston, Massachusetts, pages 157-182
https://doi.org/10.1007/978-1-4899-1424-8

[1-11] Jonathan S. Yedidia, William T. Freeman, Yair Weiss (2003). Understanding belief propagation and its generalizations. Exploring artificial intelligence in the new millennium, 236-239. ISBN:978-1-55860-811-5

[1-12] Ronald D. Anderson, Robert D. Mackoy, Vincent B. Thompson, Gilbert Harrell (2004). A Bayesian Network Estimation of the Service-Profit Chain for Transport Service Satisfaction. Decision Sciences, 35(4), 665-689
https://doi.org/10.1111/j.1540-5915.2004.02575.x

[1-13] Alessio Ishizaka, Ashraf Labib (2013). A hybrid and integrated approach to evaluate and prevent disasters. Journal of the Operational Research Society, 65(10), 1475-1489
https://doi.org/10.1057/jors.2013.59

[1-14] Sucheta Nadkarni, Prakash P. Shenoy (2001). A Bayesian network approach to making inferences in causal maps. European Journal of Operational Research, 128(3), 479-498
https://doi.org/10.1016/S0377-2217(99)00368-9

[1-15] Dinh S. Nguyen (2015). Application of Bayesian networks for product quality management in a multistage manufacturing process. In 2015 IEEE International Conf. on Industrial Engineering and Engineering Management (IEEM), Singapore
https://doi.org/10.1109/IEEM.2015.7385878

[1-16] Musfiqur Rahman, Shamim Ripon (2014). Using Bayesian Networks to Model and Analyze Software Product Line Feature Model. In International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pages 220-231, Springer, Cham
https://doi.org/10.1007/978-3-319-13365-2_20

[1-17] Nobuhiko Kondo, Toshiharu Hatanaka (2019). Estimation of Students' Learning States using Bayesian Networks and Log Data of Learning Management System. International Journal of Institutional Research and Management, Vol.3, No.2, pp. 35-49
http://www.iaiai.org/journals/index.php/IJIRM/article/view/460

[1-18] Yuri Hamada, Tatsuya Maruyama, Hiroko Shoji (2019). Pattern Classification of Value Creative Consensus Building Process in Case of Multiple-choice. International Journal of Affective Engineering, Volume18, Issue 3 Pages 129-136
https://doi.org/10.5057/ijae.IJAE-D-18-00027

[2-1] Thomas Astebro, John L. Michela (2005). Predictors of the Survival of Innovations, J. Prod. Innov. Manag., 2005, 22, pp. 322-335
https://doi.org/10.1111/j.0737-6782.2005.00129.x

[2-2] Nobuaki Arai, Hironori Takuma, Hideo Kameyama (2014). Study on an Evaluation Indicator to Promote Commercialization of R&D Results. Kagaku Kogaku Ronbunshu 2014, 40, pp. 143-145
https://doi.org/10.1252/kakoronbunshu.40.143

[2-3] Stephen B. Olsen (2003). Frameworks and indicators for assessing progress in integrated coastal manage-ment Initiatives, Ocean Coast. Manag. 2003, 46, pp. 347-361
https://doi.org/10.1016/S0964-5691(03)00012-7

[2-4] Nakamura A., Kameyama H., Ohara S. (2001). The Practical Structuring for the Stake-holder Management in ODA -The Optimization for Consensus Building Process in En-vironmental and Social Consideration, J. Int. Assoc. P2M 2011, 6, pp. 15-28
https://doi.org/10.20702/iappmjour.6.1_15

[2-5] Nakayama M., Chyoki S., Oyamada Y., Sekiya T., Mizobe K., Takuma H., Kameyama, H. (2015). Proposals for the Popularization and Establishment of Community-Based Micro-Hydropower Generation Systems. Kagaku Kogaku Ronbunshu

2015, 41, pp. 173-177
https://doi.org/10.1252/kakoronbunshu.41.173

[2-6] Takuma H., Masayuki H. (2015). Discussion of the Value Indicators for Associating Projects with Programs, J. Int. Assoc. P2M 2015, 10, pp. 23-34
https://doi.org/10.20702/iappmjour.10.1_23

[2-7] Lee S. (2011). Bayesian Network Representation, University of Washington, Seattle, WA, USA, 2011
https://courses.cs.washington.edu/courses/cse515/11sp/class2-bayesnet.pdf

