2nd European Workshop on Chance Discovery and Data Synthesis (EWCDDS16)

29 Aug, 2016

World Forum
Johan de Wittlaan 42-44, 2517 JR Den Haag, Netherlands

Program Committee

Akinori Abe (Chiba University, JAPAN) **co-Chair**
David Bergner (Sofia University in Palo Alto, USA)
Lorenzo Magnani (University of Pavia, Pavia, Italy) **co-Chair**
Peter McBurney (King's College London, UK)
Yukio Ohsawa (The University of Tokyo, Japan) **co-Chair**
Shusaku Tsumoto (Shimane University, Japan)
Katsuyoshi Yada (Kansai University, Japan)

ECAI2016
2nd European Workshop on Chance Discovery and Data Synthesis (EWCDDS16)

SCHEDULE

08:55–09:00 **Opening remarks**

09:00–09:30 **Clustering Documents Using Structural Similarity Based on Case Sets —Applied for Technological Problems from Patents—**
Hitomi Yanaka and Yukio Ohsawa

09:30–10:00 **Chance Curation in Virtual Cognitive Niches.**
Selene Arfini, Tommaso Bertolotti, and Lorenzo Magnani

10:00–10:30 **Automatic Identification of Trigger Factors: a Possibility for Chance Discovery.**
Marharyta Aleksandrova, Armelle Brun, Oleg Chertov, and Anne Boyer

10:30–11:00 **Coffee break**

11:00–11:30 **Disasters and Transformation of Daily Life: Implications for Issues in Risk Management.**
Yumiko Nara

11:30–12:00 **Emergence of Option Prices in Markets Populated by Portfolio-Holders.**
Sarvar Abdullaev, Peter McBurney, and Katarzyna Musial

12:00–12:30 **On the logical and ontological treatment of IMDJ data.**
Akinori Abe

12:30–14:00 **Lunch break**

14:00–14:30 **Study Chance Discovery in Temporal Linear Non-Transitive Logic with Agent's Knowledge**
Vladimir V. Rybakov

14:30–15:00 **Preliminary Case Study about Analysis Scenarios and Actual Data Analysis in the Market of Data**
Teruaki Hayashi and Yukio Ohsawa

15:00–15:30 **Coffee break**

15:30–17:30 **Mini session on Data Curation in the Market of Data**
organised by Teruaki Hayashi

17:30–17:35 **closing workshop**

ECAI2016
2nd European Workshop on Chance Discovery and Data Synthesis (EWCDDS16)

CONTENTS

Clustering Documents Using Structural Similarity Based on Case Sets -Applied for Technological Problems from Patents-

Hitomi Yanaka and **Yukio Ohsawa**[1]

Abstract. The description of technological problems in patent documents is important to understand the motivation of the invented technology. Understanding the motivation helps us to analyze trends of the technologies contained in a set of patent documents. Here, we approach the classification of documents based on analogy of structures of the problem descriptions. The purpose of this study is to develop a method for patent classification, with the use of hierarchical clustering based on the structural similarity of problems to be solved by the patented invention. First, we present an approach for extracting predicate-argument structures in the contents of patents. Second, we propose the similarity function to measure the structural similarity between the case sets. The result of the questionnaire survey showed that the structural similarity between patent documents can be calculated with the use of the predicate-argument structures. Furthermore, the survey indicated that comprehension of document structures can be increased by reading the documents reconstructed by the predicate-argument structures.

1 INTRODUCTION

Patent documents provide relevant knowledge of previous inventions and their motivations. New inventors can take the advantage of those previous inventions and learn possible trends of technological problems for future inventions. Due to the variety and large amount of documents of previous inventions, the classification of those helps the search of a relevant document for a new context and problem. Currently, existing standard codes of classification systems are used to support patent document search. Patent examiners can search similar patent documents written in different technical words by using classification codes. There are different kinds of classification systems. For example, IPC (International Patent Classification) is a classification system to grant the classification code in the patent document in accordance with the hierarchy of technical content. Some classification systems are proposed in each region. For example, FI(File Index) is used in Japan, ECLA(European Classification) is used in Europe, and USPC(U.S. Patent Classification) is used in the U.S. Patent examiners use these different classification systems depending on their purposes. Furthermore, these classification codes have been updated manually by experts. For improving the robustness and efficiency of the patent classifications, a method to classify patents automatically based on the technological problems has been demanded.

[1] Department of Systems Innovation, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, email: h2.yanaka@gmail.com, ohsawa@sys.t.u-tokyo.ac.jp

A patent map is also a method of patent classification. Here patent information is collected for a specific purpose of use and depicted in a visual form of presentation such as a chart, matrix, graph, or table. Fig.1 shows an example of patent matrix map of fuel cell. Like this, the trend of target technology fields can be visualized by the patent map. However, details of patent documents contained in each bubble cannot be identified at a glance. In addition, target technology fields have been also specified manually by experts.

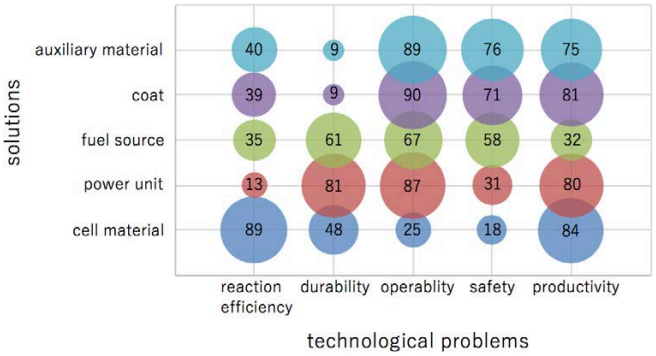

Figure 1. Example of matrix map of fuel cell: the size of pie chart shows the number of patents (in accordance with this technology and the effectiveness of patent). The figure in a pie shows the number of registered documents.

In patent analysis with these methods, analogy is expected to be useful for solving problems in technology development. The analogy is the process toward understanding the problem and solving it from the relation between the base (sometimes called the source) and the target. In patent analysis, the base is the patent data that is already familiar with, whereas the target is the technological problem that we want to understand and solve. In analogy, the correspondence between the target and the base, if noticed, triggers the solution of the target.

According to the structure mapping theory [6][3], there are two kinds of similarities: superficial similarity and structural similarity. The superficial similarity is characterized by elements contained in the target and the base. For example, let us take the following two sentences:

Sentence 1: Water is cold.

Sentence 2: Water is liquid.

Then, both sentences have the same word “water” and the superficial similarity is completed. Relation of the sentences is written below.

$cold\,(water) \rightarrow liquid\,(water)$

The structural similarity is characterized by the primary or higher-order relationships between elements included in them.

Sentence 3: A planet revolves around the sun.

Sentence 4: An electron revolves around the atom.

Then, both sentences have the same structure, "A revolves around B" and structural similarity is completed. Relation of the sentences is written below.

$revolve - around\,(planet, thesun)$

$\rightarrow revolve - around\,(electron, atom)$

MAC / FAC (Many Are Called, but Few Are Chosen) model has been proposed as a model to search common elements between the base and the target based on the superficial similarity, and to assess the validity of the reasoning by evaluating the structural similarity between the base and target [5]. If two patents of different fields of technology are similar in problem structures, they are different in the superficial similarity but similar in the structural similarity. In MAC / FAC model, the structural similarity cannot be found at first and it is difficult to recognize the relations between different fields of technology. According to the previous study about chance discovery[10], a hidden relation provides new knowledge. Therefore, if we can find a relation between patents based on the structural similarity, we can regard a chance discovery inference in which creative technology strategy can be generated. In addition, clustering and classifying patents based on the structural similarity help to find the relations between them. Therefore, we build the hypothesis that it is possible to produce a creative strategy of technology to see the visualization of patents classified by the structural similarity of technological problems.

In this study, we proposed a method for clustering and classifying patents as a method to support the creation of technical development strategy, based on the structural similarity of texts expressing technological problems to be solved by invention. In addition, we proposed a method of visualization of patents to be able to grasp the contents of the technological problems of each patent in a cluster at a glance.

2 RELEVANT STUDY

2.1 Knowledge Representation by Predicate-argument Structure

In English-speaking countries, some researchers suggested an analogy method, which converts a sentence to a logical form by syntactic analysis, in the process of question answering and recognition of implicational relations [9][12]. The previous study[9] showed that predicate logic formation helps to pinpoint exact answers for questions and justify answers on a state-of-the-art question answering system. Furthermore, the other previous study[12] showed that predicate logic formation indicates robust inference and improve to recognize textual inferences by machine learning.

Inspired by such previous work, we proposed a method to support analogy for humans with the use of predicate-argument structures. Predicate-argument structures represent what arguments are related to a predicate, and forms a basic unit for expressing the meaning of a sentence. We express sentences of technological problems by the combination of important predicate-argument structures. In this study, we focus on a first-order predicate logic formation to understand technological problems easily. As a simple and human-friendly format, the predicate-argument structures are composed of verbs, nouns, and the cases of nouns. By looking at technological problems written by the predicate-argument structures and comparing the structures with each other, patent examiners can find the structural similarity between the problems. Therefore, we have the hypothesis that with the use of predicate-argument structures, the examiners can easily use analogy and understand how each patent document approaches technological problems.

2.2 Measurement of Structural Similarity between Documents

The previous study [2] evaluated the method to calculate the structural similarity between sentences consisting of two words by conversion to tuples from an entity-relation graph. It considers structures of documents as an entity-relation graph, in which vertices correspond to entities and edges correspond to lexical-syntactic patterns that represent semantic relations between entities. Then, the previous study proposed numerous kernel functions to measure the degree of analogy between two tuples. This method does not assume a particular relation representation. However, a dependency relation between words can help to increase the accuracy of measuring the structural similarity between documents. Therefore, we use a dependency relation between the predicates and the variables of predicate-argument structures in documents to measure the structural similarity between the documents.

To identify relationships between predicate-argument structures, we focus on the case grammar theory. The case grammar theory is a theory that deals with the essential predicate-argument structure. It describes the logical form of a sentence in terms of a predicate and a series of case-labeled arguments such as agent, object, and location[4]. In Japanese, the case is represented by case particles, such as "ga" (which means subjective case), "wo" (which means accusative case), and "ni" (which means objective case). In previous study, wide case frames are automatically constructed from the web corpus.[7] The concept of case frames is based on the hypothesis that a couple of a verb and its closest case is explicitly expressed on the surface of text, and can be considered to play an important role in sentence meanings. Therefore, the couples of verbs and their closest cases are aggregated for each usage of the verbs, and basic case frames are generated. Then, the basic case frames are clustered to merge similar case frames in a thesaurus and wide case frames are generated.

In this study, based on the concept of case frames, we treat the structural similarity of documents as the similarity of the combination of cases in the documents. In order to measure the structural similarity, we calculate a distance between documents as a distance between the case sets in the documents. In the case set, we regard a combination of a case particle and a noun which has been used in predicate-argument structure in documents as one element. The next chapter describes the details.

3 METHOD

In this study, we used unexamined Japanese patent applications, which were published from 2013 to 2015 and contain the word "condiment" in the content of "problems to be solved by the invention" as datasets to make clusters and visualize. The number of the documents was 185. The proposed method consists of four steps shown in Fig.2. Below let us describe details of each step.

3.1 Summarization of Technological Problem

We extract the technological problem from the content of "problems to be solved by the invention." We approach this problem by summarizing the content. The summary of the content is constructed by

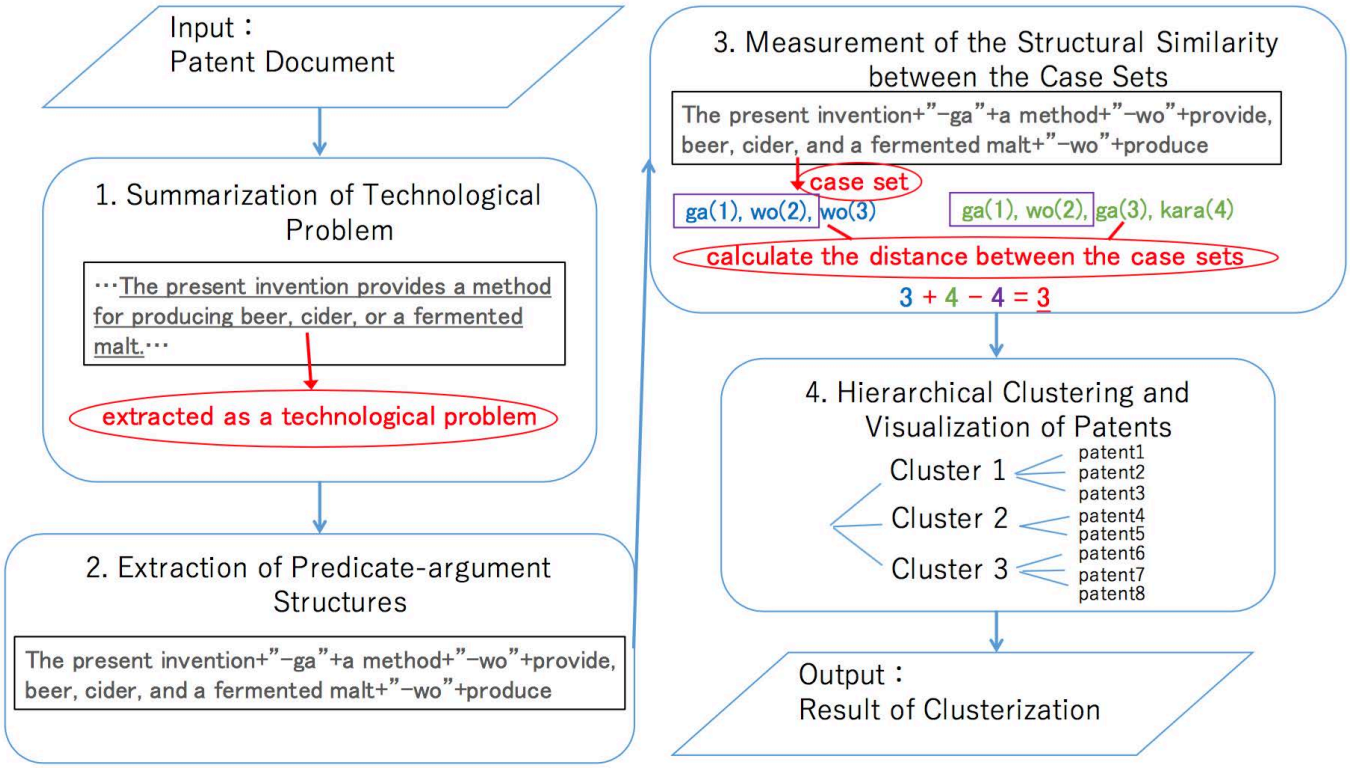

Figure 2. Flow chart of the proposed method

selecting the most important sentence from the content. In the previous study[11], the scoring method by the amount of information in the sentence is proposed to select the most important sentence. This method is based on the hypothesis that an earlier appearance of a word is more informative and scoring a sentence with a combination of appearance position and the frequency of words. The method to calculate the score of a sentence s, is described as Eq. (1).

$$score(s) = \frac{1}{||s||} \sum_i log freq(w_i) pos(w_i) \quad (1)$$

In Bag-of-Words model, the probability of a word w in a sentence s can be measured by its frequency as $freq(w)/||s||$, where $freq(w)$ indicates the frequency of w in s and $||s||$ indicates the total number of words in s. w indicates the i th word in s. $pos(w)$, which is the weight of the positions of w in a sentence s, is defined by the position of the appearances of word w in s. The word score $pos(w)$ is calculated by a geometric sequence, on the assumption that the score of every appearance of a word is the sum of the scores of all the following appearances of it. This sequence is defined as $f(w, s, i)$ in Eq.(2) for the i th appearance of word w in s.

$$f(w, s, i) = f(w, s, i+1) + f(w, s, i+2) + \cdots + f(w, s, n) \quad (2)$$

The content of "problems to be solved by the invention" represents the technological problem, starting from what previous methods could not solve, followed by details, concluding with the purpose of the patent. Important words tend to appear at the beginning and repeatedly in the content. Therefore, the scoring method is adequate to summarize the content. The sentence of the maximum $score(s)$ in Eq. (1) is here selected as the summary. The length of the summary is fixed as around 100 words to recognize at a glance.

3.2 Extraction of Predicate-argument Structures

The summary of a technological problem is expressed as the combination of predicate-argument structures. We use dependency parsing of the summary and extract predicate-argument structures. As a framework of the technological problem, predicate-argument structures are composed of nouns, case particles, and verbs. For morphological analysis, we use Juman[8] and for dependency parsing, we use KNP[7]. Both of Juman and KNP are appropriate for Japanese language. If consecutive nouns including prefix or suffix are contained in the sentence, we regard them as a whole word. Numbers, pronouns, and syncategorematic words are excluded from the extent of research object. We extract categorematic verbs as predicates. When the verbal auxiliary, "nai" (means adding negative) followed after the verb, we mention it behind the verb. We analyze the result of dependency parsing then define each verb written in root as a predicate, and relevant nouns as values of variables of the predicate.If a certain technological problem contains two predicate-argument structure A, B and both of them contains two variables, the technological problem is represented by this format below. Each word is connected by the symbol " + " and each predicate-argument structure is connected by the symbol " , ".

$noun_{1A} + case_{1A} + noun_{2A} + case_{2A} + verb_A,$
$noun_{1B} + case_{1B} + noun_{2B} + case_{2B} + verb_B$

3.3 Measurement of Structural Similarity between Case Sets

More than one predicate-argument structure must be included in a document. As the predicate-argument structure is the framework of the document, the distance between the documents assumes to be the

same as a total of the distances between their predicate-argument structures. Therefore, we calculate a distance between two patent documents as the distance between the combinations of cases in the documents. By this calculation, the distance between two documents can be determined according to their predicate-argument structures.

The case set is defined as a combination of case particles and variables in a sentence. If a certain sentence i contains n variables, the case set of the sentence set_i is represented as Eq.(3). Here, a variable in the predicate-argument structure val_n is digitized by an appearance order in the sentence. When the same noun appears more than once, the first appearance order is adopted. $case_n$ is defined as the type of case particles related to val_n in the predicate-argument structure. In Japanese, there are 9 case particles: *wo, ni, ga, de, no, yori, he, ya, kara, to*.

$$set_i = \{case_1(val_1), case_2(val_2), \cdots, case_n(val_n)\} \quad (3)$$

The distance of two documents i, j is calculated by the number of elements of the difference set of the case sets as follows:

$$dist_{i,j} = set_i \setminus set_j \quad (4)$$

Let us consider the distance of two Japanese sentences represented by the format of predicate-argument structures below.

1. A+“ga”+B+“ni”+C, A+“ga”+D+“wo”+E+“ni”+F, D+“ga”+G+“ni”+H+“wo”+I
2. J+“ga”+K+“ni”+L, J+“ga”+M+“kara”+N+“he”+O

In the first sentence, there are 3 predicate-argument structures and 8 variables. The first predicate-argument structure and the second predicate-argument structure contain the same variable “A”. The second predicate-argument structure and the third predicate-argument structure contain the same variable “D”. The case set of the first sentence is written as Eq.(5).

$$\{\text{“}ga''(1), \text{“}ni''(2), \text{“}ga''(1), \text{“}wo''(3), \text{“}ni''(4), \text{“}ga''(4), \text{“}ni''(5), \text{“}wo''(6)\} \quad (5)$$

In the second sentence, there are 2 predicate-argument structures and 5 variables. The first predicate-argument structure and the second predicate-argument structure contain the same valuable “J”. The case set of the second sentence is written as Eq.(6).

$$\{\text{“}ga''(1), \text{“}ni''(2), \text{“}ga''(1), \text{“}kara''(3), \text{“}he''(4)\} \quad (6)$$

Therefore, the difference set of these two case sets is calculated as Eq.(7) and the structural distance of the two sentences is calculated as 7.

$$\{\text{“}wo''(3), \text{“}ni''(4), \text{“}ga''(4), \text{“}ni''(5), \text{“}wo''(6), \text{“}kara''(3), \text{“}he''(4)\} \quad (7)$$

3.4 Hierarchical Clustering and Visualization of Patents

Hierarchical clustering by using average linkage is selected as a method of cluster analysis of technological problems of patents. The reason why hierarchical clustering is selected is that the relationship between clusters is visualized by it. Then, we propose a method of visualization of technological problems. The format of output data is JSON, which is a lightweight data-interchange format. The embedded structure of the data is composed of the number of the cluster, the technological problem solved by each patent document, and the patent number in sequence. In visualization, we use D3.js, which is a JavaScript library for visualizing data with JSON, HTML, and CSS.

4 QUESTIONNAIRE SURVEY

4.1 Questionnaire Hypothesis

We carried out a questionnaire survey to evaluate the validity of clustering and usefulness of visualization. We have three hypotheses as follows:

1. The distance between the case sets in documents reflects the structural distance between documents.
   In this study, we regard the structural distance between documents as the distance between the case sets in documents. To verify the validity of this hypothesis, we should confirm that the cluster is clustered based on the structural similarity.
2. Representing technological problems as predicate-argument structures helps to understand technological problems easily and accurately.
   Shown in the previous study[9], predicate-argument structures can help to find exact answers for questions. Therefore, participants can understand the structures of technological problems more easily by reading the documents formed with predicate-argument structures than by reading the plain texts. Also, the awareness of the causal structures of technological problems can be increased by the formation of predicate-argument structures.
3. Representing technological problems as predicate-argument structures helps to support analogy and to find common problems in different fields of technology.
   In relation to the second hypothesis, participants can find common problems in different fields of technology more easily by reading the documents formed with predicate-argument structures.

4.2 Questionnaire Details

To verify these hypotheses above, a questionnaire survey conducted to 18 participants in twenties in June, 2016. All relevant consent and human subject approvals were obtained for this experiment. We selected two clusters from the result of the cluster analysis and made a questionnaire form with each cluster. The questionnaire 1 was made with the cluster which contains three documents and the questionnaire 2 was made with the cluster which contains five documents. Half of all participants were given the questionnaire 1 and the other half were given the questionnaire 2. The reason why we selected different size of the clusters is that we have a hypothesis that it is more difficult to capture the structural similarity of documents if a number of documents increases and then people prefer to refer predicate-argument structures to answer the analogy question.

We presented the participants both the original documents and the result of the cluster analysis. Then, we asked 3 questions for the group of original documents and the result of the cluster analysis. The 3 questions are shown below.

Part 1. Comparing the two documents, as seen from the structure of each document, choose a word that corresponds to the word in one document from the words in the other. An example question is shown below.

Question: Comparing No.2015104322 with No.2015112075 in Fig.3, please choose a word that corresponds to the ”taste ” of No.2015104322 from the words in No.2015112075.

Choices: flavor, liquid condiment, denseness, smell, flavor, vegetable, rich, problem

The correct answer: flavor

This part contains two multiple choice questions. The correct answer is determined by the discussion among the author and two patent

experts engaged in the intellectual property department of the food company for more than three years.

Part 2. Please answer the technological problems which may be common in the cluster. (example answer: All of the sentences refers "a method of manufacturing a composition about the food" as a common issue.) If you cannot find common technical problems, please answer "nothing."
This part is free description type of the question. To make it easier to answer this part, an example answer is described in it.

Part 3. Which type of question is easier to answer, plain texts or predicate-argument structures?

The part 2 is a question to verify the hypothesis 1 and hypothesis 3. The part 1 and 3 are questions to verify the hypothesis 2.

5 RESULT AND DISCUSSION

5.1 Result of the Clusterization

Fig.3 shows the overall and one cluster of the result of the cluster analysis. The number of clusters was determined as 30. The number of elements in each cluster was in the range of 1 - 136. The number was inclined to one element and the number of clusters which contain only one element was 21. However, this tendency can be seen because clustering is based on the structural similarity of technological problems and many documents with different structures are included in datasets. Therefore, as mentioned at 4.1, the validity of the result is evaluated by the results of the questionnaire.

5.2 Result of the Questionnaire Survey

Table.1 shows the percentage of correct answers between plain texts and predicate-argument structures at Part 1. The percentage was calculated by the average of two questions.

