Hindawi
Advances in Human-Computer Interaction
Volume 2018, Article ID 9054685, 12 pages
https://doi.org/10.1155/2018/9054685

Research Article

How to Understand Belief Drift? Externalization of Variables Considering Different Background Knowledge

Teruaki Hayashi and Yukio Ohsawa

Department of Systems Innovation, School of Engineering, Tokyo, Japan

Correspondence should be addressed to Teruaki Hayashi; teru-h.884@nifty.com

Received 28 June 2018; Revised 1 November 2018; Accepted 21 November 2018; Published 4 December 2018

Guest Editor: Rafal Rzepka

It is necessary to make decisions by integrating appropriate information that is not used in daily life in disaster prevention before, during, and after disasters. Despite this, it is difficult for people to make use of appropriate information under circumstances where various kinds of information are complicated. People can be in an agitated state in which they do not know what will happen. In this paper, we define this situation as Belief Drift (BD) and discuss what kinds of data should be acquired to understand situations of BD because factors causing BD may be diverse. We collected explanations of BD from researchers with different background knowledge and discussed sets of variables inferred by VARIABLE QUEST (VQ). VQ is the inferring method for variables unifying cooccurrence graphs of variables in the datasets. The results indicate that common variables are externalized from the different explanations of BD by researchers with different background knowledge. Results suggest that, even if the terms used to explain the state of BD differ, the data acquired to understand BD are common.

1. Introduction

Projects seeking to protect society from disasters are progressing globally. The Department of Homeland Security, the Department of Energy, and so on in the United States have conducted infrastructure protection projects since 2003 [1, 2]. Concerning predicted disasters, technologies to evaluate risks have been developed systematically, including vulnerabilities and interdependencies of 12 essential infrastructures. Conversely, the Japanese government has tackled problems of information infrastructure. Resilient systems considering a variety of cellular phones and smartphones have been studied academically and practically from the viewpoints of the possibility of message transmission, stability, and the reliability of information [3, 4]. Especially for disaster prevention ahead of the Tokyo 2020 Olympic and Paralympic Games, the government has provided the Disaster Prevention Portal for foreigners [5].

However, while the robustness and resilience of the infrastructure have increased, victims of natural disasters have not been able to reach favorable decisions because of a state of desperation from anxiety. The kinds of information that have been delivered to those who were anxious during and after disasters constitute an urgent issue. For example, in the examinations of internal radiation exposure after the accident of the First Nuclear Power Plant in the Fukushima Prefecture, the affected area was divided into residents that were hardly affected by radiation exposure and those that were noticeably affected [6, 7]. In the medical examinations after the accident, rapid chronic diseases were evident [8]. Nara mentioned that, although recognition and anxiety exist, the information transmitted by risk management institutions cannot gain the confidence of Japanese residents [9]. Information and the media are diverse, which makes it difficult to obtain consistent messages [10]. Even in the Chernobyl nuclear power plant accident, the relationship between the accident and carcinogenesis is scientifically uncertain [11].

In disaster prevention before, during, and after disasters, it is crucial to create a methodology that provides appropriate integrated information used in daily life. In such situations, people cannot establish certain beliefs because of inconsistent and diverse information. For example, when inconsistent information is given such as "that town is polluted and cannot live," or "this city is safe" from disasters, it evokes

anxiety and anger in information senders and surrounding people and causes more meaningless messages to proliferate [12]. Research mentions that anxiety amplifies continuously and has taken root among residents. Depending on conditions, differing information may spread as inconsistent information, which confuses people because they cannot be confident as to what to believe. We defined this situation as Belief Drift (BD) and began this project to study this [13]. The essential mission of our project is to establish the fundamental methodology and systems of information generation, propagation, acceptance with reliability, usefulness, and consistency for people who are anxious during and after disasters.

To detect BD and create a system providing appropriate information for people whose beliefs are drifting, we acquired data to understand situations of BD. However, there is a problem of a vocabulary gap among different researchers. When disciplinary fields differ, there are many different factors to consider in actions. Based on these challenges, because BD is a multidisciplinary question, it is difficult to form a common recognition if background knowledge of researchers differs. Also, if background knowledge differs, the representation of situations also differs, which makes it difficult to decide what kind of data acquisition is sufficient to achieve a purpose.

In this paper, as a precursor for creating our methodology, we collected explanations of BD from researchers with different background knowledge and discussed what kind of data (sets of variables) we should acquire to detect BD using VARIABLE QUEST (VQ). VQ is the method of inferring Variable Labels unifying cooccurrence graphs of variables in the datasets. Variable Labels (VLs) are the names/meanings of variables in datasets. Our approach is to input the explanations of BD to VQ and obtain sets of VLs as data related to BD. We compared and discussed the differences among terms in the explanations and obtained VLs to understand the different perspectives of researchers.

The remainder of this paper is organized as follows. In Section 2, we explain the details of our approach to detect the situation of BD. In Sections 3 and 4, we show the techniques used in the experiment, i.e., Data Jackets (DJs) and VQ. Section 5 describes the purpose and experimental details. Section 6 shows the preliminary experiment for testing the performance and the reproductivity of our proposed approach using test data. In Section 7, we show the results of the analysis, and we discussed them in Section 8. Finally, Section 9 concludes with a brief review and discussion of future work.

2. Our Approach

Humans have acquired and used various data for making decisions in business, politics, or, even, daily activities. However, agents in different fields have different background knowledge. They sometimes use different terminology to explain the same concept or event. Conversely, different agents use the same term to understand different events globally. Thus, it is possible that recognition of the situations may differ by agents' background knowledge. From the studies in cognitive science, it was shown that two problem solvers might construct different facts even if they observe the same data because of the different perspectives provided by their contexts and background knowledge [16]. Metcalfe explained the same phenomenon from the field of economics. Decision makers in society make the most rational choices individually. However, they may recognize different worlds, even though they see the same world because of their background knowledge and available opportunities [17]. Boisot and Canals explained the difference between data, information, and knowledge, and specific types of utility [14]. They demonstrated that data is a property of events and things in the world, and information, by contrast, depends on expectations or states of knowledge.

Figure 1 represents the different conditions of data, information, and knowledge of the agent globally. The model was proposed to understand data, information, and knowledge as different economic factors. Based on this model, the recognition of the agent globally proceeds as follows.

Step 1. World events produce stimuli.

Step 2. The agent receives external stimuli through perceptual filters and acquires data (note that perceptual filters have limitations and cannot attain all stimuli from events).

Step 3. Obtained data are converted into information through conceptual filters (considering the previous studies, mechanisms by which different agents may recognize different worlds while looking at the same event are affected by the conceptual filters).

Step 4. The agent refers to background knowledge and recognizes events globally (the actions toward the world in Figure 1). Also, the recognition gives feedback on the perceptual and conceptual filters (the expectations from the agent's knowledge to each filter in Figure 1).