[2-8] Yedidia Jonathan S., Freeman William,T., Weiss Yair (2001). Understanding Belief Propagation and its Generalizations, Mitsubishi Electric Research Laboratories, Cambridge, MA, USA, 2001
https://www.merl.com/publications/docs/TR2001-22.pdf

[2-9] Nguyen D.S. (2015). Application of Bayesian networks for product quality management in a multistage manufacturing process. In Proceedings of the 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, Singapore, 6-9 December 2015
https://doi.org/10.1109/IEEM.2015.7385878

[2-10] Carmel A. Pollino, Owen Woodberry, Ann Nicholson, Kevin Korb, Barry T. Hart (2007). Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment, Environ. Model. Softw, 2007, 22, pp. 1140-1152
https://doi.org/10.1016/j.envsoft.2006.03.006

[2-11] Zhepeng Li, Xiao Fang, Xue Bai, Olivia R. L. Sheng (2016). Utility-Based Link Recommendation for Online Social Networks. Manag. Sci. 2016, 63, pp. 1938-1952
https://doi.org/10.1287/mnsc.2016.2446

[2-12] Geng Cui, Man L. Wong, Hon-Kwong Liu (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. Manag. Sci. 2006, 52, pp. 597-612
https://doi.org/10.1287/mnsc.1060.0514

[2-13] Kazuo J. Ezawa (1998). Evidence Propagation and Value of Evidence on Influence Diagrams. Oper. Res. 1998, 46, pp. 73-83
https://doi.org/10.1287/opre.46.1.73

[2-14] Jinhua Pan, Huaxiang Rao, Xuelei Zhang, Wenhan Li, Zhen Wei, Zhuang Zhang, Hao Ren, Weimei Song, Yuying Hou, Lixia Qiu (2019). Application of a Tabu search-based Bayesian network in identifying factors related to hypertension. Medicine (Baltimore), 98(25), e16058
https://dx.doi.org/10.1097%2FMD.0000000000016058

[2-15] Zhuang Zhang, Jie Zhang, Zhen Wei, Hao Ren, Weimei Song, Jinhua Pan, Jinchun Liu, Yanbo Zhang, Lixia Qiu (2019). Application of Tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification. Scientific Reports, 9, 6251
https://doi.org/10.1038/s41598-019-42791-w

[3-1] Iwata T., Sawada H. (2013). Topic model for analyzing purchase data with price information. Data Mining and Knowledge Discovery, 26, 559-573.
https://doi.org/10.1007/s10618-012-0281-y

[3-2] Fei-Fei L., Perona P. (2005). A Bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, California.
https://doi.org/10.1109/CVPR.2005.16

[3-3] Porter M. E., Kramer M. R. (2011). The big idea: Creating shared value. How to reinvent capitalism and unleash a wave of innovation and growth. Harvard Business Review, 89, 1-2.
https://hbr.org/2011/01/the-big-idea-creating-shared-value

[3-4] Tsochantaridis I., Joachims T., Hofmann T., Altun Y. (2005). Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research, 6, 1453-1484.
http://www.jmlr.org/papers/v6/tsochantaridis05a.html

[3-5] Linderman S., Adams R. P., Pillow J. W. (2016). Bayesian latent structure discovery

from multi-neuron recordings. In Advances in Neural Information Processing Systems 29.
http://papers.nips.cc/paper/6185-bayesian-latent-structure-discovery-from-multi-neuron-recordings

[3-6] Heller K. A., Ghahramani Z. (2005). Bayesian hierarchical clustering. Proceedings of the 22nd International Conference on Machine Learning (pp. 297-304).
https://doi.org/10.1145/1102351.1102389

[3-7] Arun R., Suresh V., Madhavan C. E. V., Murthy M. N. N. (2010). On finding the natural number of topics with latent dirichlet allocation. In Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (pp. 391-402). Hyderabad, India.
https://doi.org/10.1007/978-3-642-13657-3_43

[3-8] Cao J., Xia T., Li J., Zhang Y., Tang S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing - 16th European Symposium on Artificial Neural Networks, 72(7-9), 1775-1781.
https://doi.org/10.1016/j.neucom.2008.06.011

[3-9] Deveaud R., SanJuan É., Bellot P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17, 61–84.
https://www.cairn.info/revue-document-numerique-2014-1-page-61.htm

