Prediction of Business Partners Using an n-Gram-Based Approach that Combines a Network Model and Linear Model of a Supply Chain

Hajime Sasaki[1], Ichiro Sakata[1,2]
[1] Policy Alternatives Research Institute, The University of Tokyo, Tokyo, Japan
[2] Dept. of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

***Abstract*—Supply chains are viewed as networks of both goods and services, and knowledge and information. Their knowledge will be potential resources for new business relationship with hidden partners; however, many companies find it difficult to develop new opportunities. A recommendation of a potential partner is helpful for regional revitalization. Research into supply chains has shifted from a linear model to a network model. A network model using graph theory can topologically explain a supply chain, whereas, in a linear model, there are some differences with the real world because of a lack of information. In this study, we propose a prediction model for a new business partnership with predictors extracted from network and linear approaches used in combination for the prediction of performance and interpretability. Our dataset consisted of a network of 327,012 transactions among 131,192 companies in Northeast Japan, which was retrieved from supplier-customer relationship data provided by Teikoku Databank, Ltd. Network centralities were extracted as topological features from the network of each company. Trigram relationships were also extracted from the network motif so that the logistic flow to a company from its supplier could be used to predict customers as business partners. The results showed that the performance of the proposed model was excellent, with a high contribution of probability extracted from the trigram relationships. From this perspective, we found that the information of logistics flows is a critical factor for predicting a potential partner, even in a network model.**

I. Introduction

A. Interfirm relationship mechanisms

In corporate activities, making decisions on which company should be a customer and from which company to procure materials is an important element for directing a business strategy. The development of new customers is always an intention for all companies. Many companies believe that the information they are actively collecting is for marketing development. However, there is a possibility that there may be potential trading partners that can enhance the value of their company outside the scope of their current sales activities. Along with a diversification of consumer needs, there are many cases where new value can be created by developing into different industries. Identifying such potential business relationships is limited because only closed information is available within the company's traditional industry.

If a company identifies a better partner than its current partner, the decision to switch the transaction immediately is correct from a rational economic perspective. However, in practice, companies do not frequently switch partners. One of the reasons that it is not a reasonable transaction is because of the reciprocal cooperative relationship between suppliers and customers, that is, a concept called "keiretsu" [1], which is defined as the structure of interfirm transactions in which a large number of fixed suppliers continuously supply intermediate goods [2].

In Japan, in particular, the influence of the power balance on the structure of interfirm transactions governed by the concept of keiretsu is strong. It is said to be a source of strength compared with the type of open system trading structure in Europe and the United States [3] [4]. Several studies have been conducted in interfirm relational theory on keiretsu. In particular, discussion from the transaction cost perspective, resource dependence perspective, and institutional perspective is active [1][5][6][7][8][9].

The transaction cost perspective considers the formation mechanism of interorganizational relationships from a cost perspective. From this perspective, when the total sum of the costs related to the transaction is low, this means that a business relationship is formed for the purpose of resource procurement [1]. Conversely, when such a transaction cost is high, the company intends to make resources in-house; that is, the control of transaction costs determines a boundary of the organization [1][5][6].

The resource dependence perspective is a theory that considers the formation mechanism of interorganizational relationships as the power balance of resources [7][10]. Organizations need to acquire the resources necessary for their survival from external organizations; however, if they depend on these organizations too much, they lose their management initiative [1]. A difference in production experience between manufacturers and suppliers, and relative differences in skills often define such a power balance [11].

The institutional perspective is a theory that considers the formation mechanism of interorganizational relationships from the viewpoint of institutional homogenization and ensuring legitimacy for survival [9]. Because decision-making for corporate transactions is balanced by such complex interfirm relationships, it can be understood that simple economic rationality only cannot explain decision-making.

Additionally, actual interfirm relationships are not always reasonable because of a bias in human decision-making. For example, a company seeks a partner that has the processing

978-1-890843-36-6 ©2017 PICMET

technology necessary to manufacture the company's products. If the company does not know that there is a partner with the required technology in the vicinity because of the availability heuristic, the company has to engage with a well-known distant partner. This event occurs without the company paying an information cost, which makes the transaction cost small in the short term; however, there are cases in which the company has to pay a large cost in the long term, overall. In particular, if such short-distance exchanges that should exist are neglected, this can also result in a lost opportunity for the economy of the entire region [12] [13].