The purpose of this study is to understand BD. The kinds of data we should observe and collect to understand BD are problematic. It is essential to understand what kind of data the agent acquires to understand unknown phenomena. In this study, the agent is an observer, and the unknown phenomena are the BD. When we apply this model to our research subject, there is a possibility that the data acquired for understanding BD with different background knowledge may differ. Conversely, there is a possibility that the data considered essential for observing BD is common even if agents have different background knowledge. Based on the above discussion, we can summarize the hypotheses of this paper as follows:

Hypothesis 1. Even if the agents have different background knowledge in understanding BD, important data for understanding BD is common.

Hypothesis 2. Because the agents have different background knowledge to understand BD, important data for understanding BD differs.

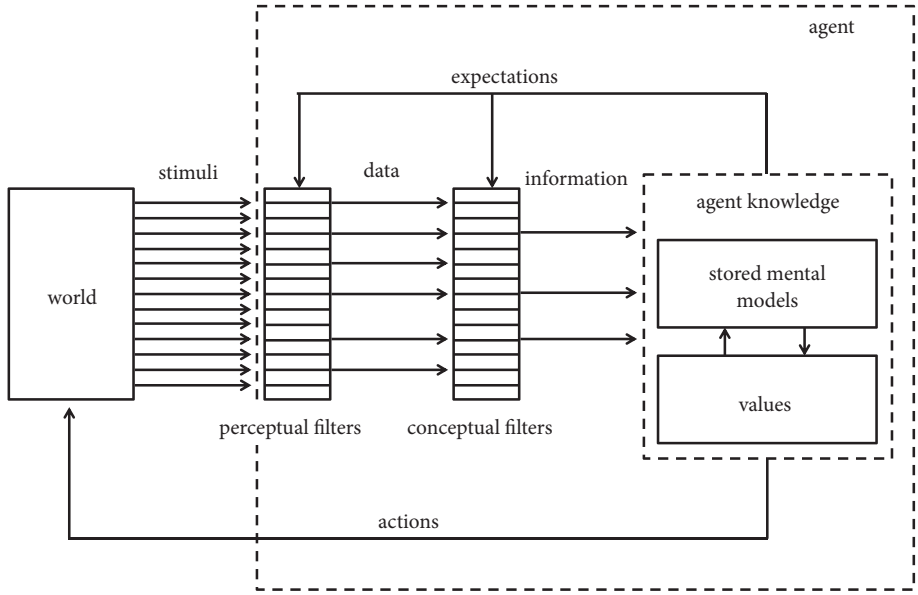

Figure 1: The agent-in-the-world model [14].

To verify the hypotheses above, we collected explanations of BD from researchers with different background knowledge and compared the terms used to explain BD. Because it is difficult to observe the background knowledge of the researchers directly, we assume that explanations of BD represent background knowledge. Also, it is difficult to observe information on data important for understanding the events directly; we use Data Jackets (DJs) as summary information on data and Variable Labels (VLs) in DJs as the detailed information about variables in data. The detailed explanations about DJs and VLs are explained in the next section.

3. Data Jacket (DJ)

Data Jacket (DJ) is a technique used for sharing information about data and for considering the potential value of datasets with the data being hidden. The idea of DJ is to share "a summary of data" in natural language as meta-data without sharing specific data [18]. Sharing the summaries of data as DJs enables data holders to provide information on their data, reducing the risk of data management, cost, and privacy. Also, data users can easily find data related to their interests through descriptions of DJs [19]. Table 1 is an example of DJ of "Vegetable Production in Japan." In DJs, variables are described by Variable Labels (VLs). A VL is the name/meaning of variables in datasets. In DJs, variables in data are summarized as VLs, which are the meta-data of variables and values in datasets. For example, the dataset "Vegetable Production in Japan" includes VLs "location of producer," "weekly (or even daily) production of each producer," "weekly expenditure on vegetable production," "selling prices," and "the type of vegetable."

In this research, background knowledge is given by sentences (set of terms), and the data important to observe the situation of BD requires a set of variables. DJs and VLs are summarized information described in natural language, and over 1,000 pieces of information on data and roughly 5,000 VLs have been collected from different domains. Although larger published databases such as DBpedia [20] are provided as Linked Open Data, they specialize in publicly available data. DJ is not limited to public data and contains information about variables from private companies and individuals. To understand the situation of BD, it is reasonable to use the dataset including various information on variables.

In this paper, to understand what kinds of data (sets of VLs) should be acquired to understand situations of BD, we use the outlines of data and VLs in DJs as corpus data of VARIABLE QUEST explained in detail in the next section.

4. VARIABLE QUEST (VQ)

4.1. The Overview of VQ. VARIABLE QUEST (VQ) is the network visualization system of VLs using the matrix-based inferring method of VLs [15, 21]. VQ represents the cooccurrence and the frequency between VLs in DJs. The cooccurrence of VLs is a feature in which there is a highly frequent pair of VLs appearing simultaneously in the data, e.g., "latitude" and "longitude," or "name," "age," and "nationality." VQ introduces the function of the fundamental matrix-based algorithm to infer VLs from outlines of data (ODs) whose VLs are missing or unknown. VQ has two important models to infer VLs as follows.

Model 1. Datasets are similar when their information for explaining data is similar.

Model 2. Datasets have similar VLs when the similarity of datasets are higher.

Table 1: An example of DJ and VLs.

Item	Content
Title of data	Vegetable Production in Japan
Outline of data (OD)	Vegetables are expensive in Japan compared with other countries. Therefore, the production of vegetables is significant for this country. This database records detailed information on the location, quantity, type, and time of vegetable production. Certain institutions could use the data to analyze factors that impact vegetable price and production. Methods could be created to reduce prices, or to profit from reasonably allocating production.
Variable Labels (VLs)	"Location of producer," "Selling prices," "The type of vegetable," "Weekly expenditure on vegetable production," "Weekly (or even daily) production of each producer."
Sharing Policy	Undecided
Formats of data	CSV
Types of data	"Numerical values," "Table," "Text"
How to collect data	Collect from individual sellers in the local market, agricultural firms, and distributors.

Table 2: The examples of top-ten VLs inferred by VQ.

OD_1		OD_2	
Inferred VLs	Similarity	Inferred VLs	Similarity
Number of births	0.349	North latitude (degrees)	0.361
Number of deaths	0.349	Installer of the device	0.361
In-migrants	0.335	Address of the seismic intensity	0.361
Fatalities	0.335	East longitude (degrees)	0.361
Out-migrants	0.335	North latitude (minutes)	0.361
Population	0.321	The name of the seismic intensity	0.361
Number of households	0.318	The pronunciation of seismic intensity	0.361
Population (male)	0.318	Match level	0.361
Population (female)	0.318	Earthquake number	0.360
Fertilities	0.318	Position number	0.360

VQ introduced the bag-of-words and vector space model [22] for creating the corpus from the training data (DJs with VLs). In the preprocessing steps, VQ conducts the morphological analysis of the text of ODs, extracting words, removing stop words, and restoring words to their original forms.