Table 1. Percentage of correct answers between plain texts and predicate-argument structures

	PT	PS	Chi-square test
Questionnaire1	27.8	38.9	$p < 0.05$
Questionnaire2	50.0	77.8	$p < 0.05$

PT:plain texts, PS:predicate-argument structures

Shown in Table.1, the percentage of correct answers was increased by reading the documents formed with predicate-argument structures. In addition, the difference in the percentage of correct answers was significant in the chi-square test. (significant level $p < 0.05$) This indicates that the use of predicate-argument structures is effective for humans to understand the structure of documents.

Table.2 and Table.3 show common technological problems answered at Part 2 in Questionnaire 1 and Questionnaire 2.

Table 2. Common technological problems answered at Questionnaire 1, Part 2

	Common problem	Number
PT	improvement of taste	5
	property change of material	2
	production of condiment	1
	analysis of condiment	1
PS	improvement of taste	8
	property change of material	1

PT:plain texts, PS:predicate-argument structures

Table 3. Common technological problems answered at Questionnaire 2, Part 2

	Common problem	Number
PT	improvement of taste	3
	preservation method	2
	prevent deterioration of liquid	2
	solving problems of food additives	1
	production of condiment	1
PS	improvement of taste	6
	solving problems of food additives	2
	production of condiment	1

PT:plain texts, PS:predicate-argument structures

The technological problem of each patent in the clusters used in the questionnaires is an improvement in the taste of different condiments with different materials. Hence, the improvement of the taste can be considered as a common structure of technological problems. Shown in Table 2 and Table 3, the improvement of the taste was answered as a common technological problem in both of the questionnaires. This result suggests that the distance between the case sets in documents reflect the structural distance between documents. In addition, comparing plain texts with predicate-argument structures, the answers were less varied in predicate-argument structures. This indicates that the common structure of technological problems should be more easily grasped from documents formed by predicate-argument structures. Furthermore, comparing the two questionnaires, the answers were more varied in the questionnaire 2. This implies that the common structure of technological problems should be more easily grasped if the number of sentences for comparison is smaller.

Table.4 shows the percentage of selected answers at Part 3.

Table 4. Percentage of selected answers

	PT	PS	Chi-square test
Questionnaire1	55.6	44.4	non-significant
Questionnaire2	55.6	44.4	non-significant

PT:plain texts, PS:predicate-argument structures

Shown in Table.4, the readability was no difference between predicate-argument structures and plain texts. Furthermore, the preference of predicate-argument structure was no difference by the degree of the difficulty of the analogical problem. There are two reasons for this. First, some participants pointed out they didn't understand the meaning of the symbol " + " and " , ". It indicates that some participants couldn't grasp the structure of documents from predicate-argument structures. Because the questionnaire did not contain the explanation about how to form predicate-argument structures, the explanation should be added to it. Second, some participants said conjunctions in plain texts helped to understand the structure of documents. Therefore, the relations between predicate-argument structures should be represented in them.

6 CONCLUSION

This study proposes a novel method for text clustering. Compared with the previous method of frequent term based text clustering[1], extracting predicate-argument structures from patent documents and calculating the distance between patent documents based on case sets in the predicate-argument structures is a new approach. In the questionnaire survey, some participants found the same common problems from the cluster. This indicates that the proposed method can

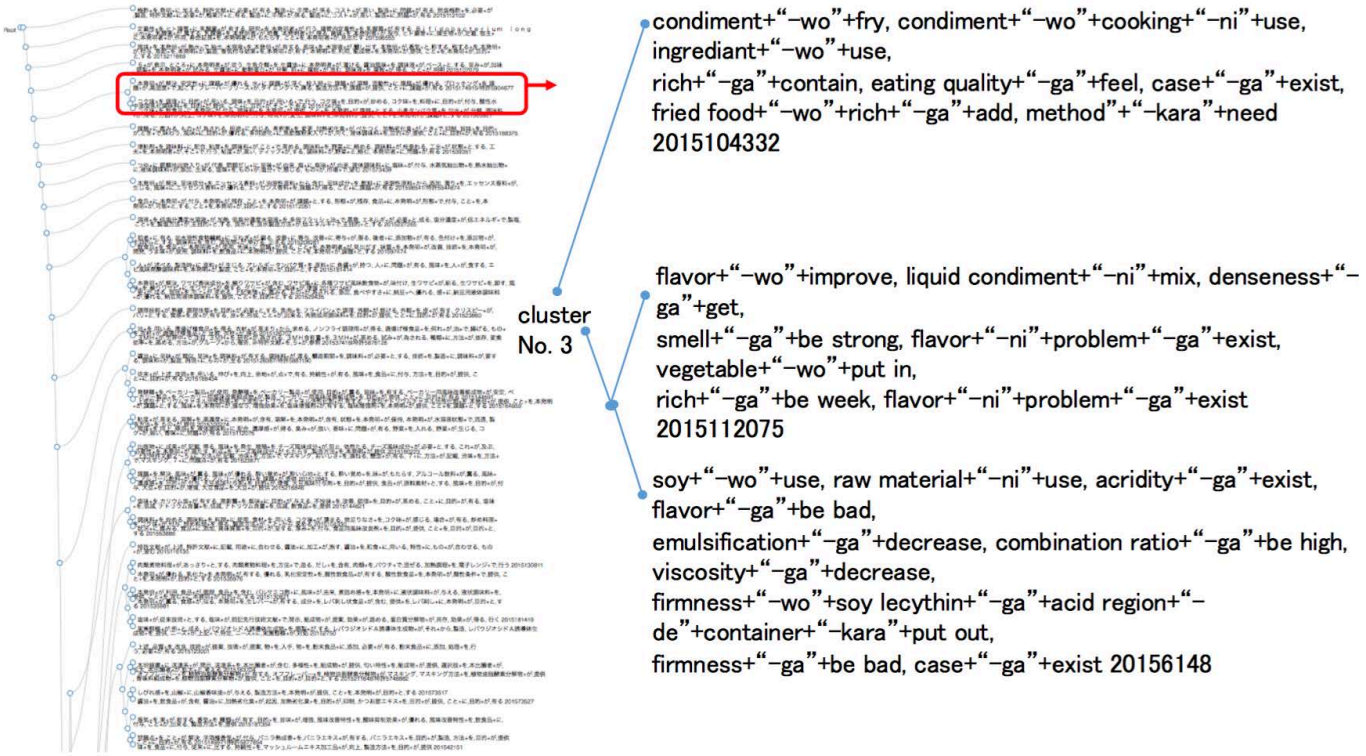

Figure 3. Overall and one cluster of the result of the cluster analysis

be a method of document clustering based on the structural similarity. However, relations between clauses are not reflected in the structural similarity of documents because this proposed method cannot consider relations between clauses such as subordinate clause and parallel clause. Therefore, an issue in the future is to design the distance between the documents, considering both relations between predicate-argument structures and relations between clauses.

Furthermore, the questionnaire survey also indicates that the percentage of correct answers can be increased by reading the documents formed by predicate-argument structures. Therefore, this visualization method is thought to be effective to support to search some patent documents relating to inventors' own technological problems. However, the survey also indicates that the readability is no difference between predicate-argument structures and plain texts. It is due to the lack of explanation of the questionnaire and there is a need for improvement of the questionnaire survey. Also, it is necessary to improve the format of predicate-argument structure easily to understand the structure for humans.

ACKNOWLEDGEMENTS

This work was supported by CREST, Japan Science and Technology Agency. This paper has been accomplished under collaborative research project with Toppan Forms, Tokyo, Japan.

REFERENCES

[1] F. Beil and M. Ester, 'Frequent term-based text clustering', *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.436–442, (2002).

[2] D. Bollegala, M. Kusumoto, Y. Yoshida, and Kawarabayashi K., 'Mining for analogous tuples from an entity-relation graph', *Proceedings of the 23rd international joint conference on Artificial Intelligence*, pp.2064–2077, (2013).

[3] B. Falkenheiner, K. Forbus, and D. Gentner, 'The structure mapping engine: Algorithm and examples', *Artificial Intelligence*, **41**, pp.1–63, (1989).

[4] C. Fillmore, *The Case of Case*, Universals in Linguistic Theory, New York, 1968.

[5] K. Forbus, D. Gentner, and K. Law, 'Mac/fac: A model of similarity-based retrieval', *Cognitive Science*, **19**, pp.141–205, (1994).

[6] D. Gentner, 'Structure-mapping: A theoretical framework for analogy', *Cognitive Science*, **7**, pp.155–170, (1983).

[7] D. Kawahara and S. Kurohashi, 'A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis', *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp.176–183, (2006).

[8] S. Kurohashi and M. Nagao, 'Japanese morphological analysis system jumanversion 3.61', *Proceedings of the 20th National Conference on American Association for Artificial Intelligence*, (1999).

[9] D. Moldovan, C. Christine, Sanda H., and Steve M., 'Cogex: A logic prover for question answering', *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.87–93, (2003).

[10] Y. Ohsawa and P. McBurney, *Chance Discovery*, Advanced Information Processing, Berlin: Springer, 2003.

[11] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, 'A study on position information in document summarization', *Proceedings of the 23rd International Conference on Conputational Linguistics*, pp.919–927, (2010).

[12] R. Raina, A. Ng, and C. Manning, 'Robust textual inference via learning and abductive reasoning', *Proceedings of the 20th National Conference on American Association for Artificial Intelligence*, pp.1099–1105, (2005).

Chance Curation in Virtual Cognitive Niches

Selene Arfini, Tommaso Bertolotti, and Lorenzo Magnani [1]

Abstract. In this paper we consider chance curation (the task of easing chance-discovery activities for users) as far as it concerns information sharing in online communities, understood as virtual cognitive niches. Virtual cognitive niches can be considered as digitally-encoded collaborative distributions of information and pieces of knowledge into the environment. The scope of chance curation within these contexts depends on the quality of the information externalized, the aptness of the information-sharing mechanisms to the agents' purposes and the control they can implement over the system. Online communities, as socially biased networks, provide more ways to connect the users to each other than to control the quality of the information they share and receive. We contend that this social bias shapes chance curation into the discrimination between what Nagy and Neff called "imagined affordances" (the combination of users' perceptions, attitudes and expectations over the functionality of a particular technology) and what we can call "critical chance", which are event-related data that conceal a particularly good opportunity or a particularly dreadful risk for the agent and her surroundings.

Introduction

The notion of curation in the framework of chance discovery was first introduced by Abe in 2010 [1], where he reviewed various types of curation as *display strategies* in particular contexts, such as exhibitions, galleries, archives and museums. The task of promoting and enabling the availability of certain products of artwork and artifacts to appropriate audiences was the key to comprehend the aided connection between users and the displaying framework as a performance of chance discovery[2]. He also proposed an interesting list of features for curation in chance discovery:

- Curation is a task to offer users opportunities to discover chances.
- Curation should be conducted with considering implicit and potential possibilities.
- Chances should not be explicitly displayed to users.
- However, such chances should be rather easily discovered and arranged according to the user's interests and situations.
- There should be a certain freedom for user to arrange chances. [1, p. 797]

Abe's definition is particularly interesting as far as online communities such as social networking websites are concerned because they are engineered not only as to be "fool proof," but to naturally co-opt the inferential patterns developed by human beings in settings of real-life cognition (and hence chance discovery), for instance social-cognition and one's natural disposition towards sharing [25].

In order to better spell out our vision, this paper will be divided in three parts, in which we will examine aspects of chance curation performed in online communities. In the first section, we will frame online communities as virtual cognitive niches: for this to make sense, we will briefly explain the notion of cognitive niche also by relying on the concept of *affordance* that, although not properly belonging to the chance-discovery paradigm, has already fruitfully interacted with the latter. In the second part, we will then focus on the particular forms of explicit and implicit communication performed in virtual cognitive niches, which consent the execution of different chance curation strategies with respect to "concrete" cognitive niches: the focus on the virtual domain, the docility-based relation with truth and the social virtues of fallacies. In the third part we will discuss about the consequences of the use (and abuse) of those chance curation strategies in order to enrich the activity of sharing diverse types of information in online communities.

We will claim that, if chance-curation strategies contribute to the implementation of social and epistemological affordances and chances, they can also lead to a controversial, and maybe unexpected, two-sided outcome. On the bright side, they foster social activities between members of enormous communities and contribute to the fine-tuning of information sharing so that it can, for instance, assist responders in case of emergencies, defeating provincial and national boundaries [8]. In this sense, it promotes the propagation of what Nagy called "imagined affordances", which are the combination of users' perceptions, attitudes and expectations over the functionality of a particular technology [16]. On the dark side, we claim that it also promotes biased epistemological judgments over the information the users receive and share, inasmuch as the communication of data is adjusted on the interests and motivations of the singular users. This manipulates the understanding of the quality of the information the agent has at her disposal and can indeed lead to the gain of what we have called "critical chances", which are event-related data that conceal a particularly good opportunity or a particularly dreadful risk for the agent and her surroundings.

In order to begin our analysis, let us start by displaying particular strategies of chance curation performed in both cognitive niche and virtual niche framework.

1 Introducing Cognitive Niches

Niche theories are a cluster of more or less interrelating approaches bridging biology, cognitive science and philosophy, exploring the relationship between agents and their environment. Originated in biol-

[1] Department of Philosophy, Education and Economical-Quantitative Sciences, University of Chieti and Pescara, Italy. Department of Humanities and Computational Philosophy Laboratory, University of Pavia, Italy. Emails: selene.arfini@gmail.com (corresponding author), bertolotti@unipv.it, lmagnani@unipv.it.

[2] Indeed, as clearly shown by Ohsawa and Fukuda [19], one of the essential goal for chance discovery studies is to understand the emergence and availability of hidden stimuli towards new chance embedded in communication and displaying strategies (such as visualization techniques).

ogy in the early XX century, niche theories would stress the functional notion of *niche* to explain *how* a species occupied its environment in opposition to the geographical notion of *habitat* [22]. The niche *constructivist* approach [17] went further, claiming that organisms actively modify their environment in ways that affect the local selective pressure, to the point of establishing an ecological inheritance system.

Cognitive niche theories originated in the "philosophical sector" of cognitive science, to stress how human beings' relationship with their environment was essentially information-based, as their success depended mostly on elevated cognitive capabilities [26, 21]. Andy Clark's constructivist take on cognitive niche stressed the local dimension, akin to the biological one: cognitive niches are *constructed* by human actors by externalizing knowledge into the surrounding environment [9, pp. 256–257].

Clark exemplifies this by telling of a bartender in a busy bar who, upon receiving orders from tables, arranges glasses by shape, and adds decorations such as straws and cocktail umbrellas, *in order to better remember the next drinks she has to mix without having continuously looking at the order list.* It is not the human species's cognitive niche concerned here, it is the bartender's own niche that she is herself constructing.

For our purpose it is also interesting to underline two particular features of cognitive niches, which have been distinctively provided by the initiators of cognitive niches theories, Pinker, Tooby and DeVore [21, 26], who describe cognitive niches as a prerogative of the human species as a cognitively proficient species. According to Tooby and DeVore, the cognitive niche is that in which human beings apply an instrumental intelligence in order to uncover and exploit, in a persistent way, cause-effect models of the external world [26]. Specifically, since the human cognitive system is "knowledge or information driven", they highlight the role of the cognitive niche as the environment in which the employment of those cause-effect models of the world represent guides for prejudging which courses of action will lead to which results. Pinker suggested how human beings' primary reliance on information and knowledge in the cognitive niche made them "informavore" [21]. With this term he highlighted how gathering and exchanging information is the substantial activity that sustains and modifies the welfare of cognitive niches.

The description of cognitive niches as structures distributing information and knowledge which strongly situate human cognition and decision making within their environment, has already been successfully connected to the framework of chance-discovery [13, 2, 6]. Chance curation, as the activity of offering users the opportunity to discover chances – understood as events with a "significant impact on a human's decision making" [18, 20] – has an environmental (eco-cognitive) dimension and can be rightly seen as a part of cognitive niche construction, or at least as strictly interrelated with the latter [15]. This kind of tasks can be conceived as as safeguarding the discovery and exploitation of cause-effect models of the external world in order to guarantee agents to gather and distribute information into the cognitive niches, and can also improve the richness of a cognitive niche. This is of particular importance if we consider the epistemological and cognitive role of chance curation in the particular framework of virtual cognitive niches, which can be generally described as collaborative distributions of information and pieces of knowledge into the environment by means of digital encoding.

The activity of cognitive niche construction reveals something important about human and animal cognitive systems. One of the main tenets of this approach is that humans do not retain in their memory an explicit and complete representation of the environment and its variables, but they actively manipulate it by picking up information and resources upon occasion. As already argued ([14]), chances – understood as events with a "significant impact on a human's decision making" – are data, or clusters of data, bearing a strong affinity with the concept of *affordance*, introduced within Gibson's ecological psychology ([11]): it is thus possible to rely on such concept in order to better understand the human part of chance discovery.

Gibson defined "affordance" as what the environment offers, provides, or furnishes. For instance, a chair affords an opportunity for sitting, air breathing, water swimming, stairs climbing, and so on. Gibson did not only provide clear examples, but also a list of definitions that may contribute to generating possible misunderstanding: 1) affordances are opportunities for action; 2) affordances are the values and meaning of things which can be directly perceived; 3) affordances are ecological facts; 4) affordances imply the mutuality of perceiver and environment.

It is important to stress that the notion of *chance* and that of *affordance* are not mutually interchangeable. While it could be said that all chances – as relevant for one's decision making (and hence one's behaviour) – are affordances, conversely not all affordances rise to the level of chances. It is nevertheless possible to elaborate on a shared characterisation of affordances and chances, in their setting a relationship between an agent, her knowledge, and her environment.

Indeed, chance-discovery and chance-curation could embody the natural follow-up to affordance theory: chance-discovery and curation is indeed about the discovery/construction, via a human-computer interaction and through effective procedure of data analysis and crystallisation, of new complex affordances, offering unforeseen possibilities for decision making and action.

2 Virtual Cognitive Niches and their Domains

The development of new informational environments through digital technology can be framed through niche constructing theory. Seeing cognitive niches construction as a human prerogative, cognitive niches theories permit to analyze the specific traits that have established the human ecological and evolutionary success.[3] Constructing *virtual* cognitive niches is, indeed, one of the most interesting ecological dynamical behavior that our species alone has shown. The virtualization of niches starts from the creation of meta-environments through the employment of computers and the Internet, which are able to interact with our natural environment (or to simulate it), but that can be manipulated in a much easier way, depending on the distribution of information that is permitted [6].

The virtual cognitive niches created through digital technologies go beyond traditional ecologies, their ontologies and what they can afford. They are the extension of cognitive niches, through an informatization of the ecological space. In other words, the human ability of gathering and exchanging information and knowledge from the environment, the ability to alter the environment so that it better serves cognitive scopes, is amplified in the framework of virtual cognitive niches. Indeed, in virtual cognitive niches, tasks of chance curation can affect users with an implicit larger range, because instances of knowledge distribution represent the sole acts of ecologi-

[3] Whether cognitive niches are a human prerogative is a debated topic. Clark himself, in his definition, refers to "animals" but his examples concern only human beings. Bertolotti and Magnani argue for the possibility of overlap between low-level cognitive niches and advanced ecological ones [7] What is beyond argument, though, is the fact that human beings *master* cognitive niche construction.

cal and cognitive importance. In virtual cognitive niches there is no gap between information and matter. Matter is reduced to coding, and the only "spatial requirement" is the memory available to host the coding. According to Clark, cognitive niches are structures built by animals in order to transform problem spaces "in ways that aid (or sometimes impede) thinking and reasoning about some target domain or domains" [9, 256–257]. In this case, virtual cognitive niches do not only have a proper target domain in the information contained in the digital reality, but also afford problem solving in the ecological reality to which the digital niche refers.

Specifically, one category of virtual cognitive niches is of particular relevance for the investigation of chance-curation tasks: online communities, such as social networking websites, newsgroups, online chat rooms, forums, and so on.[4] They are online-based platforms where individuals interact, either through anonymous avatars or actual profiles with a networks of connections, sharing personal information and contents. These are cognitive niches inasmuch as they provide users with ways to gather and exchange information relevant for their decision-making – otherwise said, they can be seen as chance repositories. Indeed, online communities as virtual cognitive niches modify the social pressure of the environment through the employment of forms of explicit and implicit communication performed in the online world. This situation calls for different chance curation strategies with respect to "material" cognitive niches [15].

2.1 Chance Curation Strategies in Online Communities

One of the first chance curation strategies that is fostered in online communities as virtual cognitive niches is what we can call the "focus on the virtual domain", which can be investigated through the implementation of Clark's constructivist take on cognitive niches. Clark describes cognitive niches as the structures that are built by animals in order to transform problem spaces "in ways that aid (or sometimes impede) thinking and reasoning about some target domain or domains" [9, 256–257]. In the case of online communities as virtual cognitive niches the user's thinking and reasoning affected are directed to two main target domains. On the one hand, there is the *virtual domain*, which is structured on the online platform and include its objects and tools, the virtual personas of the users, and the information shared, usually as "posts". On the other hand, there is the *actual external domain*, which includes the actual agents using the online platform and the contents of the posts shared in online communities. Indeed, we should specify that the information embedded in posts does not always belong just to the virtual domain (that is they are not only users' comments about some other user or notes about some item on the network's page), but they also refer to the external reality of the actual world. Indeed, one of the most relevant feature of contemporary online network is the extended possibility of sharing information and data regarding news, political events, scientific discoveries, and so on. Moreover, these information and data refer to the external reality which also encompasses online community as debatable objects: so, in an online network such as Twitter we can find a post that links to an online journal's opinion column regarding the usability and usefulness of Twitter itself. The contents of that post does not belong to the virtual domain of Twitter even if it is shared on its platform. Thus, the two domains are diverse even if they are intertwined in the users' perception.

One chance curation strategy that is fostered by the structure of online community is orienting the focus of the users on the virtual domain. The virtual domain is indeed loaded with cognitive artifacts that implement the communication and the sociability of the users sharing a particular network (as two-people and group chatrooms, more or less public personal pages and profiles, group selection sharing, and so on). These tools contribute to generate what are generally called *imagined communities* [4, 5], which are the projection of actual communities in first-person point of view of each users. In this sense, every information regarding other users are interpreted by the first person perspective of the user according to the beliefs she has regarding the information she shared first.[5]. Therefore, the focus on the virtual domain regarding the social-oriented information about the users is maintained through the projection of the imagined community and through tools (as friendly-user chat, standard format for news feeds and profile) that help to keep in mind that the virtual personas that are present on the social network site are not utterly equivalent to the external and actual agents who use the platform. They are their virtual versions, constructed on the base of the virtual profile structure, who communicate through the use of the specific tools of the platform. The actual people who share the personal information on online communities, as a physical and multidimensional agents (who do not have the possibility of spreading her opinions and thoughts at a vast public in actual reality), are different from their virtual version. The focus on the online domain permits to the agent to keep in mind the difference between the two and adopt an adequate behavior in the cognitive virtual chance, where the formalities and the hierarchical levels depend on the digital connection between people and the communication is not always and only directed to the interested target (for example in a shared public post).

Moreover, while it is obvious that the virtual domain of online communities encourages the distribution of social-oriented information (personal information of the users, interests, goals and preference for interaction), information regarding the external domain can be exchanged in order to support, change and improve the quality of the communication on the online platform. These type of information can be oriented to trigger social mechanisms of the platform (as political opinions on the base of personal experience and character, which ask for feedback and discussion) or based entirely on events, data and facts of public interest, which belong to the external reality domain. These are considered interesting by the user and shared as texts, images or links to the relative pages. Also in this occasion, the structure of online communities fosters a focus on the virtual context. The virtual niches become places where users can share interesting facts and items which help her to display what her interests and opinions are. In addiction, all the pieces of contents are displayed with the same format and style, making easy to have a quick picture of what the users would like to speak about or comment on the online platform. The focus on the virtual domain consents to elaborate new communication strategies and relationships between users of the platform, making more interactive the imagined communities they perceive.

The focus on the virtual domain, fostered by the structure of the networks, suggests also to the users to build a docility-based relation with truth and to employ some useful fallacies, which embed social virtues on the framework of online communities.