B. Connecting potential relationships

Companies must always collect information to prepare for market expansion and future transaction discontinuities. At the present time, products that consist of a single technology are rare, and many companies are required to innovate through collaborations with other industries [14]. Many companies at the present time have to overcome the challenge of open innovation; however, they have been dependent solely on their own industries to date, and they struggle to obtain information on other industries. Additionally, local governments are obliged to provide appropriate information to companies that belong to economic zones that they own. However, in the recent explosive increase in the amount of information, transaction costs are increasing because companies are overwhelmed by the vast amount of information.

By contrast, information processing is becoming more feasible at a progressively lower cost by an increase in the performance of computers, disclosure of data, and improved analytical techniques. Companies can make more rational decisions if they can contribute to reducing transaction costs through information technology. In particular, it is desirable in many respects to make it possible to identify potential partners with which companies do not currently have business relationships, but with which they can have business relationships in the future. This not only leads to new sources of information on sales expansion for companies, but also a risk hedge when current business relationships are lost in the future.

Based on this background, there are many studies that focus on business relationships in terms of connections in the network. Although, large companies and good-performing companies are important, not all companies are large scale; thus, increasing companies' ability to connect is important in the concerned area and worthy of attention; not only sales, but also companies' ability to connect is an important aspect to consider. We can observe this phenomenon, which has been shown in many regions, and recent attempts have been made to raise the emergence of innovation in these regions by positively searching for such enterprises as policy inputs and identifying potential transactions.

A method that applies statistical machine learning to the supply chain has been proposed [15][16][17] to predict and recommend such potential transactions using the informatics method. Such a proposed method for connecting new potential partners not only provides information that contributes to the expansion of sales channels for companies but also makes it easier for local governments to manage economic zones, and enables appropriate information provision services. In that respect, the recommendation of potential partners is helpful for regional revitalization.

C. Linear Supply Chain and Network Supply Chain

Research on supply chains has traditionally conducted analysis using linear structure models [18]. A linear structure supply chain model is characterized by an expression of the flow of material among companies in one direction. However, such a linear model extremely simplifies the actual state of the supply chain, and makes it difficult to fully express the complicated relationship between the flow of goods and information [19]. Under such circumstances, analytical methods to understand the flow of goods and information between companies as a network are being used to more appropriately describe the complexity of supply chains. In particular, social network analysis (SNA) is being used in such cases. SNA, which originated from the graph field in mathematics, is suitable for modeling complex societies, and it can express the actual situation more adequately than supply chain analysis using a linear model [20]. In the analysis using SNA, we refer to a company as a node, and a system with relationships between companies (i.e., business relationships) as edges is called a network. In particular, this method can quantify the topological network position of each company by calculating network centrality. By considering network centrality, it is possible to obtain an important viewpoint that small and medium-sized enterprises play an important role in regional networks [21][22]. Network centrality in intercompany transactions has been found to have a particular relationship with respect to the growth of corporate activities and the emergence of innovation [23][24][25][26].

Understanding the relationship between business transactions as a network is also meaningful from the viewpoint of the organizational mechanism of organizational relationships. For example, degree centrality and betweenness centrality represent resource dependence and power, and betweenness centrality and structural pores are negatively correlated with autonomy. By calculating such values, it is possible to quantitatively evaluate transaction relationships from multiple viewpoints [1].

Against such a background, there is research that predicts potential business transactions from the viewpoint of the network. A previous study proposed a method for recommending potential transactions with similar companies by crawling corporate official website information from the Web and calculating similarities, such as technical terms that characterize the enterprise. Additionally, using network centralities as features by focusing on the structure of regional enterprise transactions, a method for recommending potential transactions was proposed [17]. However, none of the previous research has fully considered the flow of goods and information, which is an essential element of the supply chain; thus, this practical improvement is required according to actual situations.

D. Purpose of this reserch

We propose a potential transaction recommendation method that combines the advantages of both the network model and linear model. We expect the method to not only

achieve high predictive performance but also interpretability and understandability. An indispensable element is that this is a model that makes it easy to interpret why potential business partners are recommended. Using the method of recommending similar companies by calculating the similarity of business types, there is a possibility that competitors will be recommended. By contrast, it may be difficult to understand why the recommended company became a candidate in the model with network centricity centrality as the feature quantity.

We consider information on the flow of goods and services from upstream companies to downstream enterprises as features. We expect that the features will strongly contribute to potential transaction prediction. Information on the flow of goods is essential in the linear model supply chain. The original aspect of our research is that we add logistics information from the linear model to the network model, which has been a mainstream approach in recent years, so that both accuracy and interpretability are achievable.

II. Main Idea

The main idea of our research is to use information on the flow of goods, which has not been used sufficiently to date as a feature quantity (explanatory variable) of potential business relation prediction.