Table 2 shows two examples of inferred VLs. The left column of Table 2 shows the top-ten inferred list of VLs obtained from an OD_1; "this data represents the transition of the population of each year in Tokyo, Japan." The right column is the VLs found from an OD_2 "the earthquake data in the world." We can obtain a set of VLs with similarities to the queries whose VLs are unknown. Even if a free text query does not include terms which represent VLs, VQ returns related sets of VLs with the query.

4.2. Detailed Algorithm of VQ. In this subsection, we explain the detailed algorithm of VQ. At first, VQ conducts an algorithm to calculate the similarity among training data of ODs based on Model 1. ODs are given by the sentences so that we assume that each OD is a set of terms. After conducting the preprocessing steps to ODs using a bag-of-words model, the ODs are converted into a matrix representation (a Term-OD matrix). A Term-OD matrix M ($W \times D$) consists of W-dimensional OD vectors as columns and D-dimensional term vectors as rows. Each element in the matrix M (v_{ij}) in an OD vector ($\boldsymbol{od}_j$) corresponds to the frequency with which a term (a row i) occurs in an OD (a column j) as shown in (1) and (2). Note that the subscript T on the upper-right corner of vectors represents the transposition, and the vectors are highlighted in bold in this paper.

$$M = \left(\boldsymbol{od}_1, \cdots, \boldsymbol{od}_j, \cdots, \boldsymbol{od}_D\right) \tag{1}$$

$$\boldsymbol{od}_j = \left(v_{1j} \ \cdots \ v_{ij} \ \cdots \ v_{Wj}\right)^{\mathrm{T}} \tag{2}$$

In the second step, a set of VLs in DJs is converted into a VL-OD matrix R ($V \times D$). In the training data of DJs, ODs and VLs are linked when they appear in the same DJs. Each element in the matrix R (r_{ij}) in the jth OD vector ($\boldsymbol{od'}_j$) corresponds to the frequency (0 or 1) with which the ith VL occurs in the jth OD as shown in (3) and (4).

$$R = \left(\boldsymbol{od'}_1, \cdots, \boldsymbol{od'}_j, \cdots, \boldsymbol{od'}_D\right) \tag{3}$$

$$\boldsymbol{od'}_j = \left(r_{1j} \ \cdots \ r_{ij} \ \cdots \ r_{Vj}\right)^{\mathrm{T}} \tag{4}$$

In the third step, a Term-VL matrix E ($= MR^{\mathrm{T}}$) is generated by combining a Term-OD matrix M and an OD-VL matrix R^{T}. The Term-VL matrix E is represented by (5), and the jth

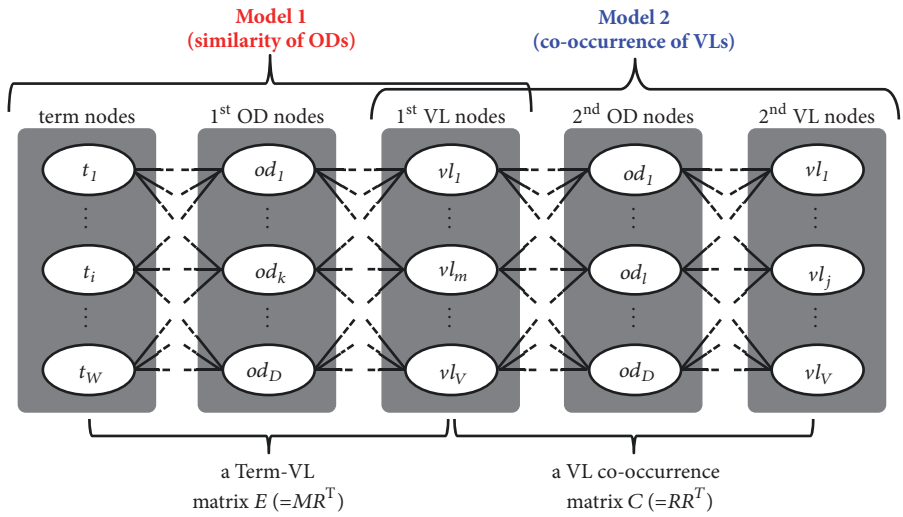

FIGURE 2: The structure of the Term-VL matrix *EC* [15] (partially modified by authors).

VL vector ($\boldsymbol{vl}_j$) is given by (6). The elements of the Term-VL matrix E (e_{ij}) are calculated by (7).

$$E = \left(\boldsymbol{vl}_1, \cdots, \boldsymbol{vl}_j, \cdots, \boldsymbol{vl}_V\right) \tag{5}$$

$$\boldsymbol{vl}_j = \left(e_{1j} \ \cdots \ e_{ij} \ \cdots \ e_{Wj}\right)^{\mathrm{T}} \tag{6}$$

$$e_{ij} = \sum_{k=1}^{D} v_{ik} r_{kj} \tag{7}$$

Each element in the Term-VL matrix E means the sum of the product of the frequency (v_{ik}) with which the ith term (t_i) occurs in the kth OD (od_k) and the frequency (r_{jk}) with which the jth VL (vl_j) links with the kth OD (od_k).

In the fourth step, we conduct the VL cooccurrence matrix C (= RR^{T}), assuming that any pair of VLs in the same DJ occurs once based on Model 2. The elements in the VL cooccurrence matrix C (c_{ij}) represent the number of DJs which include a pair of the ith VL (vl_i) and the jth VL (vl_j) as shown in (8).

$$c_{ij} = \sum_{k=1}^{D} r_{ik} r_{kj} \tag{8}$$

In the fifth step, we acquired a Term-VL matrix *EC* by a product of the Term-VL matrix E and the VL cooccurrence matrix C. The Term-VL matrix *EC* consists of V-dimensional term vectors as rows and W-dimensional VL vectors as columns. The structure of the Term-VL matrix *EC* is the same as that of the Term-VL matrix E. The element g_{ij} of the matrix *EC* is given by (9). The value is calculated by the similarities of ODs and queries, which is the function of the matrix E, and the cooccurrence of VLs, which is the function of the matrix C.

$$g_{ij} = \sum_{m=1}^{V} \left(\sum_{k=1}^{D} v_{ik} r_{km}\right)\left(\sum_{l=1}^{D} r_{ml} r_{lj}\right) \tag{9}$$

The structure of the Term-VL matrix *EC* is equivalent to the adjacency matrix of the 5-partite graph as shown in Figure 2. When OD_x whose VLs are unknown is inputted, VQ calculates a W-dimensional feature vector of OD_x ($\boldsymbol{od}_x$) referring to the dictionary (the list of terms) and the corpus. By calculating the similarities of feature vector of OD_x ($\boldsymbol{od}_x$) and each feature vector of VL ($\boldsymbol{vl}_j$) by the matrix *EC*, VQ returns a scored set of VLs with similarities. In this paper, the similarity scores of $\boldsymbol{od}_x$ and $\boldsymbol{vl}_j$ are calculated by cosine similarities (similarity($\boldsymbol{od}_x, \boldsymbol{vl}_j$) = $\boldsymbol{od}_x \cdot \boldsymbol{vl}_j / |\boldsymbol{od}_x||\boldsymbol{vl}_j|$).