[3-10] Griffiths T. L., Steyvers M. (2004). Finding scientific topics. In Proceedings of the National Academy of Sciences (pp. 5228-5235).
https://doi.org/10.1073/pnas.0307752101

[3-11] Papanikolaou Y., Foulds J. R., Rubin T. N., Tsoumakas G. (2017). Dense distributions from sparse samples: Improved Gibbs sampling parameter estimators for LDA. Journal of Machine Learning Research, 18, 1-58.
https://dl.acm.org/doi/abs/10.5555/3122009.3153018

[3-12] Cooper R. G. (2019). The drivers of success in new-product development. Industrial Marketing Management, 76, 36-47.
https://doi.org/10.1016/j.indmarman.2018.07.005

[3-13] Eliëns R., Eling K., Gelper S., Langerak F. (2018). Rational versus intuitive gatekeeping: Escalation of commitment in the front end of NPD. Journal of Product Innovation Management, 35(6), 890-907.
https://doi.org/10.1111/jpim.12452

[3-14] Gupta S., Dangayach G. S., Singh A. K., Meena M. L., Rao P. N. (2018). Implementation of sustainable manufacturing practices in Indian manufacturing companies. Benchmarking, 25(7), 2441-2459.
https://doi.org/10.1108/BIJ-12-2016-0186

[3-15] Arai N., Takuma H., Kameyama H. (2015). Effect of attorney groupings on the success rate in cases seeking to overturn trial decision of refusal of patent applications in Japan. Journal of the Intellectual Property Association of Japan, 12(1), 40-49.
https://www.ipaj.org/bulletin/pdfs/JIPAJ12-1PDF/12-1_p40-49.pdf

[4-1] Harris T. (2013). Default definition selection for credit scoring. Artificial Intelligence Research. September 2013;
http://dx.doi.org/10.5430/air.v2n4p49

[4-2] Khalilia M, Chakaborty S and Popescu M. (2011). Predicting disease risks from highly imbalanced data using random forest: July 2011; Article number 51
https://doi.org/10.1186/1472-6947-11-51

[4-3] Ali J, Khan R, Ahmad N, et al. (2012). Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues. September 2012; Vol 9, Issue 5, No3
https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees

[4-4] Jimenez A, Lazaro J and Dorronsoro J. (2007). Finding Optimal Model Parameters by Discrete Grid Search. Innovations in Hybrid Intelligent Systems. 2007; Pages 120-127
https://doi.org/10.1007/978-3-540-74972-1_17

[4-5] Yizhaki S. (1979). Relative Deprivation and the Gini Coefficient. The Quarterly Journal of Economics. May 1979; Vol 93, No2;

https://doi.org/10.2307/1883197

[4-6] Shelens J. (2014). Notes on Kullback-Leibler Divergence and Likelihood Theory. Cornell University. August 2014;
https://arxiv.org/abs/1404.2000

[4-7] S. S. Vallender (2006). Calculation of the Wasserstein Distance Between Probability Distributions on the Line. Theory of Probability & Its Applications. 2006; 18(4): 784–786
https://doi.org/10.1137/1118101

[5-1] Garg M., Kumar M. (2017). Identifying influential segments from word co-occurrence networks using AHP, Cognitve Systems Research, 47, 28-41, January 2018
https://doi.org/10.1016/j.cogsys.2017.07.003

[5-2] Angelo L., Stefan P., Fratocchi L., Marzola A . (2018). An AHP-based method for choosing the best 3D scanner for cultural heritage applications. Journal of Cultural Heritage, 34, 109-115, November–December 2018
https://doi.org/10.1016/j.culher.2018.03.026

[5-3] Mkolov T., Chen K., Corrado G., Dean J. (2018). Efficient Estimation of Word Representations in Vector Space. Computation and Language, 2013
https://arxiv.org/abs/1301.3781

[5-4] Goel A., Ganesh L., Kaur A. (2019). Sustainability integration in the management of construction projects: A morphological analysis of over two decades' research literature. Journal of Cleaner Production, 236, 117676, 1 November 2019
https://doi.org/10.1016/j.jclepro.2019.117676

[5-5] Lee H., Park G., Kim H. (2018). Effective integration of morphological analysis and named entity recognition based on a recurrent neural network. Pattern Recognition Letters, 112, 361-365, 1 September 2018
https://doi.org/10.1016/j.patrec.2018.08.015