Specifically, we use the information on the business types of three consecutive companies in a supply chain. When given the business type information of the two upstream companies, we calculate the conditional probability of the business type of the third company. For example, suppose that company A, company B, and company C have the transaction areas shown in the supply chains in Fig. 1. The business types of these companies are α, β, and γ, respectively.

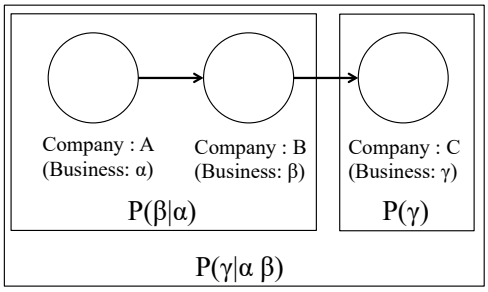

Fig.1 Conditional probabilities of business types

As shown in Fig. 1, company A is the upstream company in this supply chain, company B is midstream, and company C is downstream. In this case, at this time, the conditional probability that the company with business type β will have P(γ| α β) of having a business relationship with the company with business type γ as a customer of company B with business type α is calculated as.

$$P(\gamma|\alpha\beta) = \frac{P(\alpha\beta|\gamma)P(\gamma)}{P(\alpha\beta)} \quad (1)$$

Similarly, the conditional probability P(β|α) that the company with business type β can become the customer of the company with business type α is given by

$$P(\beta|\alpha) = \frac{P(\alpha|\beta)P(\beta)}{P(\alpha)} \quad (2)$$

We apply a model called n-gram, which is one of basic stochastic language models used in the field of natural language processing, to the supply chain to achieve a probabilistic representation of the flow of services and goods. The n-gram language model presupposes that the occurrence probability of each word depends on the immediately preceding (n-1) words. For example, when the word sequence $w_n = w_1\ w_2 \cdots \cdots w_n$ is given, the occurrence probability $P(w_n)$ can be calculated by the following:

$$P(w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2) \ldots P(W_n|w_1 \ldots w_{n-1})$$
$$= \prod_{i=1}^{n} P(w_i|w_1 \ldots w_{i-1}) \quad (3)$$

For example, when the context of "This is a cat" is given the occurrence probability P (This is a cat), it can be calculated by

$$P(This\ is\ a\ cat) = P(This) * P(is|This) * P(a|This\ is) * P(cat|This\ is\ a) \quad (4)$$

In this case, this is called the 4-gram language model. In our research, we calculate the conditional probability of three consecutive businesses types, that is, a 3-gram, and make it a feature of each item of transaction information; that is, the following equation is calculated in Fig. 1 and used as the feature quantity of the prediction model:

$$P(\alpha\beta\gamma) = P(\alpha)P(\beta|\alpha)P(\gamma|\alpha\beta) \quad (5)$$

The higher the probability, the more logistics information for the combination of businesses can be obtained. For example, it can be assumed that P("Metal processing industry", "Assembling precision equipment", "Manufacturing automobile parts") is relatively high. By contrast, it can be assumed that P ("Publishing", "Fishery", "Electric industry") is low. The flow of services and goods in the latter case is difficult to occur.

III. Methods

A. Dateset

We use a dataset that records actual intercompany transactions in the Tohoku region in Japan. The Tohoku region consists of six prefectures: Akita Prefecture, Aomori Prefecture, Iwate Prefecture, Yamagata Prefecture, Miyagi Prefecture, and Fukushima Prefecture. In particular, companies in Iwate Prefecture, Miyagi Prefecture, and Fukushima Prefecture suffered direct damage caused by the Great East Japan earthquake and tsunami that occurred on March 11, 2011, which also affected neighboring and other areas, including Tokyo. The event triggered a discussion on the necessity of a business continuity plan.

We use corporate transaction data provided by Teikoku Databank Co., Ltd. as a data source. In this data source, each item of company attribute information shown in Table 1 is recorded with the relationship representing the flow of goods

from the supplier company to the customer company as one record. Each company is anonymized and tagged by a unique identifier (ID).

A potential supplier's forecasting model is defined as a supervised model that learns that a transaction occurred in year T', although there was no transaction in year T for any intercompany relationship. Specifically, in our research, among the relationships among companies that existed in year T, the relationship in which business relationships are built for the first time in year T+1 taken as a positive example. In this data, the ratio of positive examples and negative examples is biased. Because the number of actual transactions occurring is only a portion of all combinations of all companies in this dataset, we balance the data by constructing randomly extracted negative example data of the same size as the positive sample.