In this paper, we use VQ to externalize VLs related to the state of BD, even if the terms explaining BD differ among researchers. In the next section, we explain the purpose and the method of our experiment.

5. Experimental Details

5.1. Purpose and Method. The purpose of this experiment is to acquire sets of VLs necessary to understand the situation of the users whose beliefs have drifted. To achieve this goal, we obtained linguistic sentences to explain the state of BD from researchers with different background knowledge. Using VQ with VLs, we acquired relevant data (a set of VLs) from sentences concerning BD. We attained linguistic sentences from six researchers to explain the state of BD. Three researchers are from the department of engineering, two are medical doctors, and one is a psychotherapist. All researchers hold Ph.D. in their fields. Researchers have discussed BD through several meetings for roughly one and a half years. Although six samples are relatively small, because BD is a new research topic, such samples cannot be collected elsewhere. To supplement the small number of samples, we introduce the index of commonality (10) instead of quantitative evaluation and compared degrees of variation.

We asked the researchers to write down "the state of Belief Drift you understand from your own background knowledge." To avoid the bias of others' answers, we asked them to submit the sentences separately. We input these

Table 3: Corpus statistics of VQ (parentheses represent the standard deviation).

Number of DJs	1,032
Total number of terms in DJs	27,194
Unique terms in DJs	4,971
Mean of the number of terms in each DJ	26.4 (58.5)
Maximum number of terms in DJs	1471
Minimum number of terms in DJs	1
Total number of VLs in DJs	7,029
Unique VLs in DJs	5,155
Mean of the number of VLs in each DJ	6.81 (8.80)
Maximum number of VLs in DJs	118
Minimum number of VLs in DJs	1

answers to VQ as queries described in Section 4 and obtained sets of top-30 likely VLs. Note that explanations of BD and some VLs are given in Japanese. In this paper, we translated all terms into English after the experiment.

5.2. Datasets and Method for Evaluations. To construct the corpus for VQ, we use 1,032 DJs including outlines of data (ODs) and VLs, collected from business persons, researchers, and data holders. All DJs are extracted from DJ Store which is a database with a retrieval system for DJs on the Web and is provided in RDF format [23]. Table 3 shows the statistics of corpus data. Each DJ has 6.8 VLs on average. 5,155 unique VLs are stored in total. The corpus and the dictionary were constructed from all the words in OD texts. The OD corpus consists of 4,971 unique words. We used MeCab for the morphological analysis [24], which is a common tool for analyzing morphemes of Japanese texts.

For weighing discriminative terms in the corpus, we used tf-idf in a weighting scheme [25], which is reliable for identifying distinctive terms in documents. The term frequency (tf) is the number of times a term appears in a document, and the inverse document frequency (idf) diminishes the weight of frequent terms in all documents and increases the weight of those terms which appear rarely. When inputting the sentences about BD in VQ, we removed punctuation marks and symbols in the texts as stop words, restored words to their original forms, and extracted nouns, verbs, adverbs, and adjectives as a preprocess.

To evaluate the commonality among researchers, we define the indicator of commonality. Equation (10) is an indicator for evaluating the degree of commonality of elements among clusters, which calculates the proportion of excluding terms appearing only once. T_i represents the ith set of elements, and $|T_i|$ represents the number of elements in the ith set T_i. Note that $n = 6$ and *ele* is term.

$$\text{commonality}(ele) = 1 - \frac{\sum_{i=1}^{n} \left| T_i - \bigcup_{j=1, j\neq i}^{n} T_j \right|}{\left| \bigcup_{i=1}^{n} T_i \right|} \quad (10)$$

To test the performance and the reproductivity of our approach, we first conduct the preliminary experiment in the next section by using the test data.

Table 4: Commonalities of terms and VLs using test data at random (parentheses represent the standard deviation).

Commonality	Terms	VLs
Mean	0.116 (0.037)	0.055 (0.041)
Median	0.112	0.045
Maximum value	0.208	0.200
Minimum value	0.037	0.000

6. Preliminary Study

6.1. Experiment. At first, by using the same corpus of DJs shown in Table 3, we compared the commonality of terms in DJs and VLs obtained from VQ. The purpose of this study is to understand the kinds of data we should observe and collect to understand the situation of BD. To adjust to this goal, we set three different themes ("transportation," "health," and "personal activity"). We extracted 7 DJs related to each theme, obtained top-30 VLs, and calculated the commonalities of terms and VLs. To compare the performance, we also calculated the commonalities of terms and VLs randomly. We repeated the following steps for 80 times: (1) choosing 10 DJs at random from the population of DJs, (2) inputting texts in ODs in VQ, (3) obtaining the top-30 likely VLs from each OD, and (4) calculating the commonality values of terms and VLs.

6.2. Result and Discussion. Tables 4 and 5 are the results of the preliminary experiments. Table 4 shows the commonality of randomly selected DJs. The commonality of terms is about twice as high as that of VLs on average. The median, the maximum, and the minimum of commonalities are also higher in those of terms. Moreover, there is a significant difference within comparison with a paired t-test, assuming the equal variances ($t(158) = 11.48, p < 0.01$). In other words, the commonality of VLs is significantly lower than that of terms. However, when we compared the commonalities of selected three themes, the result shows that the commonalities of VLs are 1.7 to 2.4 times higher than those of terms. This result is totally different compared with randomly selected DJs.

This result is caused by the difference between the term and VL distributions. Although the total number of terms is 27,194, the number of unique terms is 4,971, which is less than that of VLs (Table 3). That is, many terms appear more than once. Figures 3 and 4 show the distributions of the numbers of terms and VLs in the corpus. The left graphs of Figures 3 and 4 are the double logarithmic graphs. The numbers of terms and VLs in each DJ are shown on the horizontal axis and the proportion on the vertical axis. However, the probability of the frequency (k) $p(k)$ is small in the portion where k is large, and there are very few k for which $p(k) > 0$. Owing to this, the double logarithmic graph of the distribution of the frequency is weak toward the noise. Accordingly, we added an order plot on the right of Figures 3 and 4, which is equivalent to the cumulative distribution. Terms and VLs are in accordance with $p(k) \propto k^{-\gamma}$, which becomes a power distribution for both terms and VLs. The power index γ of the term distribution is 1.96 (coefficient of determination:

Table 5: Commonalities of terms and VLs of three themes.