Docility is a concept introduced by Herbert Simon [25] that describes the human agent tendency to lean on what other people say.

[4] Curation in online communities is already a topic that raises interest in business applications, especially for information delivery. For a review of different types of curation developed in business and information market, cf. [3])

[5] For instance, if the user has been honest in declaring her birthdate, she will be prone to think that also other users have been honest with the same respect [4]

The disposition is specifically related to the performance of problem-solving activities conducted on the base of social channels' suggestions. Relying on aids and resources provided by their fellows, the human agents have a major cognitive advantage: they can trust other people and so have at disposal chances that, first of all, they have never personally experienced, and, secondly, that are already available to be picked up.[6] Of course, in a concrete and world-base situation, trust is not informatively empty: one decides to trust another person, because she has reasons to do so. The agent gathers a number of clues in order to consider a particular source of information (a person, for instance) as trustworthy or not. But in an online community, trust can be a more difficult matter.

One the one hand, in an online community the shared information are not neutral – as impersonal or dispassionate: every user chooses what to share and when on the base of her interests, her desires and the effects she hopes to achieve through that particular sharing within the online community. On the other hand, every information is bound to the user who shared it: every piece of data, personal or community related, is presented in the platform because a user uploaded it and she is accountable for it. On a platform like Facebook, where the information are personally identified, this does not imply the trustworthiness of the information (the user could share it for all kinds of epistemically wrong reasons), but the trustworthiness of the social connection between the information and the virtual persona. The virtual user as vehicle for information, is a truth vector between a data and the adequacy of that particular data on her profile. This way, the users can build an online community that can provide social-based chances, using a docility-based relation with truth.

This leads the agent also to adopt a more kind perspective on the use of recognizable fallacies. Indeed, the notion of docility in online communities [25], can explains how human beings use epistemological and social resources offered by the platforms but also the tendency to fall into ecologically organized fallacious argumentations. This follows directly from the weaken structure of belief states in an eco-cognitive dimension affected by docility. As suggested by Gabbay and Woods [10], for example, there is a "doxastic irresistibility" induced by the diffusion of well spread "say so". They suggest that a docility-oriented system drives to the application of a "ad ignorantiam rule" which describes the agent's passive acceptance of information unless she has reason to retain in doing so. They wrote:

> Human agents tend to accept without challenge the utterances and arguments of others except where they know or think they know or suspect that something is amiss [10, p. 27].

This reflects the tendency of the human agent to economizes the cognitive efforts in response to a free-given flux of information. Another ecologically well-fit reaction to a docility-based environment is the application of the *ad verecundiam* fallacy: the agent accepts her sources' assurances because she is justified in thinking that the source has good reasons for them (the fallacy would be the failure to note that the source does not have good reasons for his assurances). These tendencies, which are dramatically dangerous in a scientific or political domain, are at the base of the online community interactions: in a framework where there are no impersonal communications, the validity of a comment or note is judged on the base of the trustability of the person in the particular network. In this sense, *ad verecundiam* and *ad ignorantiam*, even if are forms of fallacious reasoning, stand also for the cognitive legitimation of a space of free discussions, where trust and responsibility are weighted on the users' online accountability. They build the connections of a socially-driven system and, as such, they can be regarded as chance-curation mechanisms in this perspective.

So far we displayed how some chance curation mechanisms enacted in online communities as virtual cognitive niches help users to perceive a more interactive and honest imagined community out of the digital platform. These strategies contribute to a better distribution of information and knowledge (which refers to both domains, the virtual and the actual and external) both in terms of chances and affordances. At the same time, in rich virtual cognitive niches as online communities can be, we contend that chance curation strategies enacted by digital programmers also produce the generation of unexpected consequences related to the interaction of the users with the enhanced possibilities offered in the more complex system. First of all, programmers and designers perform chance curation by pushing the agents to elaborate the chances at their disposal through feedback processes. Secondly, but more importantly, they can offer users opportunities that the programmers did not expect to emerge, both useful and critical to the welfare of the niche. In order to speak about this problematic, in the next section, we will speak about the generation of these not foreseen possibilities in terms of "imagined affordances" [16] and "critical chances", considering the particular case of online communities communications and information-sharing during crises, which highly demands group interventions and so strong actions in rich cognitive niches.

3 Social Media and Crisis Management: Examples of Guided Chance Curation

Crises, whether natural or human-induced, cause a strong demand for chances. During a terror attack, a major incident or riot, a flood, a fire, an earthquake and so on, decision-making processes need to be quick and as much accurate as possible. Citizens need to know where they can take shelter, which areas are safe and which are to be avoided. The government needs to know as much as possible about the emergency in action in order to decide where to allocate relief personnel or police forces in case of an attack. Evidences, for instance of photographic nature, are chances. Obtaining some evidence is an event that affects decision-making, usually for better (but also for worse, in case of a false evidence).

Information posted by users over social networks and microblogging websites during a crisis is likely to include evidence that can be used either by other citizens or by the government in order to adapt their actions to the emergency. Otherwise said, social networks and microblogs become rich repositories of chances during crises: chances that just need to be discovered and exploited. The presence of a chance, even coupled the certitude that a chance is present within a certain space (be it physical or digital), does not automatically entail the exploitation of such chance. For instance, there might not be time enough to situate a potentially game-chancing chance, and the decision makers might have to rely on a weaker chance if it is easier to locate and exploit. At the same time, highly problematic situations can lead the agents to use certain objects, tools and devices at their disposal differently from usual, discovering and exploiting new chances. In rich virtual cognitive niches as online communities, this could lead to a group sharing of this discovery, implying a bottom-up modification of the digital resources and the exploitation of what Nagy and Neff called "imagined affordances" [16]. Nagy and Neff wrote:

[6] That is one of the most important assets describing cognitive economy, that is, the need to reach a sort of trade-off between the accuracy of a decision and the limited time one is bounded to.

> It can help scholars think through the way that affordances are formed in interaction between users, designers, and the physical and digital materiality of technologies. To solve this, we develop the concept of imagined affordance. Imagined affordances emerge between users' perceptions, attitudes, and expectations; between the materiality and functionality of technologies; and between the intentions and perceptions of designers [16].

Imagined affordances have been conceived in order to explain the interaction between users' social context, abilities, and purposes with technologies. On the one hand, they are the results of a productive interaction between designers and programmers' top-down manipulations of the structures of technologies and the users' bottom-up feedback activities (as use, misuse and tentative actions) on them. On the other hand, they are the implementation of "users' perceptions, attitudes, and expectations" within the possibilities and boundaries of a given technology. In order to make an example, we can speak about the process that leads to chance curation in crisis relief, which usually takes the form of enriching posts with tags and hashtags. *Tags* are additional information embedded (usually) to a picture, adding specification about the time and place where it was shot, on its subjects and if relevant on who took it. Tagging, for instance in Facebook, might directly link the post with the content of the tag (e.g. the other user's profile or, in case of a location, to other pictures coming from the same location and further information about it). *Hashtagging*, on the other hand, became originally widespread on Twitter and was later spread to other social networking websites such as Facebook: it consists in marking a post with a tag preceded by a hash symbol (#), in order to highlight its belonging to a specific topic or conversation. The action of hashtagging became widespread for a bottom-up intervention of the users on the functionality of Facebook posts: it is an imagined affordance that contributed to apply diverse mechanisms of users' enacted chance-curation strategies on online communities during crises.

Notwithstanding the fact that strategies of chance-curation aim at improving the chance distribution of a particular niche, offering tools and resources in order to enhance the niche richness (with also the implementation of imagined affordances), they can also lead to the development of what we can call "critical chances". A "critical chance" is a chance that conceal a particularly good opportunity or a particularly dreadful risk.[7] It also is the consequence of the further elaboration of chances by the users, who invest in a particular niche their expectations and interests. The dreadful consequences of a critical chance can also endanger the welfare of the niche, where some possibilities can quickly become dangerous for some users. One instance of this phenomenon in virtual cognitive niches can be traced in the 2011 Vancouver riots. Following a Hockey match, the city of Vancouver was invested on June the 15th by an unseen wave of hooliganism, vandalism and looting. Citizens reported on social media in order to support crisis responders, not only by posting images, but by tagging where was happening what and encouraging users to tag whoever they managed to recognize among the rioters. This was a great opportunity for the police forces and the citizen had explored a chance which led to a particularly good result. Rizza and her colleagues [24] provide a thorough analysis of the phenomenon. Unfortunately, while initially Vancouver Police Department asked for citizens' help in identifying the rioters, the situation soon took a grimmer outcome as the grass-roots identification process set the stage for a do-it-yourself justice. The activity of curation, carried out by enhancing posts, misfired mainly because of "unverifiable quality" of the media and "unpreparedness" of the institutions although they had requested the curation in the first place. This led to the emergence of a critical chance with a particularly bad outcome, which drove to a case of "unintended Do-It-Yourself Justice", supported by an unclear approval of a "Do-It-Yourself Society" [24, p. 52]. In spite of the partial societal failure, though, the chance curation activity was successful at letting emerge a series of chances (for restoring order) which might have gone unexploited by lack of information. To describe the failures in terms of chance curation, the activity might have pushed the intended chances towards unintended recipients who were nevertheless able to act upon them.For instance, Facebook was invented as a tool for keeping in touch with friends and acquaintances in a situation of high-and-far mobility such as the one characterizing the contemporary US. There, people attend higher education in places that are not their hometowns, and then move on to their professional careers in yet different locations. Facebook would afford answers to "What have you been up to these last few years?" or "Where did you go on holiday?". But when emergencies took place, users realized that Facebook and other social networks, developed for other scopes, could afford answering questions such as "Are you alright?", "What is happening?" (faster than traditional media), "Where should we go right now?". This was the relevant affordance imagined for distressful situations.

Concerning hashtags, part of the curation process involves making hashtags as informative and less ambiguous as possible. This is particularly challenging in the phase when hashtags emerge spontaneously, and are not enforced from some authority or authoritative group.[8]

> Such ad hoc creation of hashtags contributes considerably to Twitter's ability to respond speedily and effectively to major breaking news and other acute events, of course: within minutes of major events [. . .], relevant hashtags had emerged and began to carry the latest mainstream news stories, first-hand updates from affected locals, and commentary from the wider Twitter user community. This emergency is by no means always linear and unproblematic, of course – competing hashtags including #Oslo, #Osloexpl, and #Oslobomb carried news of the bombing in downtown Oslo, for example –, but in most cases, a gradual convergence of conversations into no more than a small handful of hashtags can be observed; standard network effects (which mean that the hashtags with the largest number of participants also contain the greatest volume and best quality of information) tend to apply. Additionally, key messages are often made visible to all the followers of competing hashtags by including both those hashtags in the same tweet (or by users manually adding other hashtags as they retweet the original message) [8, p. 7].

[7] We drew the name from the role-play game lexicon, where a roll of the dice could lead to a critical Success or a critical Failure. Generally, when throwing a twenty-faces die, the faces between 17 and 20 represents a critical success – the game master determines what happens, but it is always something good. The higher the roll, the better it gets for you. Instead, the faces between 1-4 represent a critical failure: the game master determines what happens but it is always something bad – the lower the roll, the worse the result. [12]

[8] A case of superimposed hashtag for emergency relief is situated in the June 2016 floods in Northern France. An association of online volunteers, *Volontaires Internationaux en Soutien Opérationnel Virtuel*, asked users to take pictures of flood-related emergencies (not of the flooding rivers themselves) and to post them online with the hashtag #VISOV so that they could aggregate and check them.

Spontaneous hashtags on social media appeared also to indicate resources-as-chances during crises, and not only to circulate evidence: the November 2015 terror attacks in central Paris left hundred of people stranded and unable to return to their homes in the middle of the night. In a grass-roots emergency response, many Parisians volunteered to host affected people. Social networks where the ideal setting for signaling this availability, but the chance had to be curated in order to facilitate recognition, and the #portesouvertes (open doors) quickly circulated. Chance curation, always relating to crises response, can also be "superimposed" or "guided." Examples can be drawn from the 2015 terror attacks that shook France, first in January and the already mentioned ones in November. In both cases, a collaborative navigation app, Waze, was put under scrutiny when authorities asked not to signal roadblocks and police cars as these informations might be used by terrorists on the run to avoid apprehension.

This reliance on Social Networks to enhance crisis response is a fair example of imagined affordance and the diffusion of so-called "critical chances". They are ways of perceiving possibilities in a certain artifact which were not intended by its developers.

Last but not least, the Facebook Safety Check tool[9] can be framed in this analysis of chance curation as crisis response in online communities: people rely on Facebook, a dominant social network in most of the world, as provider of chances to know whether dear ones (or mere acquaintances) are hurt in case of major incidents. Usually people are expected to state they are alright, either spontaneously or after being prompted publicly or privately by someone. Facebook developers curated this chance by introducing the Safety Check. In case of major mishap in an area where the user had been previously localized, Facebook asks the user to confirm she is alright, and then publicly reports that she logged herself as safe: interestingly, this can be seen as the adoption, by the developers, of an imagined affordances into the set of intended affordances of an artifact.[10]

4 Conclusion

In this paper we have analyzed the activity of chance-curation performed in rich virtual cognitive niches, considering the particular case of online communities. First, we have presented a comprehensive notion of virtual niches, considering the literature on cognitive niche and niche construction. Then, we have presented the interesting feature of online communities as rich repositories of affordances as chances, that can be offered through top-down manipulations of the environment performed by the designers and programmers, or bottom-up elaborations of the platform users. In particular, we have discussed the double-domain feature of virtual cognitive niches, the docility-based relation between users and the social virtues displayed in this context by the use of some fallacies. Considering chance-curation activities as both the top-down manipulation of the environment in order to offer chance to the users, and the users feedback to the referred framework in order to exploit some particularly useful chances, we have explored the case of chance-manipulation operated in online communities in order to respond to a crisis. This consented us to analyze the peculiar phenomenon of the exploitation of chances, the programmers didn't expect to emerge, both as imagined affordances and critical chances.

[9] https://www.facebook.com/about/safetycheck/

[10] The Facebook Safety Check and other social media technologies for crisis response will be analyzed in depth in a forthcoming research project coordinated by Caroline Rizza [23].

REFERENCES

[1] A. Abe, 'Curation in chance discovery', *2010 IEEE International Conference on Data Mining Workshops*, 793–799, (2010).

[2] A. Abe, 'Cognitive chance discovery: from abduction to affordance', in *Philosophy and Cognitive Science*, eds., Magnani L. and Li P., 155–172, Springer, Verlag, (2012).

[3] A. Abe, 'Data mining in the age of curation', *IEEE 12th International Conference on Data Mining Workshop*, 273–279, (2012).

[4] A. Acquisti and R. Gross, 'Imagined communities: awareness, information sharing and privacy on facebook', *Privacy Enhancing Technologies*, **4258**, 36–58, (2006).

[5] B. Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Revised edn. Verso, London and New York, 1991.

[6] T. Bertolotti and L. Magnani, 'The role of cognitive niches in mediating knowledge, entropy and violence', *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 954–959, (2013).

[7] T. Bertolotti and L. Magnani, 'Theoretical considerations on cognitive niche construction', (2016). Forthcoming.

[8] A. Bruns, 'Ad hoc innovation by users of social networks : the case of twitter', in *ZSI Discussion Paper*, ed., Soziale Innovation Centre for Social Innovation, 1–13, Vienna, Austria, (2012).

[9] A. Clark, 'World, niche and super-niche: How language makes minds matter more', *Theoria*, **54**, 255 – 268, (2005).

[10] D. M. Gabbay and J. Woods, *The Reach of Abduction: Insight and Trial*, volume 1 of *A Practical Logic of Cognitive Systems*, Elsevier, Amsterdam: North Holland, 2005.

[11] J. J. Gibson, 'The theory of affordances', in *Perceiving, Acting and Knowing*, eds., R. E. Shaw and J. Bransford, Lawrence Erlbaum Associates, Hillsdale, JN, (1977).

[12] A. Hackard and S. Jackson. GURPS lite an introduction to roleplaying. http://www.sjgames.com/gurps/lite/. Accessed: 2016-06-16.

[13] L. Magnani, 'Chance discovery and the disembodiment of mind', in *Proceedings of the Workshop on Chance Discovery: from Data Interaction to Scenario Creation, International Conference on Machine Learning (ICML 2005)*, eds., R. Oehlmann, A. Abe, and Y. Ohsawa, pp. 53–59, (2005).

[14] L. Magnani, 'Creating chances through cognitive niche construction', in *Knowledge-Based Intelligent Information and Engineering Systems*, eds., B. Apolloni, R. J. Howlett, and J. Lakhmi, 917 – 925, Springer, Berlin Heidelberg, (2007).

[15] L. Magnani and T. Bertolotti, 'Selecting chance curation strategies: Is chance curation related to the richness of a cognitive niche?', *International Journal of Knowledge and System Science*, **4**(1), 50 – 61, (2013).

[16] P. Nagy and G. Neff, 'Imagined affordance: Reconstructing a keyword for communication theory', *Social Media + Society*, 1–9, (2015).

[17] F. J. Odling-Smee and M. W. Feldman K. N. Laland, *Niche Construction. The Neglected Process in Evolution*, Princeton University Press, Princeton, 2003.

[18] Y. Ohsawa, 'Introduction to chance discovery', *Journal of Contingencies and Crisis Management*, **10**(2), 61–62, (2002).

[19] Y. Ohsawa and H. Fukuda, 'Chance discovery by stimulated groups of people', *Journal of Contingencies and Crisis Management*, **10**(3), 129 – 138, (2002).

[20] Y. Ohsawa and P. McBurney, *Chance Discovery*, Springer, Verlag, 2003.

[21] S. Pinker, 'Language as an adaptation to the cognitive niche', in *Language Evolution*, eds., M. H. Christiansen and S. Kirby, 16–37, Oxford University Press, Oxford, (2003).

[22] A. Pocheville, 'The ecological niche: History and recent controversies', in *Handbook of Evolutionary Thinking in the Sciences*, eds., T. Heams, P. Huneman, G. Lecointre, and M. Silberstein, Springer, (2015).

[23] C. Rizza. OSMOSIS – Open and Smart Management of vOlunteerS and vIctims in disaster Situation. ANR Project, 2016. Forthcoming.

[24] C. Rizza, A. G. Pereira, and P. Curvelo, '"Do-it-Yourself Justice": Considerations of Social Media use in a crisis situation: The case of the 2011 Vancouver riots', *International Journal of Information Systems for Crisis Response and Management*, **6**(4), 42–59, (2014).

[25] H. Simon, 'Altruism and economics', *The American Economic Review*, **83**(2), 156–161, (1993).

[26] J. Tooby and I. DeVore, 'The reconstruction of hominid behavioral evolution through strategic modeling', in *Primate Models of Hominid Behavior*, ed., W. G. Kinzey, Suny Press, Albany, (1987).

Automatic Identification of Trigger Factors: a Possibility for Chance Discovery

Marharyta Aleksandrova [1 2] and **Armelle Brun** [1] and **Oleg Chertov** [2] and **Anne Boyer** [1]

Abstract. Pattern identification in datasets has been the focus of many research works. Data mining, through association rules mining, is one of the best known approaches. In this paper, we introduce a new pattern, referred to as "set of contrasting rules". Contrary to most of the patterns from the state-of-the-art, this pattern has the characteristic of being made up of a set of rules. It has also the advantage of structuring rules into sets. One main originality of this pattern is that it allows to easily identify trigger factors: factors that can bring some event state changes (chances in terms of chance discovery). In real applications, this pattern can thus be used to influence the values of some attributes or to impact the human decision-making process, what is shown through the experiments conducted on a real dataset of census data.

1 INTRODUCTION

We live in the era of information society, where the manipulation of information is extremely important in all spheres of life: politics, economics, education, cultural activities, etc. [25]. Nowadays it is impossible to imagine that any important decision either at the enterprise or at the governmental level will be taken without the conduction of a prior analysis. For instance, enterprises often analyse the behavior of their customers and the general market tendencies for planning the type and the amount of goods to produce; for governmental planning, census data is an indispensable source of information. Often, even when taking personal decision, prior investigations can be done. For example, when choosing a movie to watch, one may look through the review comments.

One of the core bases of the information society is the availability of various data, used for information production. It is known that the amount of data grows very fast. According to the research of International Data Corporation (IDC), the number of digital bytes produced by the humanity doubles every 2 years [15]. This data comes from various sources. We can mention sensors in smart houses or in the industry, commercial and personal photos/videos, site logs, behavioral information concerning certain accounts associated with real users, specially collected data (like poll results or demographic data), text messages or comments, etc. Electronic devices capable to capture many activities (like change of geographical position, web-search on specific terms, organizer notes) become our daily life companions; this allows to collect personal data. It is obvious that such a wide variety of data is a precious source of useful information. However, in order to mine this information from "raw" data (not pre-processed data), the usage of appropriate tools or methods is required.

In many applications, it is important to identify trigger factors: those factors that explain or that can bring some events or system state changes. Identification of trigger factors can be of particular interest for chance discovery. Indeed in socio-demographic applications the trigger factors can be considered as those, that can influence the human decision-making, what corresponds to the definition of a *chance* in chance discovery: "a chance means to understand an unnoticed event/situation which can be (uncertain, but) significant for making a decision" [20]. Also trigger factors can be used to influence the appearance of rare or novel events, what is also a subject of the chance discovery theory. In general, researches from different domains like medicine [23, 13], sustainable entrepreneurship [7], finance [4], etc. are interested in the identification of such trigger factors, mainly through data analysis.

There exists a wide variety of data analysis methods. Data mining techniques are of particular interest, as they were designed to discover interesting and unpredictable patterns and interconnections [11], contrary to statistical methods, that are mainly used to check the validity of predefined hypotheses [12]. One of the very popular data mining techniques is association rules mining, where an association rule is a pattern of the form "if X then Y". Association rules were originally proposed to identify associations and dependencies within elements of a dataset. They were also used to identify cause-effect relationships in many applications: occupational accidents construction industry [5], traffic safety problems [9], or to detect sick and healthy conditions in males and females [19]. So, we can consider exploiting association rules mining to identify factors that explain or that trigger others (trigger factors). In the above rule, an occurrence of X explains/triggers an occurence of Y.

The contribution of this paper lies in the introduction of a new type of a pattern, referred to as "set of contrasting rules". The original aspects of a "set of contrasting rules" pattern is that it is made up of a set of rules. "Set of contrasting rules" patterns have several advantages: 1) by identifying and grouping highly informative rules, *i.e.* that highlight differences between groups of elements in the dataset, they allow to structure the huge set of association rules; 2) as these patterns highlight the differences between data groups, through the introduction of the notion of varying and invariant attributes; these differences can be interpreted as trigger factors, *i.e.* factors that can influence the value of some attributes, or the move of elements of the dataset between groups. To the best of our knowledge, no approach in the literature allows to identify such trigger factors, directly from the patterns or from rules.

The rest of the paper is organized as follows. Section 2 presents benchmark on the association rules discovery. Section 3 introduces

[1] Université de Lorraine - LORIA, Campus Scientifique, 54506 Vandoeuvre les Nancy, France, email: {marharyta.aleksandrova,armelle.brun,anne.boyer}@loria.fr

[2] National Technical University of Ukraine Kyiv Polytechnic Institute, 37, Prospect Peremohy, 03056, Kyiv, Ukraine, email: chertov@i.ua

"set of contrasting rules" pattern: its definition, the way it is mined and the way it is used. Then, Section 4 focuses on the experiments and shows that the proposed pattern can be used for the identification of factors, capable to influence human decision-making. Finally, we discuss and conclude our work in Section 5.

2 RELATED WORKS ON ASSOCIATION RULES MINING

The problem of association rules mining was introduced by Agrawal et al. in [2], with the aim of mining transactional databases of basket data and finding the associations of items that are bought together. Since then, association rules have been successfully used in a wide area of applications starting from medicine [21] to the analysis of students enrollment data [1] and building recommender systems [28].