B. *Feature setup*

A predictive model of a potential counterparty is defined as a two-class classification that specifies what type of feature quantity is effective in the data of a pair for actual transactions. Each business relationship consists of a pair that defines an upstream company as supplier and a downstream company as a customer. For the obtained network data, node attributes and edge attributes are calculated and used as feature quantities. Network centrality presented in Table 1:

TABLE 1 NETWORK INDICES.

Network indices.	Description
Degree centrality [in/out]	The number of its adjacent edges. It is recognized the company has the potential to affect others through operational decisions and strategic behavior [19]
Closeness centrality [27][in/out]	How many steps is required to access every other vertex from a given vertex. [28]
Betweenness centrality	How often a node appears on the shortest paths between two other noted in a network.
Eigenvector centrality [29]	Eigenvector centrality is different from degree and closeness centrality by emphasizing the importance of others that are connected to a focal firm [30]
PageRank	One of a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them [31]. This index identifies a company with transaction by companies that are themselves frequently transacted.
Bonacich power centrality [31]	Corresponds to the notion that the power of a vertex is recursively defined by the sum of the power of its alters.
Within-module degree[32]	How well-connected node i is to other nodes in the cluster
Participation coefficient[32]	How well-distributed the links of node i among different clusters

Next, in the obtained enterprise transaction network, 3-gram information is extracted for each main business. Because the input data in natural language is a sentence, it is technically easy to extract three consecutive words., because corporate transaction data stores the presence or absence of a business relationship between two specific companies, it is difficult to obtain three or more consecutive nodes directly. Therefore, by constructing transaction data once as a network, a specific subnetwork is extracted from the network. A network motif is a subnetwork that satisfies a certain condition. For example, the 3-node network motif is given in Fig. 2, where ID: 1 represents three node relationships without any links among them and ID: 4 represents only one node that has two transaction links with others. We extract the network motif ID: 6 in Fig. 2, where ID: 6 indicates that one node has a direct link to the other node. The node also has a direct link to another node. It represents linear material flow in the supply chain among the nodes. By extracting all motifs from the dataset, we obtain the main business information of three consecutive companies. We calculate the conditional probability from the obtained main business 3-gram information and incorporate it into the feature amount as an edge attribute.

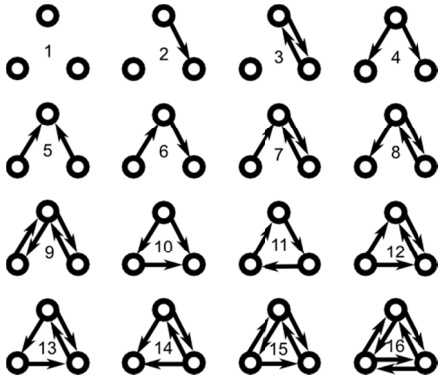

Fig2. Network motifs [33]

C. *Modeling and evaluation*

We propose five models according to combinations of feature quantities described in Table 2. Model A uses the 3-gram probability as the main feature. In addition to Model A, Model B adds network centricity to the feature quantities. Model C is given company attribute information in addition to Model B. We build these three models and compare their predicted performance.

TABLE 2 TYPE OF MODEL

Features	Cooperate attributes	Network centralities	3-gram probabilities (Proposed)
Model_A	*		
Model_B		*	
Model_C			*
Model_D	*	*	
Model_E	*	*	*

We use a random forest as a classifier. Random forest is an ensemble machine-learning algorithm that improves generalization ability by integrating multiple weak classifiers by decision trees. Random forest has simple learning methods, but can perform better identification and prediction than general decision trees. It extended to multi-class problem with different approach than CHAID. Furthermore, since it can

analyze nonlinear relationships, there is room to exceed the limit of linear regression. Random forest has the characteristic that over-fitting does not occur even if the size of the forest is enlarged.

To evaluate the model, precision, recall, and the F-measure are used. Precision represents the number of true positives over the total number of true positives and number of false positives. It defines the percentage of all retrieved data that are relevant. Recall represents the number of true positives over the total number of true positives and number of false negatives. It defines the percentage of relevant classifications that are successfully predicted by the model relative to all relevant data that exist. Generally, precision and recall are in a trade-off relationship. The F-measure is obtained by calculating the harmonic mean of precision and recall.

For evaluation of the model, Precision, Recall and F-measure are used. Precision represents the number of true positive over the number of true positive and the number of false positive. It means percent of all retrieved data that are relevant. Recall represents the number of true positive over the number of true positives and the number of false negatives. It means percentage of relevant classifications successfully predicted by the modem, relative to all relevant data that exist. In general, precision and recall are in a trade-off relationship. The F-measure is obtained by harmonic mean of Precision and Recall.