	"transportation"		"health"		"personal activity"	
	Terms	VLs	Terms	VLs	Terms	VLs
Commonality	0.152	0.360	0.153	0.264	0.132	0.229

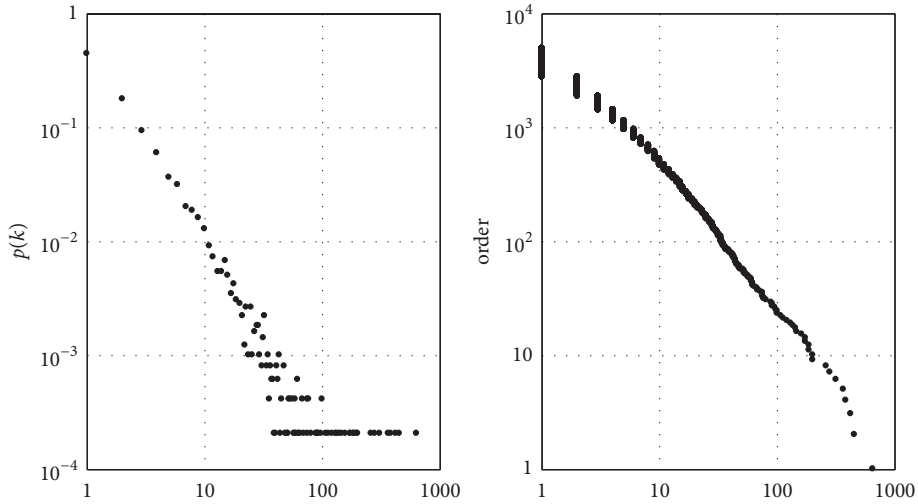

Figure 3: Distributions of the terms in Corpus.

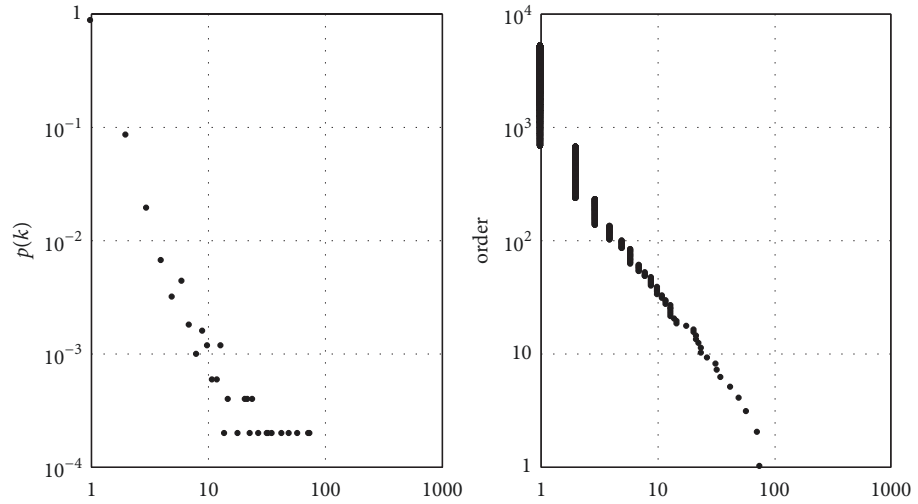

Figure 4: Distributions of the VLs in Corpus.

0.96), and that of the VL distribution is 3.09 (coefficient of determination: 0.72).

The power index of VL distribution is larger than that of terms, which shows that the frequency of many kinds of VLs is rather small. When we randomly choose DJs, the probability of obtaining common VLs using VQ is extremely low. That is, the perspective of the event is so different that we cannot acquire the common VLs. However, by setting the themes, even if the commonality of terms to explain the events is low, the VLs that can be acquired are common to some extent. In other words, the constraints to the perspective of an event by the themes have a function to share data necessary for understanding the event even if the terms expressing the event are somewhat different. For example, if you input the terms "car" and "vehicle" related to the theme "traffic," you can get the relevant VLs "traffic volume" or "location." VQ may be able to bridge the vocabulary gap and obtain the common VLs from different terms when the theme is common.

The purpose of this paper is to recognize the situation of BD in common and to clarify the data (set of VLs) to be acquired. Based on the result and the discussion on the preliminary experiment, it is appropriate to acquire a sentence to explain the BD in the main experiment, obtain VLs from VQ, and compare the commonalities. The same applies to the data on "transportation," "health," and "personal activity"; it

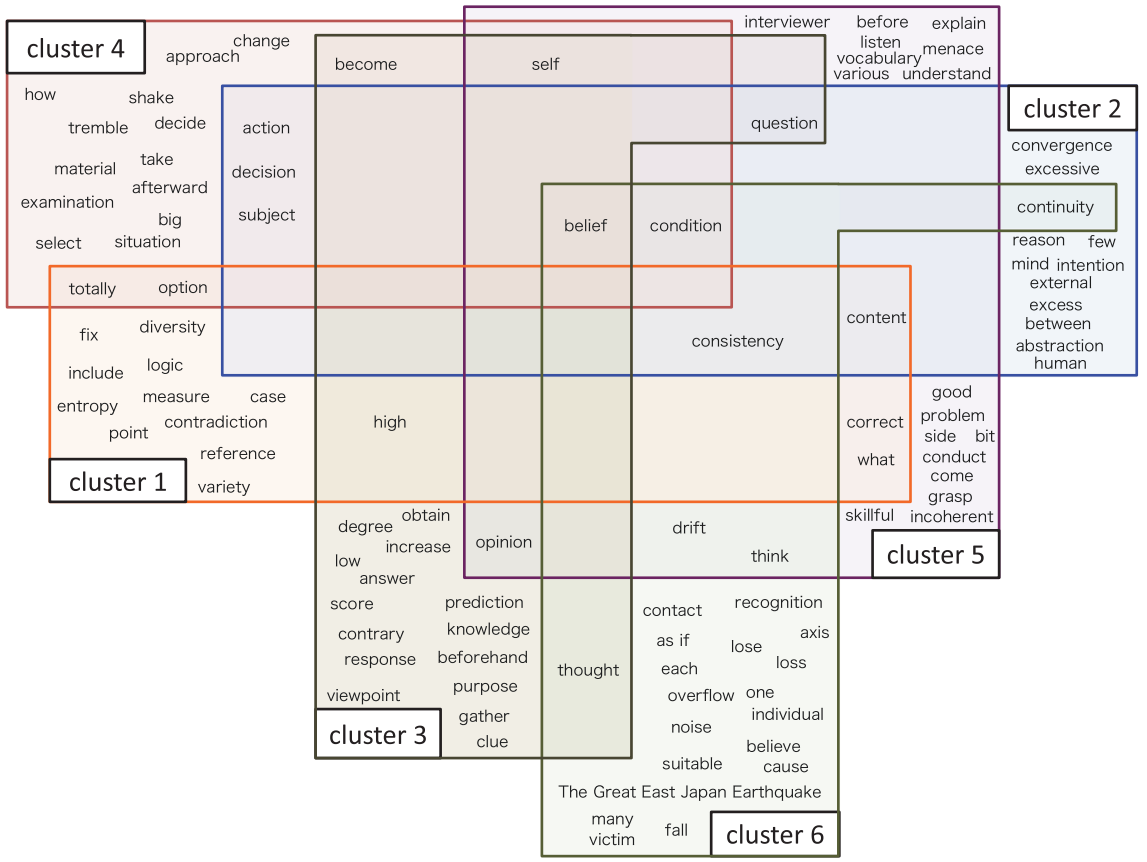

Figure 5: The clustering of terms in sentences about Belief Drift.

can be said that our proposed method works effectively in the situation of BD if the commonality of VLs is higher than the commonality of terms. We can expect the reproducibility of the method in other themes as well. We examine how this result contributes to specifying the data (set of VLs) necessary to explain the situation of DB in the main experiment of the next section.