Let D be a dataset defined on a set of n attributes $\left\{A^1, A^2, \ldots, A^n\right\}$. For each attribute A^j there is a set of possible values. We will refer to this set as the domain of the attribute A^j, denoted by $domain\left(A^j\right)$. The attributes can be either categorical or continuous. For example, considering census data, the attribute *education* is categorical with $domain\left(education\_level\right) = \{school, college, bachelor\_degree, master\_degree, PhD\}$ and the attribute *age* is continuous with $domain(age) = [0, 100]$. The domain of *age* can be discretized on intervals, such as [0,20],]20,60],]60,100], resulting in $domain\left(age\right) = \{[0, 20],]20, 60],]60, 100]\}$. Assume that all elements in the dataset D are organized into k mutually exclusive groups $G_1, G_2, \ldots, G_k$ with $G_1 \cup G_2 \cup \ldots \cup G_k = D$ and $G_i \cap G_j = \emptyset, \forall i \neq j$. Let the group of an element (its class) be specified by the value of the target attribute A^G with possible values $g_1, g_2, \ldots, g_k$.

We will call each pair $\{attribute, value\}$ an item. A set X of items is called an itemset. By $supp_D\left(X\right)$ we denote the support of the itemset X in the dataset D. The support value is calculated using the following formula $supp_D\left(X\right) = \frac{count_D(X)}{|D|}$, where $count_D\left(X\right)$ is the number of elements in D containing X and $|D|$ is the number of elements in D.

An association rule is an induction rule of the form $X \rightarrow Y$, where X and Y are itemsets and $X \cap Y = \emptyset$. X is the left-hand side (LHS) of the rule, also called the antecedent and Y is its right-hand side (RHS), or consequent. The support of the rule $X \rightarrow Y$ in D is calculated as $supp_D\left(X \rightarrow Y\right) = supp_D\left(X \cup Y\right)$ and its confidence as $conf_D\left(X \rightarrow Y\right) = \frac{supp_D(X \cup Y)}{supp_D(X)}$.

An example of an association rule R in a basket market dataset can be the following R : "$milk$ & $bread \rightarrow butter$", $supp_D(R) = 20\%$, $conf_D(R) = 90\%$. This rule means that 90% of people, who bought milk and bread also bought butter within the same transaction; and that products $milk$, $bread$ and $butter$ were found together in 20% of the transactions (elements) in the database.

Agrawal and Srikant proposed the first algorithm for mining association rules: *Apriori* [3]. Although a variety of modifications (Parallel Apriori [26], high-dimension oriented Apriori [16], etc.) and other techniques were proposed for mining association rules (see, for example [22, 27, 8, 18]), the *Apriori* algorithm remains highly used ([24, 10]). The main criteria used by these algorithms to filter rules are the user-specified support and confidence thresholds.

One of the major shortcomings of the various association rules mining algorithms is the large amount of rules produced, which are redundant and that have to be, afterwards, analysed by experts to identify the interesting and non-obvious ones [17]. Association rules can be also used for the identification of trigger factors (we can say that the antecedent triggers the consequent of the rule). However it should be done manually, though the analysis of the rules. To the best of our knowledge there are no specialized techniques designed for the automatic identification of trigger factors.

3 A NEW PATTERN "SET OF CONTRASTING RULES"

In this section we present the new pattern called "set of contrasting rules". This pattern is specially designed to contain trigger factors, which can influence the move of elements from one group to another. The "set of contrasting rules" pattern is made up of multiple rules. In general case, these rules do not allow to identify trigger factors individually, but only within the group.

In this section we introduce the definition of a new pattern, present the algorithm for mining it and show how this pattern can be used to identify trigger factors and chances.

3.1 Definitions

The definition of a "set of contrasting rules" pattern relies on the introduction of two new types of attributes: varying attributes and invariant attributes. An attribute is considered to be varying if its value can be changed externally to the system within the specified application task, and invariant otherwise. For example, when analyzing census data the attribute $income\_level$ can be considered as varying if, for instance, the government can provide citizens with financial assistance. At the opposite the value of the parameter $ancestry$ can not be changed, that is it belongs to the set of invariant attributes. Thereby we divide the set of all attributes (except A^G) into two subsets: the set of varying attributes and the set of invariant attributes.

Definition 1. For a specified parameter α ($\alpha > 0.5$), a set of rules $R_1, R_2, \ldots, R_k$ is called a set of α-contrasting rules if:

1. $conf(R_1) \geq \alpha$ & $conf(R_2) \geq \alpha$ & $\ldots$ & $conf(R_k) \geq \alpha$;
2. in each rule, the itemsets of the consequents are formed of only one attribute: the target attribute A^G, with different values of A^G in the set;
3. the antecedents of the rules are made up of the same attributes, within which there is at least one varying and one invariant attribute;
4. the values of all invariant attributes are the same for all rules;
5. at least one varying attribute has different values in the set of rules.

Let us consider the simple case where there are only two groups G_1 and G_2 defined on the dataset D. Thus, $domain\left(A^G\right) = \{g_1, g_2\}$. Let us analyse an example of a rule pair $R1$ and $R2$ presented in Figure 1.

1. both rules are highly confident, with the minimum confidence value $min\_conf = 0.7$ (> 0.5);
2. the consequent of the rules contains only one item: the target attribute. The value of the consequent A^G is different for $R1$ and $R2$;
3. the antecedent of the rules is composed of the same attributes, among which one is invariant ($A^{inv,1}$) and three are varying ($A^{var,1}$, $A^{var,2}$ and $A^{var,3}$);
4. the values of the invariant attribute $A^{inv,1}$ are the same for both rules ($A^{inv,1} = a_1^{inv,1}$), as well as the values of one of the varying attributes ($A^{var,1} = a_1^{var,1}$ for both $R1$ and $R2$);
5. the values of the two other varying attributes are different ($a_{10}^{var,2}$ and $a_7^{var,3}$ for the rule $R1$ and $a_5^{var,2}$ and $a_3^{var,3}$ for the rule $R2$);

Thereby, we can conclude that the pair of rules $R1$ and $R2$ form a pattern "pair of α-contrasting rules with $\alpha = 0.7$".

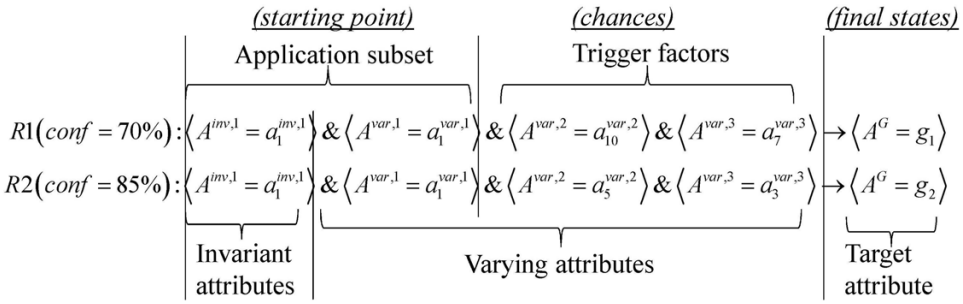

Figure 1. Example: Pair of contrasting rules

3.2 Algorithm for Mining Sets of Contrasting Rules

We now focus on the way the contrast pattern introduced can be mined.

In our work we choose to rely on the traditional Apriori algorithm and then post-process its results. The procedure that identifies sets of contrasting rules from the set L is described in Algorithm 1. As an input the algorithm receives the list of candidate rules L, and it outputs the set of sets of contrasting rules. First, highly confident rules ($min\_conf = \alpha$, see condition 1 of Definition 1) are mined using the Apriori algorithm. Second, the rules with a consequent made up of the target attribute A^G are selected to form the set of candidate rules L. Third, through a pairwise comparison of the rules in L, the sets of contrasting rules, satisfying the conditions specified in Definition 1, are found. In Algorithm 1 the function $isPairOfContrastingRules(R1, R2)$ is a boolean function, that returns $true$ if all the conditions of Definition 1 are fulfilled by the pair $R1$ and $R2$, and $false$ otherwise.

Data: L - list of discovered rules
Result: $setOfSets$ - set of sets of contrasting rules
$setOfSets = \{\}$;
for $R \in L$ **do**
  $contrSet = \{R\}$;
  $l = L - \{R\}$;
  for $r \in l$ **do**
    if *isPairOfContrastingRules(R,r)* **then**
      $contrSet = contrSet \cup \{r\}$;
    end
  end
  if $size\,(contrSet) \geq 1$ **then**
    $setOfSets = setOfSets \cup \{contrSet\}$;
  end
  $L = L - contrSet$;
end

Algorithm 1: Mining sets of contrasting rules

3.3 Identification of Trigger Factors (Chances) with Sets of Contrasting Rules

In the beginning of Section 3 we claimed that the proposed pattern "set of contrasting rules" can be used to identify trigger factors, *i.e.* factors which can influence the move of elements of the dataset from one group to another. Let us consider the pair of contrasting rules in Figure 1. Analysing these two rules, we can say that if, for the elements having $A^{inv,1} = a_1^{inv,1}$ and $A^{var,1} = a_1^{var,1}$, we force the attributes $A^{var,2}$ and $A^{var,3}$ to change their values from $a_5^{var,2}$ and $a_3^{var,3}$ to $a_{10}^{var,2}$ and $a_7^{var,3}$ respectively, then with a probability of 70%, these elements will move from the group G_2 to G_1. The move in the inverse direction will occur with a probability of 85%. Thereby, the varying attributes with different values in the pair of contrasting rules define the *trigger factors*: they can influence the move of the elements from one group to another. The invariant attributes and those varying attributes having the same values in the pair of contrasting rules specify the *application subset* that is the subset of elements, that can be influenced by these trigger factors.

Now let us analyse these 2 rules from the view point of the chance discovery theory (see [20]). Chance discovery aims at identifying multiple scenarios which have an intersection point and different final states. The intersection point (which can be hidden or unobvious) is called a *chance* and its utility is measured as the difference of the merits of final states [20]. Depending on the application task we can consider attributes which define application subset ($A^{inv,1}$ and $A^{var,1}$ in Figure 1) as a starting point of possible scenarios with final states given by the target attribute (A^G). Trigger factors in this case can be viewed as chances.

In real applications, the proposed pattern can be used to solve a wide range of tasks depending on the underlying meaning of the attributes. An example of using this pattern for the identification of factors that can influence the appearance of a certain event (a birth of the baby in our experiments) is presented in the next section.

4 EXPERIMENTAL RESULTS

The goal of the experiments conducted here is to show to what extent "set of contrasting rules" patterns allow to automatically discover highly interesting knowledge, more precisely trigger factors. The experiments are conducted on a real dataset of census data.

4.1 Dataset

The experiments are conducted on a microfile with a 5-percent sample of the California census dataset [14]. This dataset contains records of 610,369 family households (we choose to ignore subfamilies, as the number of households with subfamilies corresponds to only 3.6% of the initial sample), on more than 100 attributes.

4.2 Problem Formulation and Data Pre-processing

The goal of the conducted experiments is to *identify which factors can influence the human desire to give birth to a baby*. Thereby, the

target attribute (the attribute which we are focusing on) is related to the children in the households.

Not all of more than 100 attributes of the dataset can actually provide a significant influence on the desire to give birth to a baby. For example, it is obvious that the attribute $household\_language$, that indicates the language spoken in the family, has no influence on the families' desire to have a child. Traditional association rule mining algorithms do automatically discard statistically insignificant attributes. However, for the sake of simplicity we manually filter out the attributes. We are aware that the resulting list of chosen attributes may not be exhaustive and that some other attributes may influence the desire to have a baby. Nevertheless, as the goal of this research paper is to test the proposed pattern, but not to conduct a sociological study, we restrict the number of attributes used.

The list of the considered attributes contains the 12 following items. Their possible values, as well as the type of each attribute (invariant or varying) are given in Table 1.

- home ownership ($HouseOwn$),
- type of building ($HouseType$),
- number of vehicles available ($Vehicle$),
- spouse age (2 attributes: $HAge$ and $WAge$),
- spouse education (2 attributes: $HEdu$ and $WEdu$),
- spouse ancestry (2 attributes: $HAnc$ and $WAnc$),
- spouse class of worker (2 attributes: $HWorkClass$ and $WWorkClass$),
- husband's total income in 1999 ($HIncome$),

Table 1. Possible values of the attributes and their type; p =10,000$.

ATTRIBUTE	TYPE	DOMAIN
$HouseOwn$	var	yes / no
$HouseType$	var	NoStatHome /Apart / Att (attached house) /Det (detached house)
$Vehicle$	var	0 / 1 / 2 / 3 / ≥ 4
$(H/W)Age$	inv	young / middle-young / middle / middle-old / old
$(H/W)Edu$	inv	noSchool / school / noCollege / college / associate / bachelor / master / doctor
$(H/W)Anc$	inv	WestEurope / EastEurope / Mexico / Latino / CentralAmericaIslands / NorthAfricaAndSouthAsia / otherAfrica / otherAsia / Australia / Pasific Afro-American / OtherAmerica / NonDef
$(H/W)WorkClass$	inv	NoWork / PrivWork / GovWork / SelfEmployed
$HIncome$	var	$]-\infty, 0]$ / $]0, 1p]$ / $]1p, 2p]$ / $]2p, 4p]$ / $]4p, 6p]$ / $]6p, 8p]$ / $]8p, 10p]$ / $]10p, 20p]$ / $]20p, 30p]$ / $]30p, 40p]$ / $[40p, \infty[$
$Child$	target	YES / NO

In order to find an answer to the question that formulates our experimental goal, we divide the dataset into two contrasting groups G_1 and G_2. The first group G_1 is made up of families (elements) with one or two children aged from 0 to 2 years. The second group G_2 contains families without any children. Such restrictions on the children age are imposed in order to track the change in family state from a childless family to a family with a small child (or children). We do not consider families with elder children, as it is difficult to identify which factors triggered the child appearance some years back. Thereby, we add to the dataset the target attribute $Child$ (A^G), indicating the presence or not of small children in the family, with $domain\,(Child) = \{YES, NO\}$ (see last line of Table 1). The dataset is thus made up of 13 attributes.

To increase the reliability of the results obtained, we choose to impose some additional restrictions:

- all the families must be complete: the presence of both spouse is mandatory;
- both husband and wife must be without disabilities;
- spouse age must be within the most favorable period for having babies.

These conditions are quite relevant and obvious. For instance, it is clear that illness of the potential parents affects significantly their willingness and ability to have children.

The most favorable age bounds for having babies are 24 to 38 for men and 22 to 37 for women. Considering all the imposed above restrictions, the size of the dataset is reduced. The number of elements with $Child = YES$ and $Child = NO$ in the resulting dataset equals to 8,299 and 12,249 elements respectively. We form 5 age intervals for both husband and wife in the family. Further we will use following abbreviations: y for young, My for middle-young, m for middle, Mo for middle-old and o for old. The bounds of the most favorable age for giving birth to babies and the used age intervals (presented in the Table 2) are taken from our previous works related to this dataset [6].

Table 2. Age intervals for husband and wife

	husband	wife
young (y)	24-27	22-25
middle-young (My)	28-29	26-27
middle (m)	30-31	28-30
middle-old (Mo)	32-34	31-32
old (o)	35-38	33-37

If we divide the original dataset into sub-datasets according to the age of husband and wife, we get 25 sub-datasets. For example, we form the sub-dataset $D^{y,m}$ that corresponds to the families with a young husband (first position in the subscript of $D^{y,m}$) and middle-aged wives (second position in the subscript of $D^{y,m}$). In order to identify different patterns specific to a certain age, we conducted our analysis on these 25 sub-datasets separately.

4.3 Analyzing Sets of Contrasting Rules

We use the Apriori algorithm to mine association rules in the sub-datasets of D, with minimum confidence value equal to 0.7. After that the resulting rules are analysed by the Algorithm 1 with the goal to identify sets of contrasting rules.

Now we proceed to show what kind of information can be obtained with the help of the proposed pattern. As an example, in the Table 3 we present 3 sets of contrasting rules for 5 different sub-datasets $D^{y,y}$, $D^{My,My}$, $D^{m,m}$, $D^{Mo,Mo}$ and $D^{o,o}$ that correspond to families with husband and wife belonging to the same age group.

The antecedents of the rules in the patterns are represented in the second and third columns of Table 3. As stated in Definition 1, the antecedents of the set of contrasting rules have a common part that

Table 3. Some chosen "sets of contrasting rules" patterns, p =10,000$

$subD$	Antecedent		Consequent	Conf
	Subgroup	Trigger Factors	Child=	
$D^{y,y}$	**WEdu=school** & HIncome=]1p,2p]	Vehicle=1	YES	0.71
		Vehicle=0	NO	0.70
	WEdu=noCollege & HouseOwn=no	HIncome=]2p,4p] & HouseType=Det	YES	0.73
		HIncome=]0,1p] & HouseType=Apart	NO	0.70
	HAnc=Mexico & HouseType=Det &...	HouseOwn=yes	YES	0.75
	...& HIncome=]4p,6p]	HouseOwn=no	NO	0.82
$D^{My,My}$	**WEdu=associate** & **HAnc=WestEurope** &...	HouseType=Det	YES	0.71
	...& **WAnc=WestEurope** & HIncome=]4p,6p]	HouseType=Apart	NO	0.92
	WEdu=noCollege & **HAnc=WestEurope** & ...	HIncome=]4p,6p]	YES	0.71
	...& **HWorkClass=PrivWorke** & HouseType=Det	HIncome=]2p,4p]	NO	0.75
	WEdu=college & **HEdu=noCollege**	HouseOwn=yes	YES	0.70
		HouseOwn=no	NO	0.81
$D^{m,m}$	**HAnc=Mexico** & **WAnc=Mexico** &...	HouseType=Det	YES	0.71
	...& **WWorkClass=PrivWork**	HouseType=Att	NO	0.89
	HAnc=OtherAmerican	HIncome=]2p,4p]	YES	0.72
		HIncome=]1p,2p]	NO	0.86
	WAnc=Latino	HouseOwn=yes & HouseType=Det	YES	0.71
		HouseOwn=no & HouseType=Apart	NO	0.75
$D^{Mo,Mo}$	**WEdu=master**	HouseOwn=yes & HouseType=Det	YES	0.76
		HouseOwn=no & HouseType=Apart	NO	0.73
	WEdu=college & **HWorkClass=PrivWork**	Vehicle=2 & HouseOwn=yes	YES	0.72
		Vehicle=1 & HouseOwn=no	NO	0.78
	WEdu=associate & **WAnc=Mexico**	HouseOwn=yes	YES	0.70
		HouseOwn=no	NO	0.73
$D^{o,o}$	**HEdu=associate** & **WEdu=bachelor** &...	HouseOwn=yes	YES	0.71
	...& HWorkClass=PrivWork & HouseType=Det	HouseOwn=no	NO	0.88
	WEdu=associate	Vehicle≥4 & HouseType=Det	YES	0.70
		Vehicle=2 & HouseType=Apart	NO	0.73
	WEdu=bachelor & **HAnc=other Asia** & ...	HIncome=]10p,20p]	YES	0.70
	...& HouseOwn=yes	HIncome=]8p,10p]	NO	0.70

specifies the subgroup of the elements (starting points of scenarios); it is presented in the second column. In bold the invariant attributes are given. The antecedents have another part that differs in the values of varying attributes (this part represents the trigger factors or chances); it is presented in the third column of the table. The fourth column indicates the value of the consequent or final state (the attribute $Child$) of each rule in our patterns, and the fifth column reveals the confidence values of the corresponding rules.

When analysing the sets of contrasting rules given in Table 3, we can note that they correspond to very precise recommendations for specific subgroups of elements in the dataset. For example, let us look on the first pattern obtained for the sub-dataset $D^{y,y}$. It indicates that if we provide young families ($HAge =]24, 27]$ and $WAge =]22, 25]$) in which the wife's education level is $WEdu = school$ and husband's income is in the range $]10000, 20000]$ with a vehicle, then with a high probability (71%) they will decide to have a baby. However, in another subgroup of the same sub-dataset, which is composed of families where the wife has started the college but did not finish education there ($WEdu = noCollege$) and that do not have their own house ($HouseOwn = no$), it is not the number of vehicles that can trigger a child birth, but rather the combination of the type of the house (it should be changed from 'apartment' to 'detached house') and the increase of the income level. Also, we can see that subgroups and trigger factors (or combinations of attributes that form the trigger factors) are meaningfully different for different sub-datasets. This proves the ability of the proposed pattern to extract valuable knowledge from the dataset.

From the application point of view, the trigger factors discovered from the census data can be considered as factors that can influence human decision-making process in a certain direction. Indeed, the recommendations formed from the discovered patterns show how it is possible to increase the birth-rate, that is how it is possible to influence the decision on whether to give birth to a child. It means that discovered trigger factors can be considered as *chances* following the definition in [20]. Thereby we consider that the proposed approach can be a promising contribution to the chance discovery theory.

5 CONCLUSION

In this paper, we introduced a new pattern "set of contrasting rules". This pattern has the characteristic of being made up of several rules and is designed with the aim to directly identify specific knowledge: trigger factors and corresponding application subsets, through the introduction of the notions of invariant and varying attributes. A trigger factor is a factor that can bring some event state changes (in our case it moves elements from one group of the dataset to another). On the application level, this pattern can be interpreted as a way to search the levers of influence, which can force or trigger this move. This charac-

teristic is the key point of the proposed pattern, and according to our knowledge, there are no other patterns in the literature, dedicated to this goal. We showed that the proposed pattern can be considered as a tool for chance discovery with application subgroups corresponding to the starting points of the possible scenarios, the trigger factors to chances and the group affiliation to different final states.

The experiments conducted on a real dataset of census data demonstrate that the trigger factors can be actually identified, and that they can be easily interpreted and used to reach the desired objective. The objective of our experiments was to influence the birth rate of families. The analysis of the identified trigger factors shows that they can be considered as those that can influence the appearance of an event in the interest of the researcher.

REFERENCES

[1] Zailani Abdullah, Tutut Herawan, Noraziah Ahmad, and Mustafa Mat Deris, 'Extracting highly positive association rules from students enrollment data', *Procedia-Social and Behavioral Sciences*, **28**, 107–111, (2011).

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami, 'Mining association rules between sets of items in large databases', *ACM SIGMOD Record*, **22**(2), 207–216, (1993).

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al., 'Fast algorithms for mining association rules', in *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pp. 487–499, (1994).

[4] H Kent Baker, E Theodore Veit, and Gary E Powell, 'Factors influencing dividend policy decisions of nasdaq firms', *Financial Review*, **36**(3), 19–38, (2001).

[5] Ching-Wu Cheng, Chen-Chung Lin, and Sou-Sen Leu, 'Use of association rules to explore cause–effect relationships in occupational accidents in the taiwan construction industry', *Safety Science*, **48**(4), 436–444, (2010).

[6] Oleg Chertov and Marharyta Aleksandrova, 'Fuzzy clustering with prototype extraction for census data analysis', in *Soft Computing: State of the Art Theory and Novel Applications*, 289–313, Springer, (2013).

[7] Progress Choongo, Elco Van Burg, Leo J Paas, and Enno Masurel, 'Factors influencing the identification of sustainable opportunities by smes: Empirical evidence from zambia', *Sustainability*, **8**(1), 81, (2016).

[8] Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon, 'Rapid association rule mining', in *Proceedings of the tenth international conference on Information and knowledge management*, pp. 474–481. ACM, (2001).

[9] Karolien Geurts, Geert Wets, Tom Brijs, and Koen Vanhoof, 'Profiling of high-frequency accident locations by use of association rules', *Transportation Research Record: Journal of the Transportation Research Board*, 123–130, (2003).

[10] Zhenhai Guo, Dezhong Chi, Jie Wu, and Wenyu Zhang, 'A new wind speed forecasting strategy based on the chaotic time series modelling technique and the apriori algorithm', *Energy Conversion and Management*, **84**, 140–151, (2014).

[11] Jiawei Han, Micheline Kamber, and Jian Pei, *Data mining: concepts and techniques*, Elsevier, 2011.

[12] David J Hand, 'Statistics and data mining: intersecting disciplines', *ACM SIGKDD Explorations Newsletter*, **1**(1), 16–19, (1999).