IV. Result

As an original dataset, we extracted 327,012 transactions and 131,192 companies that have headquarters located in any of the six prefectures in the Tohoku region and all companies across the country that had transactions with these companies as of 2014. A summary of the number of transactions and companies are shown in Table 2.

A summary of company attributes included in this data set is shown in Table 3. Also, those included in the top 10 in the business types are shown in Table 4.

TABLE 2. DATA SET IN EACH YEAR

	2010	2011	2012	2013	2014
Number of transactions	291,808	312,864	313,762	321,780	327,012
Number of companies	128,890	132,616	132,075	133,926	135,257

TABLE 3 SUMMARY OF COMPANY ATTRIBUTES IN 2014 DATASET.

Attributes	Min	Median	Max
Sales revenue [JPY]	0	141	130.1M
Capital [JPY]	-289.4M	0	141.6M
Employee [person]	0	7	194,688
TDB-Score	0	46	93
Number of Office	0	2	24,204
Number of factory	0	1	117

TABLE 4. TOP10 BUSINESS TYPE IN THE DATASET IN 2014 DATASET.

Business types	Freq.
Wooden building Contractors	5,569
Civil Contractors	4,284
General cargo automobile transportation	3,278
Automobile general maintenance	2,469
Electrical wiring Contractors	1,985
General tube Contractors	1,946
Civil engineering and construction services	1,804
Interior Contractors	1,722
Earthwork, concrete construction	1,587
Water supply and drainage and sanitation Contractors	1,550

In this dataset, the number of relationships in which the transaction occurred for the first time in 2011 despite no transactions in 2010 was 53,622 as shown Table 5. As a result of randomly generating the same number of negotiated intercompany relationships with this transaction as a positive example, the analysis dataset was 107,243 transactions (=53,622+53,621) for dataset (2010-2011).

TABLE 5. NUMBER OF POSITIVE EXAMPLE AND NEGATIVE EXAMPLE

Year T Year T'	2010 2011	2011 2012	2012 2013	2013 2014
Positive example	53,622	40,537	41,225	41,120
Negative example (Balanced)	53,621	40,536	41,222	41,117

We constructed a network from the transaction dataset, extracted the corporate network motif, and further transformed it into a business type network motif. The 3-gram probability shown in the formula was calculated. In table 6, 3-grams of the top 10 combinations and their probabilities are presented in descending order of frequency.

TABLE 6 TOP-10, FREQUENT TRI-GRAM IN A DATA SET (2013-2014)

Business_α	Business_β	Business_γ	Freq.	P(αβγ)
Fresh seafood wholesale	Fresh seafood wholesale	Fresh seafood wholesale	44,457	0.021
Other fishery food products manufacturing	Fresh seafood wholesale	Fresh seafood wholesale	35,629	0.017
Sake production	Liquor wholesale	Liquor retail	13,679	0.007
Automobile general maintenance	Automotive Wholesale	General cargo automobile transportation	13,109	0.006
Fresh seafood wholesale	Fresh seafood wholesale	Other fishery food products manufacturing	10,248	0.005
Fresh seafood wholesale	Fresh seafood wholesale	Dry matter Wholesale	9,582	0.005
Automobile body maintenance	Automotive Wholesale	General cargo automobile transportation	8,650	0.004
Pharmaceutical formulations manufacturing	Pharmaceutical wholesalers	No floor clinic	8,341	0.004
Civil Contractors	State institutions	Other social insurance and welfare	900	0.04
General civil engineering and construction Contractors	State institutions	Peer group	874	0.10

For the obtained dataset, three types of models that corresponded to the table were constructed and the results of evaluating the model are presented in Table 7(Precision). Table 8(Recall) and Table 9(F-measure)

TABLE 7 PRECISION OF EACH MODEL

	2010-2011	2011-2012	2012-2013	2013-2014
Model_A	0.73	0.72	0.71	0.73
Model_B	0.73	0.73	0.74	0.72
Model_C (Proposed)	**0.94**	**0.93**	**0.93**	**0.93**
Model_D	0.78	0.78	0.77	0.77
Model_E (Proposed)	0.82	0.82	0.82	0.81

TABLE 8 RECALL OF EACH MODEL

	2010-2011	2011-2012	2012-2013	2013-2014
Model_A	0.66	0.66	0.67	0.64
Model_B	0.75	0.76	0.74	0.76
Model_C (Proposed)	0.36	0.38	0.39	0.39
Model_D	**0.77**	**0.78**	**0.78**	**0.78**
Model_E (Proposed)	**0.77**	**0.78**	**0.78**	**0.78**