7. Results

In this paper, we discuss BD by separating engineering researchers and medical researchers (a medical doctor and a psychotherapist). Figure 5 shows the results of morpheme analysis of explanations of BD collected from six researchers. One hundred and five unique terms appear in the sentences in total. Figure 6 shows the frequency of terms shared by more than two researchers. Clusters 1, 2, and 3 are engineering researchers, and clusters 4, 5, and 6 are medical researchers. Table 6 is the number of unique terms in each cluster. Since the numbers are not overly biased, we discuss differences by comparing the number of terms.

We calculate the degree of commonality of elements among researchers by using (10). As a result, the commonality of terms is 0.190 (Table 7). We also compare the commonality of terms between engineering researchers and medical researchers. We define T_{eng} ($T_{eng} = T_1 \cup T_2 \cup T_3$) as the element set of engineering researchers and

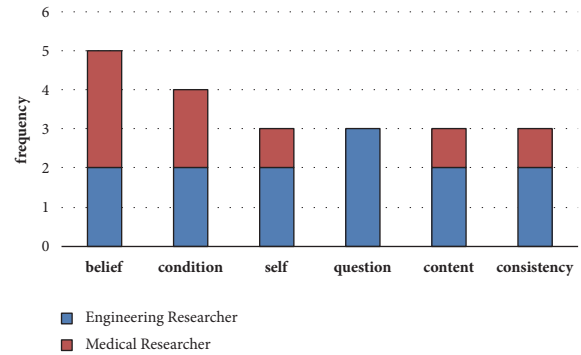

Figure 6: The frequency of terms shared by more than two researchers.

Table 6: Corpus statistics of VQ.

Cluster #	the number of unique terms
1	18
2	20
3	22
4	22
5	28
6	25

Table 7: Commonalities of terms and VLs.

	Terms	VLs
Number of unique elements	105	104
Commonality of all clusters	0.190	0.596
Commonality between clusters of Engineering Researcher and Medical Researcher	0.143	0.519

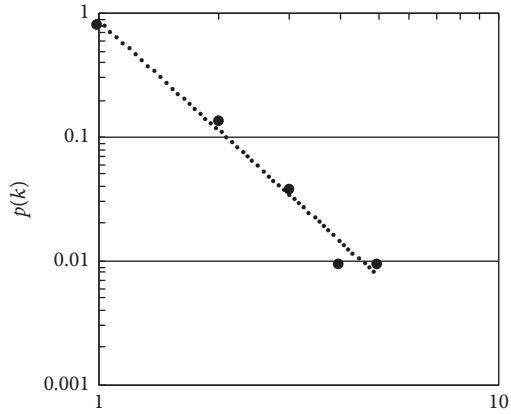

Figure 7: The distribution of terms in the logarithmic graph (the vertical axis represents occurrence probability, and the horizontal axis represents the frequency of terms).

T_{med} ($T_{med} = T_4 \cup T_5 \cup T_6$) as the element set of medical researchers. The commonality is equivalent to a Jaccard index, and it can be represented by commonality(ele) = $|T_{eng} \cap T_{med}|/|T_{eng} \cup T_{med}|$. The commonality of terms between engineering and medical researchers is 0.143 (Table 7). Figure 7 is the distribution of terms in the logarithmic graph.

Conversely, we input texts of each researcher concerning BD in VQ and obtain the top-30 likely VLs. We attained 104 unique VLs in total, which is not very different from the number of terms. Moreover, the similarity for queries of all 30 VLs is higher than 0.110. The maximum is 0.365, and the minimum is 0.113. Two VLs of 0.365 are "the specific scale of renal disease" and "inclusive scale," and nine VLs of 0.113 include "sanitary conditions" and "allergic symptoms."

Figure 8 is the result of the classification of VLs. Some VLs are identified by several phrases such as "non-psychotic psychiatric disorder (symptoms of schizophrenia)" or "ADL score." Note that, since part of the descriptions of VLs protrudes from the frames, VLs are attached with green nodes in Figure 8. Additionally, Figure 9 shows the frequency of VLs shared by more than two researchers. As in Figure 6, clusters 1, 2, and 3 are engineering researchers, and clusters 4, 5, and 6 are medical researchers. Applying (10) to the VL sets, the commonality of VLs is 0.596 (Table 7). Figure 10 is the distribution of VLs in the logarithmic graph.

8. Discussion

Looking at Figure 5, there are relatively few common terms. Only six out of 105 terms are common among over three researchers (Figure 6) and the commonality of terms is 0.190 (Table 7). The terms "belief" and "condition" are relatively common among researchers, but many terms appear only once. The power exponent of the frequency of terms is $\gamma = 2.94$ (the determination coefficient is 0.980) in Figure 7. The frequency of terms follows $p(k) \propto k^{-\gamma}$, which shows that the distribution of terms is a power distribution. The results show that common terms are rarely used to explain BD. That is, there is no commonality among researchers to explain BD.

On the other hand, compared with the commonality of terms in sentences explaining BD, the commonality of VLs is 0.596, which is 3.14 times higher (Table 7). Additionally, fifteen out of 104 VLs are common among more than two researchers (Figure 9), which is larger than for terms. This result is the same as the result when giving the theme as the constraint in the preliminary experiment. Figure 10 shows that the distribution of VLs is not a power distribution, but is the exponential distribution ($p(k') = 2.25\exp(-1.14k')$ and the determination coefficient is 0.883). This result means that, in the distribution of VLs, the extremely low frequent elements do not occupy the majority as much as the power distribution. Although there are low frequent elements in the exponential distribution, the degree to which the low frequent elements are dominant over the whole is smaller when compared to the power distribution. It shows that VLs important for understanding BD are common among researchers compared with the terms. We also compare the commonality of VLs between engineering researchers and medical researchers. Accordingly, the commonality of VLs between engineering and medical researchers is 0.519, which is 3.62 higher than that in the terms (Table 7). The commonality of the terms in the sentences to explain BD is low not only among all the clusters, but also between engineering researchers and medical researchers. However, when considering VLs, the number of common VLs increases and the commonality is higher than terms.