[13] Anders Hougaard, Faisal Amin, Anne Werner Hauge, Messoud Ashina, and Jes Olesen, 'Provocation of migraine with aura using natural trigger factors', *Neurology*, **80**(5), 428–431, (2013).

[14] "u.s. census 2000. 5-percent public use microdata sample files". http://www.census.gov/Press-Release/www/2003/PUMS5.html, 2000.

[15] Discover the digital universe of opportunities: rich data and the increasing value of the internet of things. http://www.emc.com/leadership/digital-universe/index.htm. Accessed: 2016-03-07.

[16] Lei Ji, Baowen Zhang, and Jianhua Li, 'A new improvement on apriori algorithm', in *Computational Intelligence and Security, 2006 International Conference on*, volume 1, pp. 840–844. IEEE, (2006).

[17] Sotiris Kotsiantis and Dimitris Kanellopoulos, 'Association rules mining: A recent overview', *GESTS International Transactions on Computer Science and Engineering*, **32**(1), 71–82, (2006).

[18] MA Lei, 'Association rules mining algorithm based on matrix', in *International Conference on Advances in Mechanical Engineering and Industrial Informatics (AMEII 2015)*, pp. 974–980, (2015).

[19] Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen, 'Association rule mining to detect factors which contribute to heart disease in males and females', *Expert Systems with Applications*, **40**(4), 1086–1093, (2013).

[20] Yukio Ohsawa, *Chance discovery: The current states of art*, Springer, 2006.

[21] Carlos Ordonez, Norberto Ezquerra, and Cesar A Santana, 'Constraining and summarizing association rules in medical data', *Knowledge and Information Systems*, **9**(3), 1–2, (2006).

[22] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal, 'Efficient mining of association rules using closed itemset lattices', *Information systems*, **24**(1), 25–46, (1999).

[23] Jorunn Sundgot-Borgen, 'Risk and trigger factors for the development of eating disorders in female elite athletes', *Medicine & Science in Sports & Exercise*, **26**(4), 414–419, (1994).

[24] Miao Wang, Li Chen, Yanjun Huang, Lei Zhang, Zihao Zhang, Jie Ding, and Huiliang Shang, 'The application characteristics of traditional chinese medical science treatment on vertigo based on data mining apriori algorithm', *International Journal of Wireless and Mobile Computing*, **9**(4), 349–354, (2015).

[25] Frank Webster, *Theories of the information society*, Routledge, 2014.

[26] Yanbin Ye and Chia-Chu Chiang, 'A parallel apriori algorithm for frequent itemsets mining', in *Software Engineering Research, Management and Applications, 2006. Fourth International Conference on*, pp. 87–94. IEEE, (2006).

[27] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, Wei Li, et al., 'New algorithms for fast discovery of association rules', in *KDD*, volume 97, pp. 283–286, (1997).

[28] Eva Zangerle, Wolfgang Gassler, and Günther Specht, 'Exploiting twitter's collective knowledge for music recommendations.', in *# MSM*, pp. 14–17. Citeseer, (2012).

Disasters and Transformation of Daily Life: Implications for Issues in Risk Management

Yumiko Nara [1]

Abstract. The purpose of this paper is to organize aspects of life along a temporal axis and on the basis of their constituent elements after the occurrence of the disaster; the Great East Japan Earthquake. Based on this arrangement, the author aims to investigate issues in life risk management when a disaster occurs. The content was based on the accounts of victims the author interviewed in areas stricken by the earthquake. The results revealed that life adapted to the disaster conditions (life resources place conditions on life values) is a process that causes agency to gradually recover (regulation by life values). Furthermore, when this transformation of life is re-perceived from the perspective of life risk management, at least six issues could be identified.

1 INTRODUCTION

During a disaster, we are clearly placed in a different situation compared with normal times. Earthquakes and tsunamis strike the lives and properties of individuals, who are members of society. They also damage buildings and social infrastructure that provides the framework of society, as well as organizations, which are the constituent units of society. Their damage leads to the dissolution and dysfunction of families, communities, businesses, and government agencies. As a result, social activities suffer secondary damage. The destruction of the overall social system and the breakdown of the flow of activities finally cause hardship to the life and existence of the victims who were afflicted by the disaster.

The process of temporarily social adaptation under such conditions, which differs from normal social processes, results in an emergency social system. In an emergency social system, responses different than usual are required of society, organizations, communities, and individuals. For example, organizational responses during a disaster are carried out under circumstances in which, compared with normal times, there are greater uncertainty, heightened sense of urgency, and reduced autonomy. Similarly, for individuals, procuring the usual quantity and quality of life resources becomes extremely difficult (because lifelines are usually severed). People must act under a new model of life not experienced during normal conditions. This new model includes building new life relationships while being involved in rescue and recovery activities and living in shelters.

Meanwhile, regardless of what kinds of risks are present, a natural disaster requires a length of time to reestablish life. When we hear about an earthquake (or a tsunami), we think about raw conditions for just a while immediately after it (or a tsunami) occurs. However, for those stricken by the disaster, the suffering actually lasts from several months to several years. As a result of the Great East Japan Earthquake on March 11, 2011, emergency social systems arose for all areas of life in the stricken regions. People also began their prolonged period of life affected by the disaster.

Multifaceted and comprehensive understanding of current conditions and problem-solving are essential for restoring life after a disaster. Too many life elements were damaged by the 3/11 disaster. The difficulty of comprehensively understanding life elements and connecting them systematically is still keenly felt after the earthquake.

While being conscious of this limitation, this paper seeks to clarify the following two points while focusing on natural disasters. First, I seek to understand the process of transformation of life as a result of a disaster. For this purpose, the author aims to organize aspects of life along a temporal axis and on the basis of their constituent elements after the occurrence of the disaster. Based on this arrangement, I seek to investigate issues in life risk management when a disaster occurs.

2 DAMAGE AND FRAMEWORK FOR UNDERTANDING TRANSFORMATION OF LIFE

2.1 Disaster and Time

Suffering from a disaster is the process of destruction, depravation, and loss experienced by society and its members as a result of a natural disaster such as an earthquake, volcano eruption, or flooding (Hirose 1996).

Immediately after the occurrence of a disaster, a period of emergency response ensues, followed by a period of recovery and rebuilding. The emergency response period that accompanies bewildering changes immediately after an earthquake occurs can be divided into three stages: Phase 0 – the period of disorientation (from time of earthquake to 10 hours), Phase 1 – the period of establishment of community in the stricken area (10 hours – 100 hours after the earthquake), and Phase 2 – the period of disaster utopia (100 – 1,000 hours after the earthquake). These stages are followed by Phase 3 (1,000 hours after the earthquake), the period of recovery and rebuilding (Hayashi 2003).

[1] Human Life and Health Sciences, The Open University of Japan, email: narayumi@ouj.ac.jp

Table 1. Time passage after disaster

	Level of loss	Phase
Emergency response period	・Life (loss of life) ・Property (loss of property) ・Interruption of life (disruption of daily life) ・Terror (loss of psychological calm)	Phase 0 (Period of disorientation): From occurrence of earthquake to 10 hours Victims must survive on their own power.
	・Life (loss of life) ・Property (loss of property) ・Interruption of life (disruption of daily life)	Phase 1 (Period of establishment of community in stricken area): 10 – 100 hours Primary focus is on activities to rescue victims. Obtaining information becomes possible. Organized disaster response begins.
	・Property (loss of property) ・Interruption of life (disruption of daily life)	Phase 2 (Disaster utopia period): 100 – 1,000 hours The spirit of mutual aid becomes prominent. As social functions return, obstacles to life gradually improve.
Recovery and rebuilding period	・Property (loss of property)	Phase 3 (Recovery and rebuilding period) : From 1,000 hours onward Rebuilding life and livelihood. Recovery of damaged towns. Economic revival begins.

2.2 Constituent Elements of Life

Life is composed of the following elements: life values, life norms, interest in life, and relationships in life, life activities, and life resources. These elements have the following content: (a) Life values: standards of individuals for determining how to live life better. “What is important in life?” (b) Life norms: social norms established in different situations being reorganized to meet the concrete circumstances of people who actually carry them out and the reactions of other people. Rules held by each person. (c) Interest in life: each person’s interest concerning life. (d) Relationships in life: Human relationships that bind people in all situations as prescribed by spheres in life. They include roles held by individuals. (e) Life activities: activities that put life resources into position in actual life situations. (f) Life resources: methods, tools, and information used in life, e.g. economic resources, ability resources such as physical strength and information literacy, interpersonal relationships, etc.

In life during normal times, a person’s life values regulate his or her life norms, interest in life, life relationships, and furthermore, his or her life activities and life resources. Life is formed from the aggregate of this process. When this process is accompanied by a person’s independence and planning, this human agency makes up life management (Figure 1). On the other hand – and here I state my conclusion – during times of emergencies, the opposite direction takes place. In other words, life that adapts to emergency conditions is formed by directionality in which the possession of life resources conditions life values. Of course, during normal times, resources condition other elements. However, during emergencies, this trend becomes pronounced.

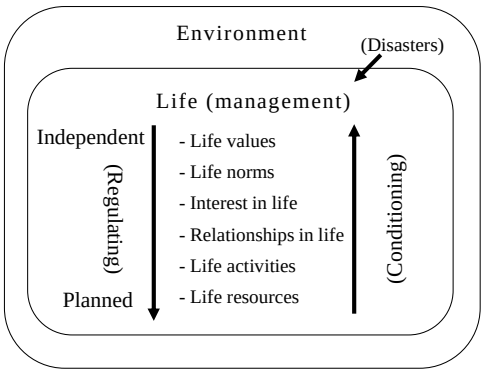

Figure 1. Life (management) and its constituent elements

3 THE GREAT EAST JAPAN EARTHQUAKE AND TRANSFORMATION OF DAILY LIFE

Using the framework described above, the author below the general transformation of life as a result of the Great East Japan Earthquake. The author entered into the field on 28th March 2011, and formed rapport with the victim by repeating communication. Afterwards, I continually (about once or twice a month from March to August 2011, about once in three or four months after September 2011) conducted interviews and made onsite observations of the damage and recovery in stricken areas along the coast. The field works were performed for approximately 36 months, in the municipalities of Miyako City, Otsuhi-cho, Ofunato City, Minamisanriku-cho, Ishinomaki City, Natori City, Iwanuma City, Watari-cho, and Minamisoma City. The author organized the contents of the interviews along a timeline. The contents of Phase 0 and Phase 1 are based on the interviewees’ recollections of the conditions at the time.

3.1 Phase 0 (from Occurrence of Earthquake to 10 Hours Later)

In this phase, all of life’s constituent elements were mobilized to preserve life. Comments from respondents included: “I escaped with just the clothes of my back.” “The water came over the levee of the rice fields. I desperately climbed the hills behind my home.” “The water surged and rose after than I expected. At first I escaped to the grounds of a temple, and the further to the cemetery above it.”

From the following comments, we see that that the extent of possessing ability resources (especially physical abilities and information literacy), information resources, and interpersonal relationship resources together had an effect the preservation of life: “The earthquake came, and immediately I thought that the tsunami would come. The appearance of the sea had changed. My wife and I escaped to higher grounds while we called out to our

neighbors." "My wife helped me and we were able to escape." In other words, being able to perceive risk by receiving signs of the coming tsunami at the level of primary information, having the physical strength to escape, and, if one did not have these two abilities, having others who compensated for their lack led to the preservation of life.

Also, as shown by the following comments, in an urgent situation, people performed roles in which they helped family members, neighbors, co-workers and patients: "The water came up above the floor. I first moved my father, who had been still sleeping, to the second floor. I then carried our dog up to second floor." "The store was swept away. I frantically evacuated all the store clerks, who were women, to higher grounds." "I was working at the airport. We have disaster response training, but I was quite scared. I evacuated the customers to higher floors." "The water kept rising. The first floor of the hospital was flooded up to the ceiling. In the muddy water, I desperately moved the patients to the second floor or above."

3.2 Phase 1 (10 hours – 100 hours)

In this phase, there was still a strong interest in preserving (saving) life, as shown by the following comments: "My home was swept away, but I kept conducting rescue as a firefighter." "I could not find my wife and kept searching."

Acquiring and distributing life resources were carried out at an extremely primitive level. "The gymnasium, was which the evacuation shelter, was cold everywhere. There was nothing to eat." "All of us shared water in plastic bottles and a little amount of food. The amount of water a day was just 1 cm in a plastic bottle." Such were the conditions, and resources from the outside were not available. A respondent said, "The Self Defense Force was amazing. I thought they were like God," showing the severity of this period.

During this time, the victims managed life after the disaster amidst building life relationships, as could be seen from the following comments: "Everyone evacuated to the temple's community center and spent time there. Neighbors took care of the elderly." "It was a relief to meet neighbors at the shelter. Everyone encouraged one another." At the same time, role conflicts began to become serious, as shown by the following comments: "My own home was swept away by the tsunami. But the city hall was always swamped by the demands of residents." "There's no way I can leave the hospital. I'm so worried about my parents."

3.3 Phase 2 (100 hours – 1,000 hours)

In this phase, individual disparities and regional disparities in acquiring life resources began to stand out. Comments from respondents included the following: "At the shelter, what was difficult was not being able to take a bath. There was also no underwear that fitted me." "I felt constricted at the shelter. I'm now having trouble with an acquaintance's family. But even there you can't stay that long. My own home was washed away, and I still can't find my wife. I don't know what to do from here on." "Help is not reaching those who have sought shelter in their own homes. I don't have a car and I can't go buy things. I'm having difficulty with food and clothes." "I want to move, but I don't have gas. My car was swept away." On the other hand, there were voices expressing contentment: "At this shelter, there are meals and we have enough to eat. The SDF came and set up a bath for us." "My brother in another prefecture lent me his car. That really helped me."

Also, changes in life activities and interest in life were often forced on the victims due to the loss of life resources: "I was a fisherman, but the harbor was destroyed. I lost my ship. I can't think about the future. I'm doing my best just to deal with each day." "What is most difficult for me now is losing my job. I don't know what to do from here on." "My job was farming, but my field was swept away. There's nothing to do every day."

In this phase, new life norms were produced and new role expectations for each person were assigned, as shown by the following comments: "Everyone helped one another at the shelter. Rules are being decided, such as who is in charge of cleaning and boiling hot water." "Everyone at the shelter really treated me (an elderly) well." "Relief volunteers really worked hard. They helped us. I'm thankful."

3.4 Phase 3 (After 1,000 hours)

When life after the disaster entered this phase, the effects of life resources, especially economic resources, on overall life became obvious. Interest in life became more concrete, but it was also affected by the disparity in possessing life resources. This could be seen by comments such as the following: "My home was swept away by the tsunami. I'm living now in my children's home in another city. I want to eventually return to my hometown, but it's financially difficult." "I want to rebuild my farm. But I'm already old and I also can't borrow the money I need, so I've given up. Right now, I'm doing a short-term job related to rebuilding after the earthquake. But I really want to farm again." "Young people are leaving the town to earn money. There's no choice, but I want to hold that down." "My wife, child, and I decided to rebuild our lives in a different place." "It really helped that we got earthquake insurance. It'll help in rebuilding our life." "My store was swept away by the tsunami. But the store clerks all survived. My family opposes this, but I have a separate plot of real estate, and I've decided to open a store again."

Also, from the following comments, we see that perception of one's own conditions of affliction depends on relative deprivation and the complex aspects of life relationships: "Supplies are being sent only to ○○ (name of an area), even though they are other areas that are equally damaged. It's unfair." "People's whose homes could be restored quickly could leave the temporary housing without even a week there."

Furthermore, in this phase, the victims looked back on the earthquake and attached meaning to it. As could be seen from the following comments, loss of and changes in elements of life, including life resources, conditioned life values: "After the tsunami, I was able to make progress step by step thanks to connections with other people. The reason I worked hard was to meet people. What changed my life was encountering others." "Anything is okay if I have my life. And after my life, if I have connections with people." "I want to preserve what I experienced, what I saw and heard, and what I thought."

3.5 Transformation in life from Phase 0 to Phase 3

The content of transformation in life from Phase 0 to Phase 3 is given in Table 2.

Table 2. Time passage of damage and daily life after the Great East Japan Earthquake

	Phase	Daily life
Emergency response period	Phase 0 (Period of disorientation): From occurrence of earthquake to 10 hours	○ All elements of life management are mobilized for preservation of life. ○ Roles are carried out under sense of urgency (helping family members, neighbor, patients, passengers) ○ The extent of waiting for life resources, especially information resources, ability resources, and interpersonal relationship resources, impacts preservation of life.
	Phase 1 (Period of establishment of community in stricken area): 10 – 100 hours	○ Concern for preserving (saving) life is still strong. ○ Life resources are acquired and distributed at an extremely primitive level. ○ Role conflicts become serious.
	Phase 2 (Disaster utopia period): 100 – 1,000 hours	○ Disparity in acquiring life resources → Disparity in life activities ○ Disparity in concern for life arises (due to conditioning of life elements as a result of resources) ○ Criteria arises for stricken areas. Victim roles are carried out in the areas.
Recovery and rebuilding period	Phase 3 (Recovery and rebuilding period) : From 1,000 hours onward	○ Meaning is formed as review of the disaster begins. → Conditions life values. ○ Actualization of concern for life grows, as does disparity in concern for life. ○ Acquisition of economic resources when it comes to life resources grows deeper. ○ Life relationships in an environment of relative predation.

4 ISSUES IN LIFE RISK MANAGEMENT INVOLVING NATURAL DISASTERS

4.1 Life Risk Management to Protect Life

From interviews of victims in the Tohoku coastal region, I came to understand aspects and processes of life affected by a natural disaster. Based on this understanding, I consider issues presented by natural disasters from the perspective of life risk management.

Interviewees have stated that suffering from a disaster continues for a long time. A major premise of this observation is “living.” Life management seeks to improve life. Risk management is positioned as a management process that contributes to this goal. Fundamentally, however, life management concerns one’s existence.

In other words, life management when it comes to natural disasters must be carried out with a focus on protecting life. In Phase 0, it is critical to analyze what happened in detail, to see which matters resulted in life and what resulted in death, and then apply the findings practically.

Stockpiling emergency food and water, confirming methods of communication between family members, and buying insurance are frequently recommended as practical disaster prevention measures that should be carried out in daily life. All of these practices are effective. However, their benefits are manifested after Phase 1. The extent of damage by a natural disaster such as an earthquake or tsunami is greatly influenced by the magnitude of its external force and society’s overall disaster prevention capability. We cannot deny that there are major limitations to management that can be carried out at the level of the ordinary citizen before and after the disaster. Even so, self-help that minimizes risk is required at the stage of Phase 0.

In the case of earthquakes, one of the highest-priority measures is to reinforce homes to withstand them. According Kobe City’s Great Hanshin Earthquake autopsy statistics, head trauma, internal trauma, neck trauma, suffocation, and external wounds caused by the collapse of buildings accounted for 83.3 percent of the causes of fatalities. Death by burning accounted for 12.8 percent, and other causes made up 3.9 percent (Hyogo Prefecture Medical Examiner, 1995). From this data, we see that securing living spaces so they do not collapse when the great external force of an earthquake hit is an urgent matter.

In the case of the Great East Japan Earthquake, on the other hand, about 90 percent of fatalities were caused by the tsunami. The step that residents should take against victimization by a tsunami is to evacuate as quickly as possible to higher grounds. The elements necessary for carrying out this measure are receiving accurate information about risk as quickly as possible, making appropriate judgments concerning the information, and then carrying out actions based on the judgments. Of course, the importance of risk information is similar for other natural disasters, such as earthquakes and typhoons. I wish to discuss this importance in detail next.

4.2 Importance of Risk Information

When a tsunami hits, victims are frequently delayed as they first confirm the occurrence of the event and then evacuate. Therefore, swift evaluation immediately after the occurrence of an earthquake is necessary. Toward this end, what is essential is risk information. Risk information that can be received by ordinary citizens is classified as primary or secondary information depending on the clear existence of others as presenters of the information and on the level of information processing. Information presented on an earthquake’s epicenter and magnitude immediately after it hits and tsunami prediction are considered secondary information.

Primary information come from natural phenomena (disaster phenomena), such as unnaturally low tides and unusual sounds of the wave. In Section 2(B)i, I described how receiving primary information was tied to swift evacuation. An interviewee in a stricken area said: “Because of my job in aqua farming, I go to the shore every day. On that day, the appearance of the sea was unusual. The waves looked different and sounded different. When they hit, they usually have the sound ‘chap, chap.’ But on that day, their sound was ‘zazazaza, zazazaza.’ They sounded as if the wind was hitting a thicket of bamboo grass. I thought it was strange, and that something was happening. So I finished my job early, and around noon I returned home, which is on a high elevation.” In addition to seeing and listening to the sea at an early stage, the respondent had a schema for understanding disaster phenomena

and a prepared action script, so he was able to appropriately evaluate primary information and take action.

To take appropriate actions based on correct judgments when natural disasters occur, it is necessary to expand one's schemas on the recognition of situations and the evaluation of actions. In other words, one must improve his or her quality of comprehension schemas and action scripts.

A comprehension schema is a pattern of deductions and inferences for predicting and evaluating not only processes but conditions. It helps a person realize what will happen next in a phenomenon. It also provides him or her with clues as to what usually accompanies such a phenomenon and its features. For example, when a landslide due to an earthquake occurs, the following signs may occur: small rocks fall, a crack opens on the side of a cliff, water gushes out, cracks appear in the ground, and eruptions and depressions occur. When one sees such phenomena and can determine that a landslide may take place, he or she is then able to save himself or herself. This is also the case with evaluating signs of an impending tsunami, such as unfamiliar sounds from the sea, rapidly ebbing tides, offshore rumblings, a major earthquake on the coast, and a slowly swaying earthquake. Having rich comprehension schemas lead to having high risk management abilities.

Action scripts are needed at the stage of deciding what responses to take next on the basis of the recognition of the risk phenomenon. An action script is a programmatic set of actions to take in a situation. An action script is a framework that is part of the knowledge structure of one's own behaviors. It is formed of a sequence of actions. Our lives are based on our action scripts, and because of them we can carry out actions in a rational order that are appropriate to the situation.

Therefore, possessing action scripts with appropriate contents for disasters leads to having high risk management abilities. Continuing with the example above, once individuals recognize the dangerousness of a landside or earthquake, the ability to flee as quickly as possible, to warn others, and to run to higher grounds far from the coast determines whether someone lives or dies.

If a person is riding the elevator when an earthquake hits, the basic principle is to press all the floors' buttons, make his body as small as possible, ride out the tremors, and get off on the nearest floor as quickly as possible. If he is trapped in the elevator, he should make contact with outside parties, avoid exhaustion, and wait for help to arrive. If he is driving a car while an earthquake is taking place, the appropriate response is to gradually slow down, stop the car on the side of the road, and turn off the engine. If leaving the car and evacuating is necessary, he should leave the key in the ignition and not lock the doors. Increasing the quality of one's comprehension schemas and action scripts by accumulating and learning from experience is a specific method of risk management against natural disasters.

Examples of secondary information related to tsunami disasters include tsunami warnings and evacuation recommendations and instructions. However, even though warnings and evacuation recommendations are issued when a tsunami occurs, there have been many cases reported where residents did not evacuate or tarried in evacuating (Matsuo et al. 2004, etc.). For example, the Japan Meteorological Agency issued a major tsunami (height of 3m or more) warning to the prefectures of Aomori, Iwate, and Miyagi on February 28, 2010, a day after an earthquake hit central Chili. However, according to the Fire and Disaster Management Agency, although the warning was issued to 500,000 residents, only 6.5 percent were confirmed to be in evacuation shelters.

The reasons for residents' sluggishness despite the announcement of tsunami warnings include the lack of understanding of disaster information and disaster phenomena, normalcy bias where disaster information is considered as "not a big deal" and thus underestimated, and the false alarm effect (the cry wolf effect) as a result of disasters that did not materialize from previous warnings (Katada and Kodama et al., 2005, etc.).