[5-6] Mikolov T., Chen K., Corrado G., and Dean, J. (2013). Efficient estimation of word representations in vector space, arXiv preprint
https://arxiv.org/abs/1301.3781

[5-7] Church K. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162, January 2017
https://doi.org/10.1017/S1351324916000334

[5-8] Jianqiang L., Jing L., Xianghua F., Masud M., Zhexue H. (2016). Learning distributed word representation with multi-contextual mixed embedding. Knowledge-Based Systems, 106, 220-230, 15 August 2016
https://doi.org/10.1016/j.knosys.2016.05.045

[5-9] Carrasco R., Sicilia M. (2018). Unsupervised intrusion detection through skip-gram models of network behavior. Computers & Security, 78, 187-197, September 2018
https://doi.org/10.1016/j.cose.2018.07.003

[5-10] Fukui K., Miyazaki T., Ohira M. (2019). Suggesting Questions that Match Each User's Expertise in Community Question and Answering Services. 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),2019
https://doi.org/10.1109/SNPD.2019.8935747

[5-11] Jing X., Wang P., Rayz J. (2018). Discovering Attribute-Specific Features From Online Reviews: What Is the Gap Between Automated Tools and Human Cognition?. Software Science and Computational Intelligence, 2018
https://doi.org/10.4018/IJSSCI.2018040101

[5-12] Jan R., Khan A. (2020). Emotion Mining Using Semantic Similarity. Natural Language Processing, 2020
https://doi.org/10.4018/978-1-7998-0951-7.ch053

[5-13] Kim S., Park H., Lee J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. Expert Systems with Applications, 152, 113401, 15 August 2020
https://doi.org/10.1016/j.eswa.2020.113401

[5-14] Jatnika D., Biijaksana M., Suryani A (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. Procedia Computer Science, 157, 160-167, 2019

https://doi.org/10.1016/j.procs.2019.08.153

[5-15] Kai H., Qing L., Kunlun Qi., Siluo Y., Jin M., Xiaokang F., Jie Z., Huayi W., Ya G., Qibing Z. (2019). Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. Information Processing & Management, 56 (4), 1185-1203, July 2019
https://doi.org/10.1016/j.ipm.2019.02.014

[5-16] Wolf L., Hanani Y., Bar K, and Dershowitz N. (2014). Joint word2vec Networks for Bilingual Semantic Representations. IJCLA, 5 (1), 27–42, Jan-Jun 2014
https://www.gelbukh.com/ijcla/2014-1/IJCLA-2014-1-pp-027-042-Joint.pdf

[5-17] Lior R., Maimon O. (2005). Clustering methods - Data mining and knowledge discovery. handbook, (Springer US), 321−352, 2005.
https://doi.org/10.1007/0-387-25465-X_15

[5-18] Chakraborty S., Paul D., and Das S. (2020). Hierarchical clustering with optimal transport. Statistics & Probability Letters, 163, 108781, August 2020
https://doi.org/10.1016/j.spl.2020.108781

[5-19] Xu Q., Zhang Q., Liu J., and Luo B. (2020). Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Systems with Applications, 151, 113367, 1 August 2020
https://doi.org/10.1016/j.eswa.2020.113367

[5-20] Kim Hy., Kim Ha., and Cho S. (2020). Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. Expert Systems with Applications, 150, 113288, 15 July 2020
https://doi.org/10.1016/j.eswa.2020.113288

[5-21] Bai L., Liang J., and Cao F. (2020). A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. Information Fusion, 61, 36-47, September 2020
https://doi.org/10.1016/j.inffus.2020.03.009

謝辞

本研究に際し、長期間かつ多方面に渡ってご指導とご鞭撻を賜りました岩下基教授に心より感謝いたします。
また、本論文の審査にて様々な視点から数多くの貴重なご助言をいただきました井上明也教授、谷本茂明教授、東京理科大学黒沢健准教授に深く感謝いたします。
さらに、本研究の契機となる取り組みでご一緒させていただき、本論文に際してもご指導や審査において幅広いお力添えをいただいた田隈広紀准教授に改めて感謝申し上げたく存じます。
最後に多忙な中で体調を気遣ってくれた妻、里美にも感謝の気持ちを伝えたいと思います。

本研究を糧として、日本における製品開発の改善と向上に寄与できるように今後も研鑽を積んでいきたく存じます。

多大なるお力添えをいただきました皆様への感謝と御礼を申し上げたく、謝辞に代えさせていただきます。