Considering both the commonalities and distributions, the results suggest that, even if the terms used to explain the state of BD differ, the data (sets of VLs) to be acquired to understand BD are common to some extent. The higher commonality of VLs is similar to the result of giving the theme as the constraint in the preliminary experiment. As we expected, the reproducibility of the method can be guaranteed when the themes are given as the constraints when we use VQ to extract the common data to be acquired to understand the certain situations. Thus, the result supports Hypothesis 1: "even if the agents have different background knowledge in understanding BD, important data for understanding BD is common." Moreover, although it is possible that the vocabulary gap may interrupt a discussion on data necessary for decision making, VQ may be able to bridge the vocabulary

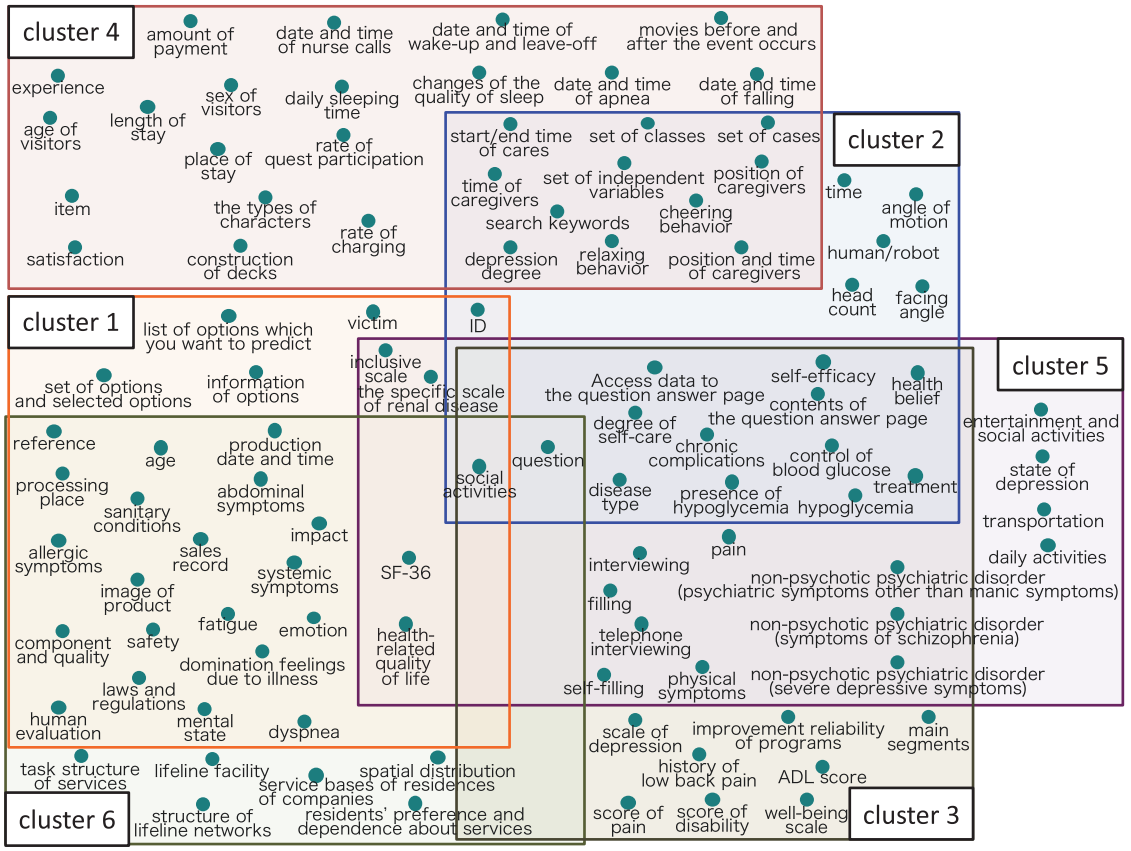

Figure 8: The clustering of VLs inferred from the terms to explain Belief Drift.

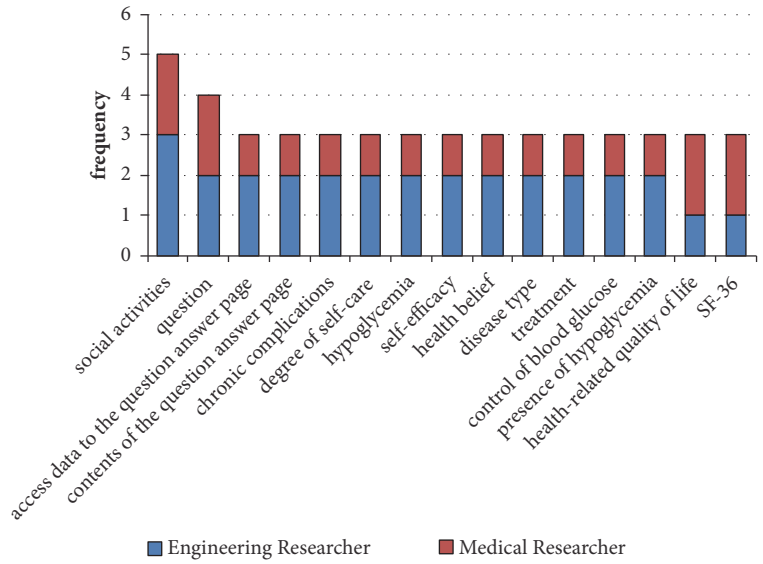

Figure 9: The frequency of VLs shared by more than two researchers.

gap and infer related VLs even if the terms used to explain the common event differ because of the various background knowledge.

The interesting point of the results is that VQ suggests "social activities," "question," "access data to the question answer page," and "contents of the question answer page" which are most common among researchers. In the research project of BD, we are analyzing text data of the question answer site to evaluate the degree of BD. The results of this paper strongly support the hypothesis that the text analysis

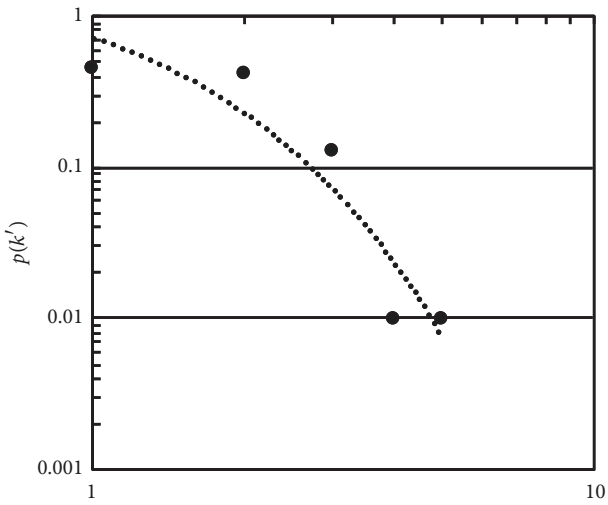

Figure 10: The distribution of VLs in the logarithmic graph (the vertical axis represents occurrence probability, and the horizontal axis represents the frequency of VLs).

of the question answer page may be useful for evaluating and detecting the state of BD.

9. Conclusion

9.1. Summary. Agents with different background knowledge may use different words to describe common concepts. Alternately stated, despite the same framework of thought or understanding, vocabulary gaps may occur when explaining common events or problems such as Belief Drift. Vocabulary gaps may make it difficult to form a common recognition for action, i.e., the acquisition of data or analysis. To bridge these gaps, we tried to extract sets of Variable Labels from different terms explaining the common concept using VARIABLE QUEST. Consequently, although the commonality of terms was low, the commonality of Variable Labels was higher. In other words, even though the terms used to explain events or problems differ, since the framework of thought and understanding are relatively the same, the Variable Labels necessary for understanding the state of BD attained higher commonality.