The effects of cognitive biases, including normalcy bias, when it comes to risk are considered serious. In the case of the Great East Japan Earthquake, in addition to normalcy bias, majority synching bias (everyone is not escaping, so it is okay not to either) and veteran bias (similar events have happened before) have been reported to contribute to delay in evaluation. When I visited a stricken area, an interviewee told me: "I see the high sea wall every day, so I thought everything would be okay. Even if a tsunami comes, things would be fine because of the sea wall. Everything had been fine before. My guard was down." This sea wall, nicknamed "The Great Wall," was built in the Tarou-cho district of Miyako City. However, the tsunami came over the 10m-high, 2.4km-long embankment. The Tarou district is an area vulnerable to tsunamis. The tsunami following the Meiji-Sanriku Earthquake in 1896 claimed the lives of 1859 residents – half of the area's population – and the 1933 Sanriku tsunami took the lives of 911 residents. To make sure that no more lives would be lost, the sea wall was built.

To swiftly evacuate residents, it is of course essential for the information presenters to create and transmit risk messages quickly, accurately and effectively. In addition, ordinary residents, who are the receivers of the information, must utilize the risk information. Based on the basic principle of "protecting your own life yourself," Katada and Kodama et al. (2005) noted that shattering a fixed image of tsunami disasters, doing away with the normalcy bias, understanding the mechanisms of how tsunamis occur, ridding oneself of overdependence on tsunami information, and improving tsunami information literacy are critical for residents.

4.3 Dealing with Role Conflicts

The lives of individuals are established as they create life relationships. This is still true during a disaster. Life relationships become concrete resources and are mobilized for the preservation of life and for living in shelters. At the same time, role conflicts may result for individuals. Life risk management must be established based on this fact.

An individual is a member of one's own family. At the same time, in many cases he or she is a member of other social systems (geographic community, workplace, etc.). As shown above, victims stricken by the disaster told of worrying about elderly parents who remained in their homes and about not being to leave their workplace even though they were concerned about their own homes. Their stories all demonstrate that they had no choice but to place higher priority on fulfilling the roles expected of them in the workplace. According to a survey by JICHIRO Miyagi Headquarters, over 20 percent of workers in municipalities such as Sendai City and Ishinomaki City in Miyagi Prefecture could rest for only a day or less during the two-month period after the earthquake.

An interviewee in a disaster area said the following: “The water came with surprising force. I have responsibility for disaster prevention in my hometown, so I went to help an elderly neighbor. As I left my home, I told my wife, ‘You get away, too.’ That was the last I saw of her.” This account shows that major tension arises between roles as community residents and as family members during a disaster, especially during the urgent Phase 0. According to the Fire and Disaster Management Agency, the number of fire corps volunteers dead or missing after the Great East Japan Earthquake was 253 in the three prefectures of Iwate, Miyagi and Fukushima. Of this figure, at least 72 died as they sought to close floodgates along the coast. Furthermore, there were many who became victims as they led residents in evacuation. Suffering due to altruistic acts of helping family and neighbors, instead of preserving one’s own lives, demonstrates the seriousness of role conflicts during Phase 0.

Resolving such tension is an extremely difficult issue. During Phase 0, carrying out “Tsunami Ten-den-ko” (immediate evacuation by everyone without concern for family or friends) is an effective method. A major premise of this measure is that individuals are responsible for becoming proficient with their own disaster response abilities, as emphasized by Katada (2011). For Phase 1 and afterwards, social and systemic studies on how resources from locations outside of stricken areas can be rapidly brought in are essential.

4.4 Increase of Resources in Each Phase

The fourth issue that emerged from the accounts of interviewees as they described how their lives were affected by the disaster is the question of how to increase resources on the basis of a reconsideration of who are the vulnerable.

In general, vulnerable people during a disaster are assumed to be the elderly and those with low physical abilities. However, anyone can be vulnerable depending on his or her possession of resources. For example, during Phase 0, even if a person is young and physically strong, if he or she lacks literacy that enables making appropriate decisions and carrying out appropriate actions on the basis of disaster information, or if his or her neighbor does not realize his or her existence, that person may experience delay in evacuating or in receiving aid, to a fatal result. The same goes for Phases 1 and 2. Also, when life after a disaster enters Phase 3, the possession of economic resources is greatly influenced by the conditions of suffering. There were victims who could compensate for the lack of these resources with interpersonal relationship resources. However, there were also victims who could not.

Everyone who lives in Japan, a disaster-prone archipelago, is a future victim. A person must undertake life risk management for all phases with the assumption that he or she may become a vulnerable person. As we have already seen, during times of emergency, available resources exert conditions on one’s life. We must consider during normal times what resources we are lacking and prepare them.

4.5 Balance of Self-Aid, Mutual Aid, and Public Aid during Emergency Times

Entities that have independence and act autonomously as constituent elements of social systems are called agents. When we consider an emergency social system from the aspect of the interactions of a variety of agents during a disaster, we see structures and functions that are different from normal times and a variety of agents at work. The first group of agents consists of individuals, families, and relatives; the second group consists of communities; the third group consists of the government; the fourth group consists of NGOs inside and outside the country; and the fifth group consists of overseas governmental, quasi-governmental, and international organizations. By taking another look at the mutual interactions between these agents from another angle, we can see them as constituting “self-aid, mutual aid, and public aid” during a disaster (Tanaka, 2007).

During a disaster, multiple agents carry out different roles compared with normal times. The ways they carry out the roles also differ. Before a disaster occurs, many residents expect much public aid, and they receive a variety of government services. However, during a disaster, the balance of self-aid, mutual aid, and public aid shifts from that of normal times. The ratio of the amount of self-aid, mutual aid, and public aid during a disaster is roughly 7:2:1, as calculated from the results of various fact-finding surveys of disasters (Kawada, 2008; Hayashi, 2003). When a disaster occurs, people have the same expectations of what the government can provide as they do during normal times. However, in actuality ordinary citizens must increase their self-aid capability. This is impossible to suddenly do when a disaster happens. Therefore, we must seek to strengthen our self-aid capability while carrying out life risk management during normal times.

4.6 Rebuilding Community and Life Risk Management

From the accounts related by interviewees, we consistently see that the strength of the community lends power to the lives of individuals during the process of moving through the phases in the aftermath of a disaster. During Phase 0, communal strength is directly involved in the rescue of lives. From Phase 1 onward, life is reorganized by the mutual aid of neighbors.

A variety of research studies have shown that high-quality social capital leads to crime prevention as well as to disaster prevention (Nara 2011, Hirschfield et al. 1997, Suzuki 2011, etc.). The Great East Japan Earthquake confirmed once again the importance of community.

To form and maintain a community, it is necessary for people to be rooted in the area. On the other hand, in areas stricken by a disaster, young people will leave to find employment elsewhere. Also, in affected areas where primary industries provided jobs, the 3/11 earthquake dealt a blow to both homes and employment. When an individual seeks to recover and rebuild his or her life, it must be accompanied by the recovery and rebuilding of the community, which is the physical and social sphere of life. For a currently stricken area, measures to rebuild employment and primary industries are necessary. For future disaster areas, improving communal strength is required during normal times.

5 CONCLUDING REMARKS

In this paper, the author organized the content of the transformation of life due to the Great East Japan Earthquake on a temporal axis and on the basis of the constituent elements of life. The content was based on the accounts of victims the author interviewed in areas stricken by the earthquake. The results revealed that life

adapted to the disaster conditions (life resources place conditions on life values) is a process that causes agency to gradually recover (regulation by life values). Furthermore, when this transformation of life is re-perceived from the perspective of life risk management, at least six issues could be identified. These challenges are: 1) carrying out life management as a management process to protect life, 2) quickly and accurately receiving risk information and making decisions based on the information, 3) recognizing the difficulty of dealing with role conflicts, 4) increasing resources based on a reconsideration of the vulnerable, 5) recognizing the balance of self-aid, mutual aid, and public aid during an emergency, and 6) the need to take measures to strengthen community and, at the same time, carry out life risk management.

To further study and carry out actual resolution of the life risk management issues described above, I wish to close this paper by discussing what should be studied for life risk management.

Under the 5th Science and Technology Basic Plan (adopted by the Cabinet on January 22, 2014), the following four principles were set as the goal for Japan: 1) a nation that achieves sustainable growth and autonomous development of community, 2) a nation that realizes a safe, secured, full, and high-quality life for its citizens, 3) a nation that leads in the resolution of global problems and the achievement of world development, and 5) a nation that continues to create intellectual property and to nurture a culture of science and technology. Furthermore, the following basic principles were established for future science and technology policies: 1) integrated promotion of science, technology and innovation (STI) policies, 2) greater priority on the roles of human resources and their supporting organizations, and 3) implementation of the STI policy created together with society. The main foci of these principles can be said to be "innovations" and "problem solving" (by government, industry and academia and in a cross-disciplinary manner).

Risk, including that of natural disasters, is a subject of academic and cross-disciplinary research and practice. Actually, risk is being dealt with in a variety of fields. Research on risk is being carried in the social sciences (economics, management studies, home economics, sociology, psychology, political science, law, ethics, education, etc.) and the natural sciences (engineering, chemistry, medicine, pharmacy, biology, agriculture, fishery, animal husbandry, forestry, earth sciences, etc.), as well in the quantitative sciences and information science (statistics, stochastics, etc.), which provide the academic foundation to the sciences. Research on risk is being conducted in such a wide range of academic disciplines because risk occurs in extremely diverse forms of human activities. Therefore, the ways of dealing with risk are also diverse.

To investigate the actual conditions of different risks, the background of their occurrences, their perception by human beings, and how to deal with them, an academic approach to risk by multiple academic domains is becoming necessary. Also, since the beginning, life is a multifaceted, comprehensive engagement. Multifaceted, comprehensive approaches are essential to risk management in life and to rebuilding life after a disaster. Here, problem-solving through government-industry-academia cooperation and cross-disciplinary cooperation not covered by academic domains are required.

REFERENCES

[1] Hirose, Hiroshi (1996), When Encountering a Disaster, The Asahi Shimbun Company

[2] Hirschfield, A. & Bowers, K. J. (1997) "The Effect of Social Cohesion on Levels of Recorded Crime in Disadvantaged Areas," Urban Studies, Vol.34, No.8, 1275-1295.

[3] Hayashi Haruo (2003), Earthquake Prevention Studies to Protect Lives, Iwanami Shoten, Publishers

[4] Katada, Toshitaka, Shin Kodama, Noriyuki Kuwasawa, and Shun'ichi Koshimura (2005), "Current State and Issues of Tsunami Prevention As Seen from Evacuation Behaviors of Residents: From Survey of Kesenuma Residents' Consciousness Concerning the 2003 Miyagi Earthquake", Journal of the Japan Society of Civil Engineers, No.789, Ⅱ-71, 93-104.

[5] Katada, Toshitaka, Motohiro Honma (2008), "Observations on Relationship between Information before Disaster and Residents' Evacuation for Tsunami Disaster Prevention", Journal of the Japan Society for Disaster Information Studies, No.6、61-72

[6] Katada, Toshitaka (2011), "Not 'Disaster Prevention with Knowledge' But 'Disaster Prevention with Attitude,", The Open University of Japan Communication On Air, No.103, 1-5.

[7] Coop Kobe, Seikyokenkyukiko (1996), Change in Life and Mutual Aid after Earthquake: Survey of Great Hanshin Earthquake Trade Union Member Survey: Midterm Report.

[8] Matsuo, Ichiro, Shunji Mikami, Hiromichi Nakamori, Isao Nakamura, Naoya Sekiya, Jun Tanaka, Saneyuki Udakawa, Hiroaki Yoshii (2004), "Tsunami Evacuation Behavior during 2003 Tokachioki Earthquake," , Disaster Information, No. 2, 12-23.

[9] Nara, Yumiko (2011), Life Risk Management, Foundation for the Promotion of the Open University of Japan.

[10] Nihei, Yoshiaki (2009), "Disaster Prevention Education Using Tsunami Evacuation Script,", Yoshiaki Nihei, ed. Psychology of Disaster Prevention, Toshindo.

[11] Noda, Takashi (1997), Disasters and Social Systems, Kouseisha Kouseikaku Corporation.

[12] Suzuki, Takeyasu (2011), Regional Disaster Prevention Capability That Overcomes Giant Disasters, Shizuoka Academic Press

[13] Tanaka, Shigeyoshi (2007), "Perspectives in Disaster Sociology," Jun Oyane, Masaki Urano, Atsushi Tanaka, Hiroaki Yoshii, eds. Introduction to Disaster Sociology (Series Disaster and Society I), Koubundou Publishers.

Emergence of Option Prices in Markets Populated by Portfolio-Holders

Sarvar Abdullaev[1] and **Peter McBurney**[2] and **Katarzyna Musial**[3]

Abstract. Options constitute integral part of modern financial trades, and are priced according to the risk associated with buying or selling certain asset in future. Financial literature mostly concentrates on risk-neutral methods of pricing options such as Black-Scholes model. However, using trading agents with utility function to determine the option's potential payoff is an emerging field in option pricing theory. In this paper, we use one of such methodologies developed by Othman and Sandholm to design portfolio-holding agents that are endowed with popular option portfolios such as bullish spread, bearish spread, butterfly spread, straddle, etc to price options. Agents use their portfolios to evaluate how buying or selling certain option would change their current payoff structure. We also develop a multi-unit direct double auction which preserves the atomicity of orders at the expense of budget balance. Agents are simulated in this mechanism and the emerging prices are compared to risk-neutral prices under different market conditions. Through an appropriate allocation of option portfolios to trading agents, we can simulate market conditions where the population of agents are bearish, bullish, neutral or non-neutral in their beliefs.

1 Introduction

The classic finance literature on derivatives is mostly based on Black-Scholes framework [1] which prices options from the perspective of no arbitrage assumption. According to this assumption, if there is a strategy with other financial instruments in the market which could simulate the payoff structure of the created financial contract, the value of such contract must be equal to the total cost of running this strategy. European option is one example of such financial contracts that gives the right to its holder to buy or sell certain asset at an agreed price in future. Black and Scholes showed that the payoff from holding an option (i.e. European option onwards) can be replicated by taking positions in two different markets: one is risk-free investments market, and the other is risky assets market. There is a mathematical solution which requires the parameters of these underlying markets to be set in order to compute the *risk-neutral* price of given option. In its initial formulation, Black-Scholes framework models the risky underlying market as Geometric Brownian Motion (GBM) which also implies the efficiency of that market. Moreover, the risk-free market was assumed to be static, so that the risk-free rate the investor has chosen to price the option remained constant throughout the lifespan of the option. Since then, similar models have been developed under different assumptions about the underlying market, and some of the important ones are Black-Scholes-Merton model which assumes that the asset prices are discontinuous [11] and Heston model which assumes that the volatility of the asset prices also changes according to certain stochastic model [7].

However it is known that contemporary financial markets are populated with heterogeneous traders using different methodologies that model the behaviour of markets. These are the crucial factors for each trader to make buying or selling decision. There have been many researches which propose agent-oriented approaches to price options, in contrast to the previous models which were directed at modeling the behaviour of the markets as a whole. In agent-oriented approaches, the behaviour of an individual trader is designed, so they can be simulated to obtain aggregate prices. This approach is also referred as *indifference pricing*, so the agents are indifferent to the exposed risk of buying or selling a contract based on their individual utility function. Another important aspect of indifference pricing is that there is no unique price as it happens to be in monolithic frictionless markets described in classic finance, but different bidding and asking quotes for each agent using different utility function. Gerber and Pafum described risk-averse traders based on an exponential utility function which could produce a bid-ask spread around risk-neutral option prices [4]. The width of the bid-ask spread could be specified by the risk-averseness factor of the exponential utility function. The other application of utility functions in pricing derivatives can be the use of intrinsic aspects of the agent's implementation such as portfolio, budget constraints, transaction costs and the other market related frictions. Carmona [2] and Henderson etal [6] provide extensive overview of indifference pricing methodologies used in practice.

Beyond designing the trader's behaviour, there has been a considerable advancement in designing market mechanisms too. Such mechanisms could determine the equilibrium prices and efficient allocations of goods from the corresponding behaviour of participating agents. Typical implementations of auctions have realised these objectives along with other important properties such as incentive compatibility, individual rationality and budget-balance. Strategyproof auctions can always guarantee truthfulness of participating agents through revelation principle and some even allow traders to be more expressive in revealing their combinatorial preferences. Parkes etal [13] and Parsons etal [14] provided up-to-date survey of different auction protocols, their designs and implementations.

The key question that we are posing in this paper is what option prices emerge if the traders come to the market already owning some option portfolios, and they also make their pricing decisions based on this fact. How would the prices be different from traditional risk-neutral prices? In this paper, we develop an agent-based system which uses direct double auction mechanism designed for trading multi-unit orders which also preserve the atomicity of orders. We

[1] King's College London, UK, email: sarvar.abdullaev@kcl.ac.uk
[2] King's College London, UK, email: peter.mcburney@kcl.ac.uk
[3] Bournemouth University, UK, email: kmusialgabrys@bournemouth.ac.uk

use inventory-based Logarithmic Market Scoring Rule (LMSR) option trader developed by Othman and Sandholm [12] to enable the option pricing based on the trader's current portfolio. Such agents price options based on the payoff structure of their current portfolio. For example, if the agent already owns short and long market positions on certain number of Out-of-The-Money (OTM), At-The-Money (ATM) and In-The-Money (ITM) options, his payoff can be different based on the underlying asset price on maturity date. And by pricing any given option with respect to his portfolio, agent computes the logarithmic score difference between his current payoff structure and the new payoff structure after buying or selling the option. We endow the LMSR traders with commonly used option portfolios such as *bullish spread*, *bearish spread*, *butterfly spread* etc. and run them in our proposed mechanism. This allows us to set up different scenarios in the market and observe the changes in option prices to evaluate their sensitivity to certain factors such as changes in the asset price or time-to-maturity of the option. The resulted prices are also compared with the theoretical Black-Scholes prices as well as their theoretical sensitivity to different factors. Besides that, we show the range of accepted bids and asks on each trading day, and analyse the relative efficiency and the distribution of differences in mechanisms budget. The key contribution of this paper is that it shows how option prices may differ from the theoretical prices when the option traders utility is based on their corresponding portfolios. Moreover, it proposes a new methodology in option pricing via double auctions which would enable the analysts to mix different indifference pricing methodologies together to obtain competitive option prices.

2 Options

Option is the type of financial derivative that enables its holder (i.e. owner) to buy or sell specified assets at certain future price to writer (i.e. issuer) of the option. Holder of the option buys for an additional cost (i.e option premium) determined by the market or the writer of the option. On the other hand, the writer of the option sells by taking future obligation to trade assets if holder chooses to exercise his right to buy or sell. Option contract must specify the underlying asset to be traded, its *volume*, *strike price* and expiration date. European options can be exercised only on their maturity date, while American options on any date until expiration. We will use only European options in the scope of this paper.

Options are defined as *put* or *call* options depending on rights and obligations that they bear. Put option gives its holder the right to sell underlying assets at agreed strike price, where the writer has the liability to buy them when holder exercises his right. Call option gives its holder the right to buy at agreed strike price, while the writer has the liability to sell. Option's value usually depends on several parameters of the underlying market such as spot asset price S_0, risk-free rate r and asset price volatility σ, and the conditions written in the option contract such as strike price K and time to maturity T. The other parameter of the asset (if it is a company stock) is the dividend it yields annually. This is normally subtracted from the overall return the asset is likely to make, but in this paper, we assume that the asset does not yield any dividend. There is an established relationship between put and call options with the same strike price and maturity date. This relationship results from the possibility of buying the one and selling the other. Using the put-call parity relationship, we can easily convert call prices to put prices, and vice versa. Therefore in our simulation, we only price call options, because the put price can be directly obtained from call price. Interested reader can look up the Black-Schole's formula[1] for risk-neutral pricing of options. Option belongs to different *moneyness* range at any given time t depending on whether its strike is greater or less than the current asset price. Table 1 summarises the options by moneyness.

	OTM	ATM	ITM
CALL	$K > S_t$	$K = S_t$	$K < S_t$
PUT	$K < S_t$	$K = S_t$	$K > S_t$

Table 1: Options by Moneyness

2.1 Greeks

Greeks analysis provides set of measurements for evaluating the sensitivity of option price on different factors in the market. They have important role in hedging portfolios and evaluating the volatility of the asset prices. We consider two of them for purpose of our analysis of option prices obtained from the simulation.

1. *Delta* $\Delta = \frac{\partial c}{\partial S}$: It measures the rate of change of the option price with respect to the change in price of the underlying asset. For example, if the delta of the call option is 0.4, then it means that if there is small change in the underlying asset's price, there will be a change in call options price in 0.4 of that amount. Delta is defined as the partial derivative of option's price function with respect to underlying asset price.
2. *Theta* $\Theta = -\frac{\partial c}{\partial \tau}$: It can be defined as the sensitivity of the option price to the passage of time, or 'time decay'. Its value is always negative, as option price becomes less sensitive to time as it approaches its maturity date. In other words, the payoff the option yields is more certain near its maturity.

2.2 Option Portfolios

We review different types of option portfolios used in practice. Traders can take different positions with options of different moneyness and create option portfolios which can align with their forecast and at same time limit their loss in case if their forecast is not true. Cohen counts more than 40 option portfolios and classifies them based on their market direction (i.e. bullishness or bearishness), volatility level, riskiness and gain [3]. We will not use all of them, but consider only the ones used in the scope of this work.

Let us consider, butterfly spread. This type of spread involves taking positions in options with three different strike prices. In butterfly call spread, trader has an estimate that the price is not going to change sharply, so he wants to stay neutral. He buys 2 call options: one ITM with low K_1 and one OTM with high K_3. At the same time, he sells 2 ATM calls with K_2, where K_2 is halfway between the range of K_1 and K_3. This spread leads to a profit if the asset price will not go far from its current spot price. It will incur in fixed loss if the asset price changes sharply in either directions. Butterfly spread can be created using put options as well. Figure 1 illustrates payoff structure of a butterfly call spread.

We can summarize the option portfolios used in the scope of this paper in Table 2 where c_A stands for ATM call, p_A ATM put, and so forth.

3 Portfolio-holding Trading Agent

In prediction markets [15], the informants are allowed to change the aggregator's payoff structure for a corresponding payment. For example, if aggregator is accepting bets for teams A and B on a football

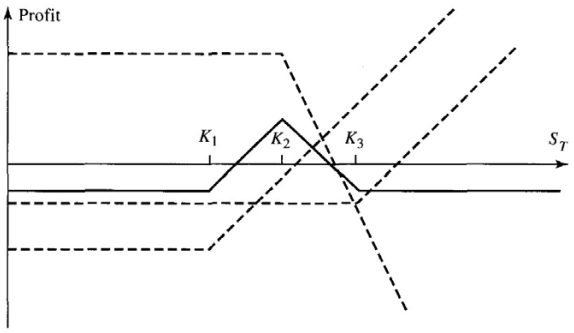

Figure 1: Butterfly Spread with Call Options.

Name	c_A	p_A	c_O	p_O	c_I	p_I
Bull Call Spread	0	0	-1	0	1	0
Bear Call Spread	0	0	1	0	-1	0
Butterfly Call Spread	-2	0	1	0	1	0
Long Call Ladder	-1	0	-1	0	1	1
Short Call Ladder	1	0	1	0	-1	0
Iron Butterfly	-1	-1	1	1	0	0
Long Straddle	1	1	0	0	0	0
Long Strangle	0	0	1	1	0	0
Short Strangle	0	0	-1	-1	0	0
Strip	1	2	0	0	0	0

Table 2: Some of the Popular Option Portfolios

match, and his current payoff structure is (300,200) meaning that the aggregator has to pay \$300 in total if team A wins, and \$200 if team B wins. However one would like to bet on team A, and he expects to receive \$50 if his bet is achieved. The aggregator changes his payoff structure to (350,200) by accepting the bet, and he also needs to decide how he can charge the client for accepting his bet. The most common method for evaluating the cost of accepting the bet in prediction markets, LMSR [5] and it is defined as a cost function for the vector of payoffs $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ on the probability space of events $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$:

$$C(\mathbf{x}) = b \log \left(\sum_i \exp(x_i/b) \right) \tag{1}$$

where $b > 0$ is a liquidity parameter. The larger values of b produce tighter bid/ask spreads, but may also incur larger worst-case losses capped by $b \log(n)$ [12].