TABLE 9 F-MEASURE OF EACH MODEL

	2010-2011	2011-2012	2012-2013	2013-2014
Model_A	0.69	0.69	0.69	0.68
Model_B	0.74	0.74	0.74	0.74
Model_C (Proposed)	0.52	0.54	0.55	0.55
Model_D	0.78	0.78	0.77	0.78
Model_E (Proposed)	**0.79**	**0.80**	**0.80**	**0.80**

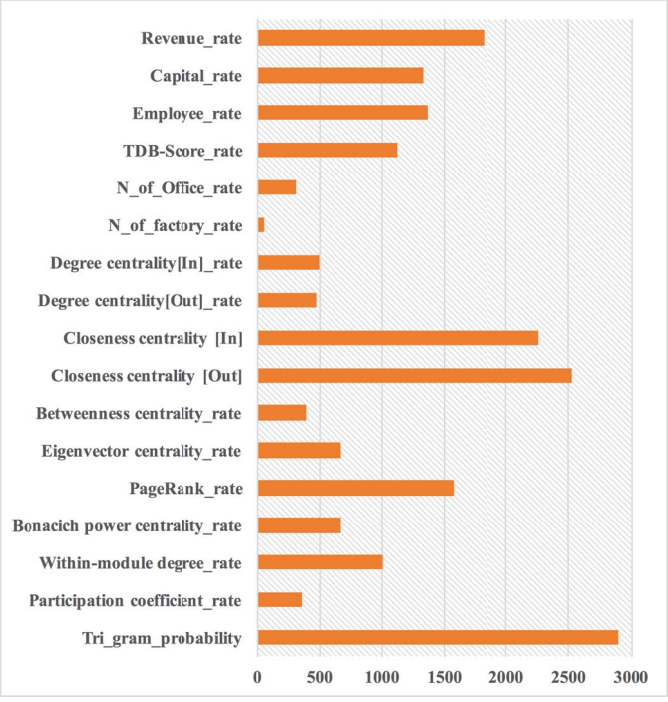

Figure3 Feature importance of Model_E in 2013 model

V. DISCUSSION

In Table 2, we can see both the number of companies and transactions tend to increase slightly. However, from Table5, it can be seen that the number of positive examples after 2011-2012 dataset is decreasing. Positive example is the number of new transactions that occurred for the first time in year T+1 among the companies that exist in year T. This can be thought of as the impact of the earthquake that occurred in 2011. As shown in Table 2, it is a phenomenon that could not be seen with the number of companies and transactions alone.

Table 4 shows that the most frequent business type in the dataset was the construction industry, for example, "Wooden building contractors" and "Civil contractors". According to the Cabinet Office, the construction industry is listed above the bankruptcy rate because of the Great East Japan earthquake in 2011. One of the reasons is that purchasing materials has become difficult [34]. Because of the earthquake, the connection with suppliers and customers that had trading relationships was broken and it was difficult for companies to find new business partners, which indicates that many companies went bankrupt. Construction demand for the Tokyo Olympic Games is increasing, in addition to earthquake reconstruction projects. Identifying potential transactions is meaningful for the entire Tohoku region economy.

The most consecutive industries in Table 6 were combinations of Fresh seafood wholesale. We can see local governments ("State institutions") were included as intermediate companies. The fact that prefectures occupied the top position for the combination of business transactions means that prefectures are key players in the entire Tohoku region. Although it is necessary to compare other areas, such characteristics cannot be seen in the metropolitan economic area. The fact that local governments are key players in business cannot deny the possibility that the awareness of self-sustainable new development is lower in corporate activities than in other regions.

The Model_C and Model_E in Table 7,8 and 9 contain the 3-gram probability as a feature quantity, which is the original aspect of this research. In addition to the performance of these models, we also discuss prediction performance by comparing with existing features in Model_A, Model_B and Model_D [17].

The precision of Model A, which considered corporate attribute information, was almost the same as that of Model B, and it is difficult to determine whether the corporate attribute information prediction made a large contribution to performance improvement in this analysis. We should pay attention to the fact that corporate attribute information is not directly related to transaction information. Even if we obtain the result that a specific feature number contributes to potential transactions, it is not necessarily the case that the appropriate and realistic trading partner is recommended for the company. Figure3 shows the importance of Model_E that includes corporate attributes, network centrality and tri-gram probability as features. When we focus on only corporate

attribute, we can see from this result the model has an internal structure that the fact that the rate of sales amount is high contribution for prediction. Such a prediction model that focuses on corporate attributes may be a model that promotes "winner takes it all." What is important in potential transaction model is not comprehensive corporate attribute values, but compatibility with the companies.