However, in order to obtain the same performance as these results, a certain amount of information on data (DJ in this paper) is required. It can be said that it is difficult to get satisfactory results if the corpus data is minimal. Also, it is necessary to test the selection protocols of themes given as constraints. In this paper, we did not discuss how much data is reasonable for obtaining results. For our future study, it is necessary to clarify the appropriate number of terms and VLs to obtain sufficient results.

9.2. Future Work. The essential mission of our project is to establish the fundamental methodology and systems providing appropriate information for people whose beliefs are drifting. The result of this experiment suggests information about variables and data we need to acquire to understand situations of BD caused by various factors. In the next stage of our project, we will obtain data according to the advice of medical researchers. Moreover, we will integrate our result with the suggestion obtained from the text analysis of the question answer page.

Data Availability

Data from Data Jacket with Variable Labels and the results used to support the findings of this study have been deposited in the Data Jacket Store repository stored in RDF/XML (http://160.16.227.37/sparql) and are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was partially supported by JST-CREST Grant Number JPMJCR1304 and JSPS KAKENHI Grants Numbers JP16H01836 and JP16K12428. We thank research members of JSPS KAKENHI Grant Number JP16H01836 for proposing the definition of Belief Drift, which worked as the initial input data for the study of this paper. We also appreciate that the tool VARIABLE QUEST used in this paper is based on the presented paper and the fruitful discussion in the MoDAT workshop in IEEE International Conference on Data Mining Workshops (ICDMW) 2017.

References

[1] "Official website of the Department of Homeland Security," https://www.dhs.gov/topic/disasters.

[2] "Official website of the Department of Energy," https://www.energy.gov/.

[3] "Basic Policy and Action Plan for Building IT Disaster-Management Lifeline," in *Proceedings of the IT Disaster-Management Lifeline Promotion Conference in Japan*, 2012, https://japan.kantei.go.jp/policy/it/_full.pdf.

[4] Ministry of Internal Affairs and Communications, "Japan's International Contribution in the Field of ICT for Disaster Management," http://www.soumu.go.jp/menu_seisaku/ictseisaku/bousai_ict/eng/.

[5] Infrastructure Ministry of Land, "Disaster Prevention Portal," http://www.mlit.go.jp/river/bousai/olympic/en/index.html.

[6] M. Tsubokura, S. Kato, S. Nomura et al., "Absence of internal radiation contamination by radioactive cesium among children affected by the fukushima daiichi nuclear power plant disaster," *Health Physics Journal*, vol. 108, no. 1, pp. 39–43, 2015.

[7] M. Tsubokura, S. Kato, S. Nomura et al., "Reduction of high levels of internal radio-contamination by dietary intervention in residents of areas affected by the Fukushima Daiichi nuclear plant disaster: A case series," *PLoS ONE*, vol. 9, no. 6, 2014.

[8] M. Tsubokura, K. Hara, T. Matsumura et al., "The immediate physical and mental health crisis in residents proximal to the evacuation zone after Japan's nuclear disaster: An observational pilot study," *Disaster Medicine and Public Health Preparedness*, vol. 8, no. 1, pp. 30–36, 2014.

[9] Y. Nara, "A Cross-cultural Study on Trust and Risk Perception among Japan, China, and the United States: Focusing on Earthquakes and Nuclear Power Plant Accidents," in *Proceedings of the 16th International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, pp. 1609–1621, 2012.

[10] K. B. Moysich, R. J. Menezes, and A. M. Michalek, "Chernobyl-related ionising radiation exposure and cancer risk: an epidemiological review," *The Lancet Oncology*, vol. 3, no. 5, pp. 269–279, 2002.

[11] H. Nakada, N. Murashige, T. Matsumura, Y. Kodama, and M. Kami, "Informal network of communication tools played an important role in sharing safety information on H1N1 influenza vaccine," *Clinical Infectious Diseases*, vol. 51, no. 7, pp. 873-874, 2010.

[12] A. Sugimoto, S. Krull, S. Nomura, T. Morita, and M. Tsubokura, "The voice of the most vulnerable: Lessons from the nuclear crisis in Fukushima, Japan," *Bulletin of the World Health Organization*, vol. 90, no. 8, pp. 629-630, 2012.

[13] Y. Ohsawa, N. Kushiro, M. Hirano, M. Tsubokura, and M. Kami, "Technologies for Generating and Providing Information to Suppress Belief Drifts and Reinforce Psychological Resilience," in *a Project of JSPS KAKENHI Grant Number JP16H0*, https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-16H01836/.

[14] M. Boisot and A. Canals, "Data, information and knowledge: Have we got it right?" *Journal of Evolutionary Economics*, vol. 14, no. 1, pp. 43–67, 2004.

[15] T. Hayashi and Y. Ohsawa, "VARIABLE QUEST: Network Visualization of Variable Labels Unifying Co-occurrence Graphs," in *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 577–583, New Orleans, LA, November 2017.

[16] Y. Hayashi, K. Miwa, and J. Morita, "A laboratory study on distributed problem solving by taking different perspectives," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 333–338, 2006.

[17] J. S. Metcalfe, *Evolutionary Economics and Creative Destruction*, Routledge, UK, 1998.

[18] Y. Ohsawa, H. Kido, T. Hayashi, and C. Liu, "Data jackets for synthesizing values in the market of data," *Procedia Computer Science*, vol. 22, pp. 709–716, 2013.

[19] Y. Ohsawa, H. Kido, T. Hayashi, C. Liu, and K. Komoda, "Innovators marketplace on data jackets, for valuating, sharing, and synthesizing data," in *Knowledge-Based Information Systems in Practice*, vol. 30 of *Smart Innovation, Systems and Technologies*, pp. 83–97, Springer International Publishing, Cham, 2015.

[20] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in *The Semantic Web*, vol. 4825 of *Lecture Notes in Computer Science*, pp. 722–735, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[21] T. Hayashi and Y. Ohsawa, "Matrix-based method for inferring variable labels using outlines of data in data jackets," in *Advances in Knowledge Discovery and Data Mining*, vol. 10235 of *Lecture Notes in Computer Science*, pp. 696–707, Springer International Publishing, Cham, 2017.

[22] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[23] T. Hayashi and Y. Ohsawa, "Knowledge structuring and reuse system design using RDF for creating a market of data," in *Proceedings of the 2nd International Conference on Signal Processing and Integrated Networks, SPIN 2015*, pp. 607–612, India, February 2015.

[24] T. Kudo and Y. Matsumoto, "Japanese dependency structure analysis based on support vector machines," in *Proceedings of the the 2000 Joint SIGDAT conference*, pp. 18–25, Hong Kong, October 2000.

[25] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.