The agent who wishes to change the payoff from $\mathbf{x}$ to $\mathbf{y}$ has to pay the difference between the costs $C(\mathbf{y}) - C(\mathbf{x})$. In our above example, given that $b = 100$, the aggregator accepting bets must charge the client $C((350, 200)) - C((300, 200)) \approx \$39$ for the bet.

The same principle can be used for the option trader who holds a certain portfolio of options that generate certain payoffs for different asset price outcomes, and prices other options from the point of his own payoff structure. The agent can virtually simulate buying or selling particular type of option and compute the changes it makes to his current payoff structure. For example, let agent take butterfly call spread by buying ITM call at strike $K_1 = 80$ and OTM call at $K_3 = 120$, and selling 2 ATM calls at $K_2 = 100$. We can compute his discounted payoffs for the range of possible prices where the asset price can end up at time T. Let this payoff structure be $\mathbf{x}$, and it can be depicted as in Table 3. The trader feels bullish and wants to buy one more call option at strike $K_4 = 130$. This should change his payoff structure to $\mathbf{y}$ as shown in Table 4.

Asset Prices	**Payoffs x**
< 70	0.00
75	0.00
80	0.00
85	4.75
90	9.51
95	14.26
100	19.02
105	14.27
110	9.51
115	4.75
120	0.00
> 125	0.00

Table 3: Potential payoffs of the trader holding butterfly call spread BEFORE buying call at $K_4 = 130$

Asset Prices	**Payoffs y**
< 70	0.00
75	0.00
80	0.00
85	4.75
90	9.51
95	14.26
100	19.02
105	14.27
110	9.51
115	4.75
120	0.00
125	0.00
130	0.00
135	4.75
140	9.51
> 145	$e^{-rT}(S_t - 130)$

Table 4: Potential payoffs of the trader holding butterfly call spread AFTER buying call at $K_4 = 130$

As it can be seen from tables 3 and 4, buying a certain type of option can significantly change the payoff structure of the trader. The agent's bid for buying an OTM option at $K_4 = 130$ can be computed from the difference of his payoff structures $C(\mathbf{y}) - C(\mathbf{x})$, and in our particular case is \$1.42 given that the liquidity parameter is $b = 2500$. We can also compute the Black-Scholes price of such option given parameters $T = 1$, $r = 0.05$, $S_0 = 100$, $\sigma = 0.02$ and it is \$2.52. This would mean that the trader places a bid for given OTM option less than its risk-neutral value.

We have simulated bids and asks for the call option with different strikes setting the liquidity parameter $b = 100$ and compared it with Black-Scholes prices. Figure 2 illustrates the bids and asks of LMSR trader holding butterfly call spread. This shows the breadth of bid-ask spread for the call option under different strikes.

To sum up, the indifference pricing methods like LMSR takes into account the agent's current portfolio and make option pricing decisions based on this information. We can use this principle in populating the option market by traders holding various portfolios, and observe the resulted option prices.

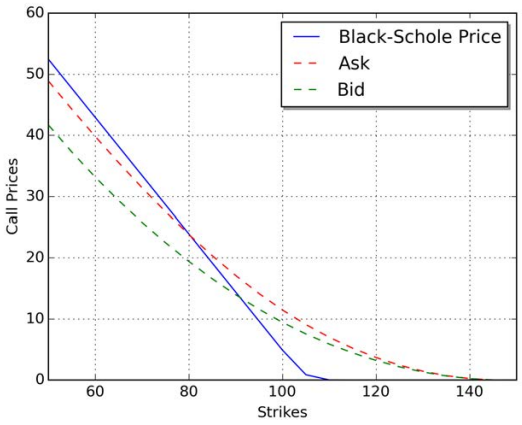

Figure 2: Bids and asks of LMSR trader holding butterfly call spread in comparison with Black-Scholes prices.

4 Direct Double Auction

We extend McAfee's double auction [10] to a multi-unit double auction, but in the process we have to give up its weakly budget-balanced property and introduce an agent who has to subsidise the exposed multi-unit bid or ask in order to preserve strategyproofness of the mechanism and the atomicity of orders. There have been a number of multi-unit double auction designs proposed previously [8, 9] which support weak budget-balance property. However these mechanisms partially satisfy the orders to spread the burden of overdemand or oversupply. In this design, we propose a multi-unit double auction that preserves the atomicity of orders at the expense of budget-balance. The reason for such a requirement is that it is crucial for the option trader who uses option portfolios. Because option portfolio determines exactly at which quantity each type of option needs to be sold or bought, trader cannot take quantity less than requested. The violation of atomicity of orders would result in the distortion of option portfolio as a whole.

Consider multi-unit bid as a tuple $\mathbf{b}_i = (b_i, q_i)$ where b_i is per unit bid, and q_i is the amount demanded. The same is defined for multi-unit ask. We can split this tuple into set of equally-valued single-unit bids $\mathbf{b}_i = \bigcup_{t=1}^{q_i} b_{i,t}$. This can be done to asks as well. Then we have complete set of bids $\mathbf{b} = \bigcup_{i=1}^{n} \mathbf{b}_i$ and asks $\mathbf{a} = \bigcup_{i=1}^{n} \mathbf{a}_i$. We can use single-unit McAfee's mechanism to find the allocation and payment. However, we can observe below that not all bids/asks can be fully satisfied.

Lemma 4.1. *In multi-unit McAfee's mechanism, there exists at most one multi-unit bid/ask which is partially satisfied, and the remaining winning bids/asks are fully satisfied.*

Proof. Let us assume that we use McAfee's matching rule for expanded set of single-unit bids $\mathbf{b}$ and asks $\mathbf{a}$ ordered subsequently by its host multi-unit bid or ask. Then we should have some k such that $b_{(k)} \geq a_{(k)}$ and $b_{(k+1)} < a_{(k+1)}$ for their constituent single-unit bids and asks. We can also claim, without loss of generality, that there exists such a multi-unit bid $\mathbf{b}_i$ such that two of its bids $b_{(k)}, b_{(k+1)} \in \mathbf{b}_i$. This would imply that $b_{(k)} = b_{(k+1)}$. However, there cannot be some multi-unit ask $\mathbf{a}_j$ having asks such that $a_{(k)} = a_{(k+1)}$, because it contradicts with $b_{(k)} \geq a_{(k)}$ and $b_{(k+1)} < a_{(k+1)}$. Hence, $a_{(k)}$ and $a_{(k+1)}$ must belong to different multi-unit asks. It must also be the case that the multi-unit ask which owns $a_{(k)}$ is fully satisfied, and so do other preceding winning multi-unit bids and asks. □

We can formulate an LP problem for for multi-unit bids and asks where $\lambda_i \in [0, 1]$ now. So it is not binary any more, and takes any value between 0 and 1. When it takes 1, the multi-unit bid/ask is fully satisfied, zero means it is rejected. But when $\lambda_i \in (0, 1)$, the agent i is partially satisfied. For given vectors of valuations and quantities (v, q), allocation rule for multi-unit double auction is:

$$\max_{\lambda} \sum_i q_i \lambda_i v_i \tag{2}$$

$$s.t. \quad \lambda_i \in [0, 1] \quad \forall i \tag{3}$$

$$\sum_i q_i \lambda_i = 0 \tag{4}$$

where $q_i \in \mathbb{Z}$ represents quantities, v_i is the agent's valuation, λ_i is an allocation decision variable.

The solution of above allocation problem can be used to find the volume demanded and supplied. Below are the formulas for computing the volumes of matched multi-unit bids V_b and asks V_a.

$$V_b = \sum_i q_i \quad \text{s.t. } q_i > 0, \lambda_i > 0 \tag{5}$$

$$V_a = \sum_i |q_i| \quad \text{s.t. } q_i < 0, \lambda_i > 0 \tag{6}$$

Let us denote the number of multi-unit bids matched (both fully and partially) as K, and for multi-unit asks L. Also Kth multi-unit bid would mean the lowest bid matched, and Lth multi-unit ask would mean highest ask matched. I denote their quoted valuations as b_K and a_L, and quantities as bq_K and aq_L respectively. From Lemma 4.1, we know that there is at most one $\lambda_i \in (0, 1)$ exists, so let us denote this as λ^*. It can also be noted that if such λ^* exists, it either belongs to Kth multi-unit bid, or Lth multi-unit ask. Now depending on whether λ^* exists, and if it exists, to whom it is assigned to, we apply appropriate payment rule. There are 3 cases that can emerge in this mechanism:

1. No λ^*: This would mean that supply and demand is matched exactly, hence $V_a = V_b$. In this case, buyers pay at b_{K+1}, sellers receive at good a_{L+1}. Because $b_{K+1} < a_{L+1}$, mechanism subsidises the deficit of $V_a(a_{L+1} - b_{K+1})$.
2. λ^* is assigned to buyer: This means that there is an over-demand, hence $V_b > V_a$. In this case, mechanism rejects Kth multi-unit bid. If there is a tie, it is randomly resolved. The remaining $K - 1$ buyers pay b_K per unit, L sellers receive a_{L+1} per unit. As the implication of Kth buyer rejection, a number of sellers at the bottom of the list can be exposed to $V_a - V_b + bq_K$ number of goods unmatched. So mechanism pays out $a_{L+1}(V_a - V_b + bq_K)$ to them. Because $b_K < a_{L+1}$ and number of full matches is $V_b - bq_K$, mechanism subsidises in total the deficit of $(V_b - bq_K)(a_{L+1} - b_K) + a_{L+1}(V_a - V_b + bq_K)$.
3. λ^* is assigned to seller: This means that there is an over-supply, hence $V_b < V_a$. In this case, mechanism rejects Lth multi-unit ask. If there is a tie, it is randomly resolved. The remaining $L - 1$ sellers receive a_L per unit, K buyers pay b_{K+1} per unit. As the implication of Lth seller rejection, a number of buyers at the bottom of the list can be exposed to $V_b - V_a + aq_L$ number of goods unmatched. So mechanism sells out in total $b_{K+1}(V_b - V_a + aq_L)$ worth of goods, and generates income. Because $b_{K+1} < a_L$ and number of full matches is $V_a - aq_L$, mechanism subsidises in total the deficit of $(V_a - aq_L)(a_{L+1} - b_K) - b_{K+1}(V_b - V_a + aq_L)$.

In above payment rules, mechanism is not only taking loss from clearing bids and asks at their offsetting prices, but also covering the exposed bids and asks resulting from the rejection of least efficient traders. Although the first part of the mechanism's loss can be insignificant in competitive markets due to narrow difference between inefficient bid and ask, the second part contributes the large portion of it, as the mechanism takes the responsibility to cover the exposed bids or asks. Given that the difference between $a_{L+1} - b_K$ is insignificant, the worst case budget-deficit for the mechanism is given below:

$$\bar{q}(K-1)(a_{L+1} - b_K) + a_{L+1}(\bar{q} - 1) \tag{7}$$

In worst case budget-deficit scenario, all buyers submit cap quantities $\bar{q}$, and Kth multi-unit bid is covered for $\bar{q} - 1$ of its bid. The mechanism rejects Kth bid, and leaves $\bar{q} - 1$ quantities for matched asks exposed. Mechanism spends extra $a_{L+1}(\bar{q} - 1)$ to cover these exposed asks. Hence it is the incentive of the mechanism to keep $\bar{q}$ as low possible to minimise its loss.

Theorem 4.1. *Proposed multi-unit double auction is Dominant Strategy Incentive Compatible (DSIC) and individual rational.*

Proof. Proof is done using Vickrey's argument. Without loss of generality, let us assume buyer i submits multi-unit bid (b_i, q_i) and $b_i > v_i$.

1. No λ^*: Then the clearing price is b_{K+1}, K buyers trade and there is no partially satisfied bid. If buyer gets fully satisfied, then $b_i \geq b_{K+1}$. So buyers utility is $v_i - b_{K+1}$, and in case if it is $v_i < b_{K+1}$ buyer gets negative utility, while if he posted v_i he would not trade and his utility would be zero. If $v_i \geq b_{K+1}$, the utility is indifferent to truthful bidding. If his bid is rejected, buyer is also indifferent.
2. λ^* assigned to buyers: Then the clearing price is b_K, $K-1$ buyers trade and there is one partially satisfied bid. If buyer gets fully satisfied, the above Vickrey's argument applies for critical bid b_K. If buyer gets rejected, he is indifferent to truthful bidding. However if buyer is partially satisfied, then $b_K = b_i > v_i$, he is rejected and he would be rejected for submitting v_i. So he is indifferent.
3. λ^* assigned to sellers: Then the clearing price is b_{K+1}, K buyers trade and there is no partially satisfied bid. The same argument for no λ^* case applies here.

In case if bidder submits $b_i < v_i$.

1. No λ^*: Then the clearing price is b_{K+1}, K buyers trade and there is no partially satisfied bid. If buyer gets fully satisfied, then $v_i > b_i \geq b_{K+1}$ and buyer has the same positive utility. If buyer gets rejected, and $v_i > b_{K+1}$, buyer misses the positive utility, otherwise he is indifferent.
2. λ^* assigned to buyers: Then the clearing price is b_K, $K-1$ buyers trade and there is one partially satisfied bid. If buyer gets fully satisfied, he is indifferent. If buyer gets rejected, the above Vickrey's argument applies for critical bid b_K. However if buyer is partially satisfied, then $b_K = b_i < v_i$, he is rejected and misses a positive utility.
3. λ^* assigned to sellers: Then the clearing price is b_{K+1}, K buyers trade and there is no partially satisfied bid. The same argument for no λ^* case applies here.

So there is a dominant strategy for buyer i, and it is $b_i = v_i$.

If buyer plays his dominant strategy, his utility is always non-negative. Hence, buyer is ex-post individual rational. Same argument applies to sellers. □

There two ways of looking at the efficiency of the mechanism we proposed. First way is computing the efficient trades happened within the mechanism. Because mechanism takes the place of Kth (Lth) rejected partially satisfied buyer (seller), the efficient trades are not lost. Hence mechanism can be considered efficient. However there is a partially satisfied bid (ask) rejected from the trade. In second way of looking at mechanism's efficiency, we can consider this rejected partially satisfied bid (ask) as the lost efficiency, because the traders are not benefiting from it. In this case, at most $\bar{q} - 1$ units of goods supposed for trade can be lost.

Also it is worthwhile to mention that the proposed mechanism is tractable, because it uses LP for determining the allocation which is polynomially solvable, and the payment rule is $O(1)$.

5 Experimental Results

We have simulated the asset prices using the Geometric Brownian Motion with a calibrated parameters according to the historic data of NASDAQ-100 index in 2014. The daily mean drift is computed as $\mu = 0.0007$, and and the volatility is $\sigma = 0.0089$. The figure 3 shows the instance of simulated asset prices that I use for all experiments. It can be seen that the initial asset price is the same as NASDAQ-100

Figure 3: Simulated NASDAQ-100 Indices

on 2 January 2014, $S_0 = \$3563.57$, and the it is $S_T = \$3597.59$ at the end of the year. This particular instance of asset price is interesting because it includes dramatic fluctuation near the end of the year. This should enable us to stress test the option pricing methods that are proposed. We also analyse option with strike \$3563.00 which is ATM. Also we use only call options for the simulation, because put prices can be directly computed from the call price using put-call parity relationship.

LMSR traders create a positive bid-ask spread, which forbids them from trading if the market is uniformly populated with LMSR traders holding the same portfolio. Therefore LMSR trader should be also simulated in mixed groups each holding different set of portfolios, and thus produce different prices. It is also important to note that LMSR trader is deterministic in their pricing, because the only factor which affects their pricing decision is their portfolio and fixed range of events horizon that they use to compute their final payoff. Therefore two LMSR traders holding the same portfolio produce same bids or same asks. To make market more heterogeneous, I use most of the option portfolios given in Table 2 to simulate traders from neutral, non-neutral, bullish and bearish perspectives. The full list of LMSR traders with the portfolios they hold is given in Table 5.

After running several experiments with LMSR traders, we found out that liquidity $b = 100$ provides reasonable range of bids and asks which are likely to produce trades in the market. LMSR trader picks random quantities between -2000 and 2000 while submitting orders,

Trader Name	Belief	Portfolio
LMSR-NEUT1	Neutral	Butterfly Call Spread
LMSR-NEUT2	Neutral	Iron Butterfly
LMSR-NEUT3	Neutral	Long Call Ladder
LMSR-NEUT4	Neutral	Short Strangle
LMSR-NON-NEUT1	Non-Neutral	Short Call Ladder
LMSR-NON-NEUT2	Non-Neutral	Long Straddle
LMSR-NON-NEUT3	Non-Neutral	Long Strangle
LMSR-NON-NEUT4	Non-Neutral	Strip
LMSR-BULL	Bullish	Bullish Call Spread
LMSR-BEAR	Bearish	Bearish Call Spread

Table 5: LMSR Traders and their portfolios

so agent's decision to buy or sell is uniformly distributed. The negative quantities stand for asks, and the positive ones are bids. Table 6 lists some experimental scenarios using different LMSR traders together.

Groups	Traders	Population
NEUT,NON-NEUT	LMSR-NEUT1	25
	LMSR-NEUT2	25
	LMSR-NON-NEUT1	25
	LMSR-NON-NEUT2	25
ALL	LMSR-NEUT1	25
	LMSR-BULL	25
	LMSR-BEAR	25
	LMSR-NON-NEUT1	25
MORE BULL	LMSR-NEUT3	10
	LMSR-BULL	70
	LMSR-BEAR	10
	LMSR-NON-NEUT3	10
MORE BEAR	LMSR-NEUT4	10
	LMSR-BULL	10
	LMSR-BEAR	70
	LMSR-NON-NEUT4	10

Table 6: Experiments with LMSR Traders

Mechanism simulates 365 trading days going up to the point the option expires. It feeds the option market with new asset price information and collects corresponding bids and asks from LMSR traders. Because mechanism is direct, it clears order in one round and switches to the next trading day. Every trading day, the traders are re-instantiated with the same distribution of portfolios so they do not remember their previous choices. Greeks are simulated by linearly changing the control factors (asset price or time to maturity) and fixing the other parameters constant, and inputting the given setting to a mechanism populated with the same traders.

Figures from 4 to 7 represent the simulation of different groups of traders given in Table 6. The yellow shaded area indicates the range of accepted orders for a given day, the blue line is the Black-Scholes price which is accepted as a benchmark model, and the red line indicates the average of clearing bid and ask prices for given day. In Figure 4 we can see the trade between neutral and non-neutral portfolio holders. It can be seen that the prices are volatile around Black-Scholes prices. This is explained using the deterministic nature of LMSR traders. The whole market consists of 2 neutral LMSR traders and 2 non-neutral LMSR traders who output all together 8 different pricing quotes, 4 for bids and 4 for asks. So naturally, one of 4 bids and one of 4 asks are used as the clearing prices for the matched orders. Because mechanism has very few choices to determine the clearing price among mostly homogeneous quotes, the option price for each trading day differs significantly. In Figure 5, we can see that the prices start with the same volatility, but upon maturity they get close to the option's theoretical price and the volatility around it subsides. Figure 6 illustrates the market mostly populated with bullish traders. In this example, there is no much volatility, and the prices are generally close to risk-neutral price. This is because the upward trend of the asset prices correspond with the traders' expectations. However in the market of bearish traders as shown in Figure 7, the call options are initially underpriced, because the direction of asset prices is opposite to traders' belief and therefore considered less profitable for them. However the prices cross the risk-neutral price only after option lives the half of its lifespan. This is because the payoff from the option becomes more certain, as the asset prices continue to grow defying the bearish belief of traders. We simulate the Greeks using

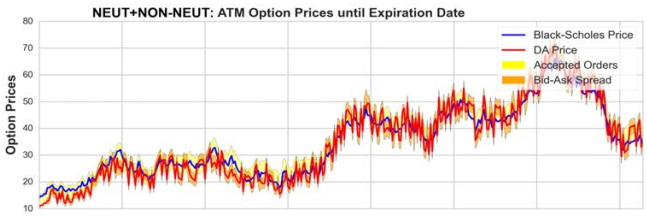

Figure 4: Market Simulation of Traders in NEUT and NON-NEUT Group

Figure 5: Market Simulation of Traders in ALL Group

Figure 6: Market Simulation of Traders in MORE BULL Group

the mixed population (i.e. group ALL) of LMSR traders for OTM, ATM and ITM calls and compare them with Black-Scholes analytical solutions. Simulation of Greeks involves fixing all parameters of the market, except the one which is tested for sensitivity. For example, the delta is measured by linearly changing the asset prices in the mechanism while fixing the passage of time and other parameters such as the population of traders, risk-free interest rates, etc.

Figure 7: Market Simulation of Traders in MORE BEAR Group

Figure 8 shows the deltas obtained from the simulation, and it can be seen that they steeper compared to Black-Scholes' analytical delta. This means that in a market populated with LMSR traders the option prices are highly sensitive to the changes in the asset price. Similar to risk-neutral pricing, the option price is highly volatile when the asset price is around its corresponding strike. The steepness of delta can be explained using characteristics of the mechanism and the LMSR traders involved. Because LMSR traders produce limited number of different bids and asks for the same option, and the mechanism has to clear the orders using the critical bids and asks, the sharp jumps in option prices are plausible. We have also simulated the option theta

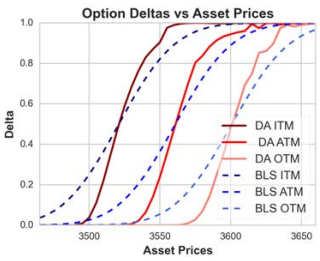

Figure 8: Option's delta

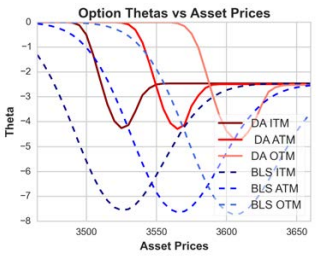

Figure 9: Option's theta

for the same configuration of the market. In Figure 9, we linearly changed the asset prices and computed the option's sensitivity to the change in time. We can see how the option price loses less than risk-neutral price when the asset price is around its strike as it approaches its maturity date. This is because LMSR traders are more inclined to their private beliefs related to the payoff from the portfolios they hold. This results in less change in option price compared to risk-neutral traders when the option is nearing its maturity date. We also observed that theta is less than the analytical solution if simulated per change in time-to-maturity.

From above simulations, we also found out that about 10% of efficient trades have been rejected by the mechanism to preserve the atomicity of other orders. It also means that mechanism had to cover up 10% of overall trades due to this rejection. Also surplus or deficiencies resulted from clearing the trades has been distributed around zero.

6 Concluding Remarks

In this paper, we have simulated LMSR-based option trading agents holding popular option portfolios in a multi-unit direct double auction to analyse the resulted option prices. Our simulation results have shown that pricing option via double auctions is a valid technique as the obtained prices were close to risk-neutral solution if the market was populated with traders having different beliefs. Besides that we were able to observe option prices under different population of traders with bearish, bullish, neutral and non-neutral beliefs modelled through their corresponding portfolios. For example, we have seen that the neutral and non-neutral traders create volatility as their portfolios consist of opposite payoff structure. Also in more bearish population of traders, the calls were initially underpriced until the option reached half-way to its maturity. When all traders are simulated together, the volatility of prices subsided as the option approached its maturity date. From Greeks analysis, we found out that the option prices are highly sensitive to the changes in the asset price, while they are less sensitive to the change in maturity date. This can be seen from the comparatively fixed width of accepted orders too in the simulation of the marketplace. Moreover we determined that in our current setting, we could generate enough volume if we set the liquidity parameter to $b = 100$ (although it is specific to particular price range of options, and not generic quality.). We also analysed the mechanism's relative efficiency and the budget balance.

This approach can be further improved using other incentive compatible mechanisms such as original McAfee's double auction with single-unit orders. Also the traders can be more sophisticated in making buy or sell decisions rather than randomly choosing the either action. Also other indifference pricing methods such as traders with exponential utility, zero-intelligence traders, etc can be simulated with our presented agents to observe their impact to the results obtained, and to analyse how they are different from the standard risk-neutral valuation techniques.