The precision of Model_C exceeded 0.93. High precision means that trade is actually occurring at a high rate among predicted results that future transactions will occur.

From the recall point of view, we observe that Model_C was lower compared with others from Table 8. Now, users who benefit from the proposed method will be interested in how much they trust the recommended information. By contrast, they are unlikely to be interested in how many of the correct new transaction are covered; that is, in a system that aims to recommend potential partners, a model with higher precision is more practically useful than a model with high recall. From such a viewpoint, it can be said that all three proposed models had higher performance than existing methods.

We see that Model_E and Model_D have the same performance. From this, it was confirmed that even if trigram probability were included in Model_D, it would not adversely affect Recall.

Given the overall performance of the model based on the F-measure, Model_E, which used corporate attribute, the network centricity and 3-gram probability features, was excellent. This model achieved higher performance than the baseline existing methods. From these facts, we could achieve the same performance using the 3-gram probability as the feature quantity. 3-gram probability is noteworthy in terms of data availability. Companies should be aware of the information on which companies they use as suppliers. In extreme terms, it means that constant prediction is possible even with only the 3-gram probability.

In this research, we could propose a model with both prediction performance and interpretability by having tri-gram probability that is more important feature than company attributes and network centralities.

VI. Conclusion

Activities to seek potential transactions are indispensable, not only in corporate activities but also for local governments. This can also be explained from the viewpoint of organizational relationship formation, including the transaction cost perspective that actual business relationships are not necessarily caused by reasonable results.

The mechanism to identify and recommend potential transactions is gaining attention as a means of entering the market, developing new innovation opportunities, hedging business continuity risk, and introducing policies that contribute to regional economic development for local governments.

In our research, we applied a machine learning method to reduce the information processing cost, and have shown that it is possible to predict a rational and novel potential supplier.

In supply chain analysis in recent years, the network viewpoint is the mainstream idea, but in our research, we aimed not only at the network but also a hybrid model that incorporates the good aspects of the traditional linear model. Thus, not only can we consider the structure of complicated transaction relationships but also obtain prediction results that are compatible with the ease of interpretation necessary for human decision-making.

Specifically, the 3-gram probability based on the conditional probability of three consecutive business types was taken as the feature quantity. The quantification of the event in which the flow of material occurs conditionally stochastically was inspired by a stochastic language model traditionally used in the field of natural language processing. The trigram probability can be interpreted as a feature that returns a low number in a combination of material flows that rarely occur. In the network model, the flow of goods has not yet been discussed sufficiently. By considering the intrinsic information of business transactions, such as the flow of goods, as a conditional probability, it is possible to achieve realistically possible recommendations for business transactions with higher accuracy. All the methods proposed in our research exceeded the baseline from the viewpoint of precision, and we confirmed that this is a useful model. In the future, we would like to further consider models that incorporate aspects of compatibility between companies in feature quantities.

The Cabinet Office in Japan began operating a system called the Regional Economy Society Analyzing System (RESAS) in 2015 [35]. This system is aimed at local governments to enable them to understand the current situation and problems in the area based on objective data to achieve local creation. The corporate transaction information provided by Teikoku Databank used in our research is also stored in the system. The API of the system was publicized and a Hack-a-thon event was held to form a public–private partnership that involved the private sector. In December 2016, the Basic Law on the Promotion of Use of Public–Private Data was enacted, and it is thought that further openness of data will proceed and public–private partnerships will explore how to use new information technology [36]. Thus, as machine learning and the use of big data are expected to increase as information that contributes to corporate management, using the method proposed in our research will provide insight becomes a trigger to be associated with a potential trading partner. We hope to lead the creation of innovation.

Acknowledgment

The dataset in this research is provided by Teikoku Databank, Ltd.

REFERENCES

[1] Akiyama, T. (2014) A Study of the Formation Mechanism of Inter-organizational Relationship by the Network Analysis.

[2] Hsu, R. C. (1999). The MIT encyclopedia of the Japanese economy. Mit Press.

[3] Cusumano, M. A., & Takeishi, A. (1991). Supplier relations and management: a survey of Japanese, Japanese-transplant, and US auto plants. *Strategic Management Journal*, *12*(8), 563-588.

[4] Dyer, J. H. (1994). Dedicated assets: Japan's manufacturing edge. *Harvard Business Review*, *72*(6), 174-178.

[5] Coase, R. H. (1937). The nature of the firm. *economica*, *4*(16), 386-405.

[6] Williamson, O. E. (1975). Markets and hierarchies: antitrust analysis and implications. *New York: The Free Pres*.

[7] Thompson, J. D. (1967). Organizations in action: Social science bases of administrative theory. Transaction publishers.

[8] Pfeffer, J., & Salancik, G. R. (2003). *The external control of organizations: A resource dependence perspective*. Stanford University Press.

[9] DiMaggio, P.J. & W.W. Powell.(1983). The Iron Cage Revisited. *American Sociological Review*, vol. 48, no, 2, pp. 147- 160

[10] Pfeffer, J., & Salancik, G. R. (1978). The external control of organisations. *New York*, *175*.

[11] Argyres, N. (1996). Evidence on the role of firm capabilities in vertical integration decisions. *Strategic Management Journal*, *17*(2), 129-150.

[12] Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360-1380.

[13] Sakata, I., Kajikawa, Y., Takeda, Y., Hashimoto, M., Shibata, N., & Matsushima, K. (2007). *Network dynamics in the twelve regional clusters*. RIETI discussion paper series 07-J-023.

[14] Kodama, Technology fusion and the new research-and-development Harv. Bus. Rev., 70 (4) (1992), pp. 70–78

[15] Guo, X., Yuan, Z., & Tian, B. (2009). Supplier selection based on hierarchical potential support vector machine. *Expert Systems with Applications*, *36*(3), 6978-6985.

[16] Guosheng, H., & Guohong, Z. (2008). Comparison on neural networks and support vector machines in suppliers' selection. *Journal of Systems Engineering and Electronics*, *19*(2), 316-320.

[17] Zuo, Y., Kajikawa, Y., & Mori, J. (2016). Extraction of business relationships in supply networks using statistical learning theory. *Heliyon*, *2*(6), e00123.

[18] Handfield, R. B., & Nichols, E. L. (1999). *Introduction to supply chain management*. prentice Hall.

[19] Kim, Y., Choi, T. Y., Yan, T., & Dooley, K. (2011). Structural investigation of supply networks: A social network analysis approach. *Journal of Operations Management*, *29*(3), 194-211.

[20] Autry, C. W., & Griffis, S. E. (2008). Supply chain capital: the impact of structural and relational linkages on firm execution and innovation. *Journal of Business Logistics*, *29*(1), 157-173.

[21] Loveman, G., & Sengenberger, W. (1991). The re-emergence of small-scale production: an international comparison. *Small business economics*, *3*(1), 1-37.

[22] Hoang, H., & Antoncic, B. (2003). Network-based research in entrepreneurship: A critical review. *Journal of business venturing*, *18*(2), 165-187.

[23] PowellWW, Koput KW, Smith-Doerr L. 1996. Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology. Adm. Sci. Q. 41:116–45

[24] Powell WW, White DR, Koput KW, Owen-Smith J. 2005. Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. Am. J. Sociol. 110:1132–205

[25] Tsai, W. 2001. "Knowledge Transfer in Intra-Organizational Networks." *Academy of Management Review* 44: 996–1004. doi: 10.2307/3069443

[26] Giuliani, E., and M. Bell. 2005. "The Micro-Determinants of Meso-Level Learning and Innovation: Evidence from a Chilean Wine Cluster." *Research Policy* 34 (1): 47–68. doi: 10.1016/j.respol.2004.10.008

[27] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581-603.

[28] Freeman, L.C. (1979). Centrality in Social Networks I: Conceptual Clarification. *Social Networks*, 1, 215-239.

[29] Bonacich, P., 2007. Some unique properties of eigenvector centrality. Soc. Networks 29 (4), 555–564.

[30] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

[31] Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 1170-1182.

[32] Guimera, R. and Amaral, L.A.N. (2005) Cartography of complex networks: modules and universal roles, Journal of Statistical Mechanics: Theory and Experiment, 02, P02001.

[33] "The 16 possible network motifs involving three nodes." From *Math Insight*. Retrieved from http://mathinsight.org/image/three_node_motifs on 8/Jan./2017.

[34] Cabinet office, Disaster Management. Retrieved from http://www.bousai.go.jp/kaigirep/gekijin/dai2kai/pdf/shiryo1-1.pdf on 8/Jan./2017.

[35] Ministry of Economy, Trade and Industry. Retrieved from http://www.meti.go.jp/press/2015/04/20150421001/20150421001.html on 8th/Jan./2017

[36] The house of representatives, Japan. Retrieved from http://www.shugiin.go.jp/internet/itdb_gian.nsf/html/gian/honbun/houan/g19201008.htm on 8th/Jan./2017.