REFERENCES

[1] Fischer Black and Myron Scholes, 'The pricing of options and corporate liabilities', *The journal of political economy*, 637–654, (1973).
[2] René Carmona, *Indifference pricing: theory and applications*, Princeton University Press, 2009.
[3] G. Cohen, *The Bible of Options Strategies: The Definitive Guide for Practical Trading Strategies*, Prentice Hall, 2005.
[4] Hans U Gerber and Gérard Pafum, 'Utility functions: from risk theory to finance', *North American Actuarial Journal*, **2**(3), 74–91, (1998).
[5] R. Hanson, 'Logarithmic market scoring rules for modular combinatorial information aggregation', *Journal of Prediction Markets*, **1**(1), 1–15, (2007).
[6] Vicky Henderson and David Hobson, 'Utility indifference pricing-an overview', *Volume on Indifference Pricing*, (2004).
[7] Steven L Heston, 'A closed-form solution for options with stochastic volatility with applications to bond and currency options', *Review of financial studies*, **6**(2), 327–343, (1993).
[8] Pu Huang, Alan Scheller-Wolf, and Katia Sycara, 'Design of a multi–unit double auction e–market', *Computational Intelligence*, **18**(4), 596–617, (2002).
[9] Simon Loertscher and Claudio Mezzetti, 'A dominant strategy double auction with multi-unit traders', Technical report, mimeo, (2013).
[10] R Preston McAfee, 'A dominant strategy double auction', *Journal of economic Theory*, **56**(2), 434–450, (1992).
[11] Robert C Merton, 'Option pricing when underlying stock returns are discontinuous', *Journal of financial economics*, **3**(1), 125–144, (1976).
[12] A. Othman and T. Sandholm, 'Inventory-Based Versus Prior-Based Options Trading Agents', *Algorithmic Finance*, **1**(2), 95–121, (2012).
[13] D. Parkes and J. Kalagnanam, 'Auctions, bidding and exchange design', Technical report, IBM Research Division, (2004).
[14] Simon Parsons, Juan A Rodriguez-Aguilar, and Mark Klein, 'Auctions and bidding: A guide for computer scientists', *ACM Computing Surveys (CSUR)*, **43**(2), 10, (2011).
[15] David M Pennock and Rahul Sami, 'Computational aspects of prediction markets', *Algorithmic game theory*, 651–674, (2007).

On the logical and ontological treatment of IMDJ data

Akinori Abe[1]

Abstract. Recently we have been able to deal with a great volume of data. Accordingly the keyword "big data" is referred to in various type of applications. For that several frameworks for dealing with such data have been proposed. For instance, during performing the game Innovators Marketplace, participants generate novel products by combining several techniques. The strategy for generating new product can be regarded as abduction. In addition, if we deal with data in a logical processing such as abduction, it will be possible to perform more sophisticated procedure. In this paper, a logical processing of data in IMDJ as well as an abductive strategy will be discussed. In addition, an ontology based data access to IMDJ for inferences will be discussed.

1 Introduction

In these several years, we should deal with a great volume of data. Accordingly the keyword "big data" is often used for various type of applications. If we can have a lot of data, it will be preferable for data analysis. However, if the data size increases, it is rather difficult to deal with such big data. Of course, we should consider the moral and humanity perspective of the data. In such situation, the data cannot be published easily. In USA, the system of open data is supported by the Transparency and Open Government [18]. In fact the data is obtained by the government. On the other hand, in Europe legislation is conducted in order to protect the personal data [22]. Thus the policy of data treatment is different in the world.

In addition, semantic web has been moved and extended to the open data system called linked data. Where data is described in the format of RDF. The format is preferable for the data processing by the computer. Thus open data have been moved to linked open data (LOD) [12]. Accordingly from the viewpoint of data treatment scheme, data tends to be opened to the public in the world.

However, it is not easy to publicate the open data. Ohsawa proposed the system "Data Jacket" [17]. It can be regarded as the system in which when data is provided, its actual content is hidden but index which shows the abstract or list of the data is shown. Ohsawa used the metaphor of CD's jacket for the explanation of Data Jacket [14]. Accordingly Data Jacket can be considered a data set which shows only index of the data. That is, we can understand data's abstract feature but we cannot know the actual contents. This follows the concept of open data, but we cannot see inside of the data. It is possible control the level of publication.

Ohsawa et al. conduct the innovation game (Innovators Marketplace) based on the framework of Data Jacket [16]. During the innovation game players combine techniques to build new products. This type of activity can be regarded as abduction. Based on the above concept, I discussed the role of abduction in the innovation game [2].

[1] Faculty of Letters, Chiba University, Dwango Artificial Intelligence Laboratory, email: ave@ultimaVI.arc.net.my, ave@chiba-u.jp

In the following, I will discuss the logical treatment of data as well as the abductive strategy in the innovation game. In addition, an ontology based data access to IMDJ for inferences will be discussed.

2 Abduction and Clause Management System (CMS)

2.1 Hypothetical reasoning (abduction)

In this section, abduction is explained as hypothetical reasoning. In fact hypothetical reasoning prepares a hypothetical base to select consistent hypothesis set for the inference. So it does not build a new product from nothing, but performs an "operation of adopting an explanatory hypothesis [19]" which is pointed out by Peirce.

A hypothetical reasoning is a reasoning strategy to deal with incomplete knowledge. This "incomplete"ness means that it is not complete in the sense that all knowledge cannot always be trusted as true. Thus the inference performs as a non-monotonic reasoning [10, 11]. The well-known systems of abduction are Theorist [20] and ALP (Abductive Logic Programming) [9]. A hypothetical reasoning is illustrated based on Theorist.

A hypothetical reasoning which is an explanatory reasoning generates (collects) a consistent hypothesis set from hypothesis candidates to explain the given observation. The generated hypothesis set can be regarded as an answer (solution) which can explain the observation. This reasoning is a reasoning dealing with incomplete knowledge. This is because the hypothesis base using in the reasoning may contain not-true knowledge (Knowledge is not true in a certain situation or more than two knowledge are not true or consistent in the same time.). The following is the inference mechanism of a hypothetical reasoning.

$$F \not\vdash O. \qquad (O \textit{ cannot be explained only by } F.) \tag{1}$$

$$F \cup h \vdash O. \qquad (O \textit{ can be explained by } F \textit{ and } h.) \tag{2}$$

$$F \cup h \not\vdash \square. \qquad (F \textit{ and } h \textit{ are consistent.}) \tag{3}$$

where F is called fact which is always true. On the other hand, h is called hypothesis which is not always true and included in the hypothesis set H ($h \subseteq H$). O is an observation to be explained. $\square$ is an empty set. When $F \cup h \vdash \square$, F and h are not consistent. In the knowledge base, inconsistent information is also included.

In addition, for hypothesis generation (selection), minimality condition should be considered. Although several problems are pointed out, it is a proper condition for the effective reasoning. In fact, for the applications such as an LSI circuit design, the minimality condition functions well. But for the application such as a natural language

processing, the minimality condition does not function well [13]. For the problem, please see [13].

2.2 Clause Management System (CMS)

Clause Management System (CMS) [21] proposed by Reiter and de Kleer is a framework for the database management. The management framework can be regarded as abduction. Since CMS shows missing and necessary minimal clauses for explanation of a certain situation. That is, if we use CMS as abduction, it is possible to generate missing knowledge for a certain inference.

An abductive reasoning in CMS is conducted as follows:

When

$$\Sigma \not\models C, \tag{4}$$

(C (observation) cannot be explained only by Σ)

CMS will generate a minimal clauses set S to Σ for satisfying the following formulae.

$$\Sigma \models S \vee C, \tag{5}$$

$$\Sigma \not\models S. \tag{6}$$

($\neg S$ is not included in Σ.)

S is called minimal support clause. And $\neg S$ can be considered as a missing hypothesis set to explain C in the world (knowledge base) of clause set Σ. Accordingly $\neg S$ can be considered as an abductive hypothesis set. For the hypothesis reasoning, it is necessary to prepare a candidate hypothesis set. However, for CMS it is not necessary to prepare such a set. However, the generated hypothesis set cannot always be correct and sometimes it is too minimal. Then if CMS is introduced, it is necessary to be careful in the hypothesis set selection.

2.3 Abductive Analogical Reasoning (AAR)

As shown above, CMS generates only the minimal hypothesis set. Thus it is not always the case we can obtain the sufficient hypothesis set. Accordingly I proposed Abductive Analogical Reasoning (AAR) [1] that logically and analogically generates missing hypotheses. Its generation mechanism is similar to CMS's. Structures of generated knowledge sets are analogous to the known knowledge sets. In the framework of AAR, not completely unknown but rather unknown hypotheses can be generated. In addition, by the introduction of analogical mapping, we can adopt new hypothesis evaluation criteria other than Occam's Razor (for instance, criteria such as explanatory coherence [23]). The inference mechanism is briefly illustrated as follows (for notations, see [1]):

When

$$\Sigma \not\models O, \qquad (O \text{ cannot only be explained by } \Sigma.) \tag{7}$$

Σ (background knowledge) lacks a certain set of clauses to explain O. Consequently, AAR returns a set of minimal clauses S such that

$$\Sigma \models S \vee O, \tag{8}$$

$$\neg S \notin \Sigma. \tag{9}$$

The result is the same as CMS's. This is not always a guaranteed hypothesis set. To guarantee the hypothesis set, we introduced analogical mapping from known knowledge sets.

$$S \mid\!\sim S', \qquad (S' \text{ is analogically transformed from } S.) \tag{10}$$

$$\neg S' \in \Sigma, \tag{11}$$

$$S' \mid\!\sim S'', \tag{12}$$

$$\Sigma \models S'' \vee O, \tag{13}$$

$$\neg S'' \notin \Sigma. \tag{14}$$

O is then explained by $\neg S''$ as an hypotheses set. Thus we can generate a new hypothesis set that is logically abduced whose structure is similar to authorized (well-known) knowledge sets.

2.4 What is abduction for?

As shown above, abduction is regarded as the inference which can generate what the user wants to generate. In addition, even if the necessary data is missing from the database, abduction can generate or suggest the necessary data. In the following sections, based on the feature of abduction shown in this section, I will discuss the inference procedure in IMDJ as abduction and the possibility of an abductive inference in a database.

3 Innovation Marketplace on Data Jacket (IMDJ)

Innovation Marketplace on Data Jacket (IMDJ) is called as Innovation Game. The Innovation Game seems a game where a new production will be obtained during the combination of various techniques, materials and previous products.

Usually the game adopt an analogous game system. It uses a large paper. KeyGraph's output is printed on a large paper (game board). For instance as shown in Figure 1, techniques and their abstract explanations are printed. In addition, techniques are linked by the links generated by KeyGraph.

Figure 1. A game board in Innovation Marketplace on Data Jacket

The mechanism of KeyGraph generation was illustrated in [15]. So I will not give detailed illustration of the generation of the links. Participants place yellow cards (showing their requirements) and blue cards (their proposal) on the game board. By the above activities, several requirements and proposals are shown. For the game that I took part in, techniques other than those printed on could be used. Such techniques can be searched by the computer software. They are

used in the previous games. Participants generate several proposal by combining techniques on the game board and additional techniques. Then several applications which will satisfy requirements are proposed.

While the game participants buy proposals according to their sense of value (How their requirements are satisfied and the proposal is good.). In addition, proposals are valued by the all participants. The best proposal can obtain money. The winner is a participant who can obtain the biggest money. One of the strategy in this game is to obtain as many money as a participant can. This fact can be regarded as an economic game. For that it is necessary to produce the best and unique proposal that can satisfy one or more requirements. This can be regarded as the best design and requirement satisfaction problem. That is a type of constraint solving problem (CSP) which can be solved by abduction. I will discuss the production strategy not from the economic viewpoint but from the abductive viewpoint.

To summarize the game, when a requirement (yellow card) is given, a proposal (blue card) will be generated accordingly. Even if the proper techniques are not on the game board, techniques which were used in the previous game can be searched and used. In this sense, the strategy used in this game can be regarded as abduction. The detailed illustration of abduction was given in the previous section. In the next section, I will show the brief strategy of the Innovation Game.

Briefly explained, in the Innovation Game participants do their best to explain the certain requirement by generating or collecting a set of hypothesis (in this case, a set of techniques) in the current environment. Many creative activities can be performed as a case of abduction. Accordingly this type of inference strategy can be considered as abduction. Of course, if the current environment changes, the strategy and the conclusion will change. Thus the inference strategy has the characteristic of non-monotonic reasoning.

4 The possibility of inference in database

4.1 The role of abduction in IMDJ

In the Innovation Game participants do their best to explain the certain requirement by generating or collecting a set of hypothesis (in this case, a set of techniques) in the current environment. Then in the game, they buy and sell the products.

Participants place yellow cards (showing their requirements) and blue cards (their proposal) on the game board. By the above activities, several requirements and proposals are shown. For the game that I took part in, techniques other than those printed on could be used. Such techniques can be searched by the computer software. They are used in the previous games. Participants generate several proposal by combining techniques on the game board and additional techniques. Then several applications which will satisfy requirements are proposed. As a game, participants evaluate proposals that satisfy their requirements. If they are satisfied with a certain proposal, they pay money to obtain it. A person who can obtain the largest mount of money can win the game. In addition, at the end of game, proposals are evaluated again by participants. Then the main activity in IMDJ is to generate hypothesis set (in this case, a set of techniques) satisfying requirements in the current environment. Many creative activities can be performed as a case of abduction. Thus this type of inference strategy can be considered as abduction. In addition, if the environment changes, we must consider non-monotonic feature of reasoning.

A strategy "generating or collecting a set of hypothesis (in this case, a set of techniques) to explain a certain requirement." can be formalized as hypothetical reasoning as follows:

$$environment \not\vdash requirement. \tag{15}$$

(In the current environment, it is not possible to explain the requirement.)

$$environment \cup techniques \vdash requirement. \tag{16}$$

(In the current environment, when we can generate a set of techniques it is possible to explain the requirement.)

$$environment \cup techniques \not\vdash \Box. \tag{17}$$

(In the current environment, the set of techniques are consistent.)

For instance, if the requirement is such as "I would like to go to see the Olympic game by taking the rest," it will perhaps mean "I would like to go to see the Olympic game without any allowance, but I do not like my secret to be disclosed to my office." Then the above situation can be described as follows:

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \cup \{show\ a\ different\ place \cup perform\ tweet\} \vdash pretend\ his/her\ place. \tag{18}$$

Techniques shown in the technical map to achieve the above proposal are "shop arrival count data" and so on. By using them, the proposal satisfying the requirements will be generated. That is, a participant generates a technique satisfying the requirements "to pretend his/her place" and to make use of the tweet technique (perhaps of the Olympic game) by manipulating "shop arrival count data" rather differently from the original usage.

Then by the process shown in [2] hypothesis (necessary technique) can be generated. In fact, by using CMS and AAR, the above strategy can be explained. For instance, in the above example, suppose we do not have a technique "perform tweet." Then

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \cup \{show\ a\ different\ place\} \not\vdash pretend\ his/her\ place. \tag{19}$$

If we follow CMS, the following formula can be obtained.

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \not\models \{\neg show\ a\ different\ place\} \vee pretend\ his/her\ place. \tag{20}$$

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \models S \vee \{\neg show\ a\ different\ place\} \vee pretend\ his/her\ place. \tag{21}$$

In order to satisfy the requirement, it is necessary to find minimal S supporting the following formula. If we can obtain $S = \neg tell$, the above formula can be satisfied.

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \models \{\neg tell \vee \neg show\ a\ different\ place\} \vee pretend\ his/her\ place. \tag{22}$$

As for the clause "show a different place," a technique such as "shop arrival count data" can be used to achieve it [2]. The most difficult problem is how to generate a clause "tell." It will be possible by the introduction of AAR to generate a clause "tell."

When the following formula is obtained;

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \models \{\neg tell \vee \neg shop\ arrival\ count\ data\} \vee pretend\ his/her\ place. \quad (23)$$

It will be necessary to search or generate a technique similar to "tell." For instance, if we interpret "tell" as "communicate (send) his/her voice via air," then "broadcast" which can be interpreted as "communicate (send) his/her voice via radio wave," (thus "tell $\approx$ broadcast") can be searched from the knowledge base in the other context or newly generated. In addition, if we interpret "tell" as "communicate phrase (including characters) via air," then "perform tweet" which can be interpreted as "communicate phrase (including characters) vie the internet," (thus "tell $\approx$ perform tweet") can be searched from the knowledge base in the other context or generated newly. The selection of hypothesis set is determined considering if the hypothesis set is consistent with "the place the Olympic game is held" and other techniques, and if it can be used in the context. If it is not consistent, it is necessary to produce new techniques or to replace with the other techniques. This is a non-monotonic aspect of this reasoning. In this case, if the personal information is "broadcasting" in "the place the Olympic game is held," the activity "broadcast" will reveal the personal information. Therefore "broadcast" is not consistent in the situation. "Perform tweet" has a rather publication feature, but it may not be so bad compared with "broadcast." Accordingly, "perform tweet" can be adopted as a similar technique as "tell." The proposal shown in the technical map is determined, because the proposer thought "perform tweet" was enough to keep the privacy of the individual. If it is not still sufficient, it will be necessary to adopt another technique that can keep the privacy of the individual better, for instance, "send a message via FB (Facebook)" ("tell $\approx$ send a message via FB"). This type of non-monotonic hypothesis selection continues until a consistent hypothesis set can be generated. In addition, when the generated hypothesis set cannot be used as it is, an analogical mapping is introduced to generate new and usable hypothesis set in the situation.

$$knowledge\ of\ the\ place\ the\ Olympic\ game\ is\ held \mid\!\approx \{\neg send\ a\ message\ via\ FB \vee \neg shop\ arrival\ count\ data\} \vee pretend\ his/her\ place. \quad (24)$$

In the above, simple analogical mappings are shown, however it can be considered more complex angelical mapping such as the structure mapping shown by Gentner [7]. Actually a analogical mapping between, for instance, "tell and broadcast" can be discussed in the structure mapping.

4.2 Abduction inside the database

As shown above, in IMDJ if a certain technique cannot be found in the technical map, it can be searched on PC if it was used in the previous game. However, sometimes we cannot determine what is missing or what is necessary to satisfy the requirement. In addition, if the inside of the database is not known, it is difficult to search such knowledge. In such case, it is necessary to generate missing knowledge. Abduction used above can perform a hypothesis generation by hypothetical reasoning and an analogical hypothesis generation based on the similar hypothesis. However, it is rather difficult to generate a bland new hypothesis in the hypothetical reasoning framework. In such cases, it will be necessary to introduce such a system or strategy as CMS [21].

In CMS as shown in the previous section, the minimal knowledge missing from the database for the explanation of something is suggested. It is regarded as the same knowledge as that obtain by abduction. A database is not always perfect, it is preferable to generate missing candidate knowledge for the explanation of the observation. In addition, it is possible to check if the database is perfect or not. Especially the database whose inside is not open to the public, such procedure is necessary.

4.2.1 Relational Data Mining

Relational data mining is the data mining technique for relational databases. In this paper I will not discuss data mining, but the following concept "relational" should be considered in an inference.

Džeroski pointed out for relational data mining that "most existing data mining approaches look for patterns in a single table,... . Relational data mining approaches, on the other hand, look for patterns that involve multiple relations from a relational database" [5]. Thus relational data mining is different from the previous[3] data mining. It performs data mining of related data based on the framework of, for instance, relational database.

For instance, the example shown in [5] is:

```
IF Customer(C1, Age1, Income1, TotalSpent1,
                                BigSpender1)
  AND MarriedTo(C1, C2)
  AND Customer(C2, Age2, Income2, TotalSpent2,
                                BigSpender2)
  AND Income2 >= 108000
THEN BigSpender1 = Yes
```

Here it is pointed out that $C1$ and $C2$ are connected (related) via the relationship $MarriedTo$. Such a relationship can be described in the Linked data, and in Data Jacket such description should be considered. For instance, analogical relationship can be described in the same way such as followings:.

```
SimilarTo(broadcast, Facebook)
```

By using this type of data, a knowledge searching in database becomes more flexible. Then a flexible reasoning can be achieved. In fact, recently a relational data system has seemed to change to Linked Data system, that was a semantic web. In the Linked Data system, concepts are written in the format of RFD. RDF will be an extended format of a relational data system. Thus we can discuss the same issue in the semantic web and its extension of Linked Data.

On the Linked Data system, the relation such as `SimilarTo(broadcast, Facebook)` can be discussed in the structure-mapping case. In fact, the relationship is described in a linked information, but the linked information can be described in more complex description such as nested and hierarchical models.

[2] It is necessary to provide a background knowledge about this knowledge to perform this.

[3] This paper was written in 2001, but it can still be said so.

4.2.2 Stream reasoning

For the linked data the possibility and necessarily of the stream reasoning is discussed [4]. For the stream reasoning, it is pointed out that it is necessary to handle massive datasets, to process data streams on the fly, to cope with heterogeneous dataset, to cope with incomplete data, to cope with noisy data. to provide reactive answers, to support fine-grained information access, and to integrate complex domain models. According to the above requirements, the basic concept of the stream reasoning was formalized. Then several stream reasoning system such as C-SPARQL [4] and EP-SPARQL [3] have been proposed. They are an extension of SPARQL which is a query language for RDF. Actually the stream reasoning is a reasoner for stream data. This type of reasoning is very important for the changing data. But the most interesting feature of the stream reasoning is that it performs an ontology based data access. Gebser et al. formalize stream reasoning based on Answer Set Programming (ASP) [6]. This system utilizes time-decaying logic programs to capture sliding window data in a natural way. Except the inference for stream data, an ontology based data accessing reasoning can be considered in inferences in IMDJ. It will be a nice idea to introduce the concept of the stream reasoning to abduction, especially CMS-based abduction. Then more flexible reasoning can be achieved. In addition, a (structural) analogical reasoning can combined to the concept of the stream reasoning. Especially, if ASP-based stream reasoning is considered, a very useful system can be constructed.

4.3 Ontology based data access and reasoning in IMDJ

As shown above, the concept of the stream reasoning is suitable to reasoning in IMDJ. The above reasoning can be achieved by the concept of the ontology based data access. In the system of ontology, data are described in relationships which are hierarchical and linking relationships. If we can follow such relationships, even if we do not know the inside of the database, a certain inference can be achieved by abduction, and the necessary technique can be suggested. By a certain control, such suggestion can hide an actual the information of the inside of the data. Thus we can evaluate the necessity data in the database.

When the database is described in RFD format, if the user input query, the proper data extraction and inference will be performed by stream reasoning. In addition, if we introduce the reasoning system such as CMS and AAR, the above inference will be possible.

Thus for the in database reasoning, the concept of the stream reasoning can be adopted and the framework of RFD format can be applied to excecuting the analogical mapping in the reasoning system such as CMS and AAR.

For IMDJ Hayashi proposed to create a database for accurate extraction and reuse of knowledge by converting scenarios with RDF [8]. Accordingly knowledge contained in scenarios can be described structurally in RDF description. This description can be used for knowledge recommendation system in which when a user inputs a new scenario on the Web browser, the system returns related knowledge retrieved from RDF stores. This mechanism is achieved by SPARQL which is a query language for RDF. Our final aim is to achieve abductive reasoning which can generate missing or new hypothesis in IMDJ.

5 Conclusions

In this paper I discussed the logical treatment of data in IMDJ. First, I reviewed my previous work which is abductive consideration of the performance in IMDJ. Then I proposed the consideration of the concept of the stream reasoning and the concept of relational data mining. Where relationships between data are considered in an inference. In fact, all data should be regarded as patterns that involve multiple relations. Thus for reasoning, it will be necessary to consider such relationships. For that the data format of RDF for the database description can be adopted. Then the ontology based data access can be introduced during abductive reasoning. Accordingly we can perform flexible inference in IMDJ's proposal generation. Then the possibility that for the in database reasoning, the concept of the stream reasoning can be adopted and the framework of RFD format can be applied to excecuting the analogical mapping in the reasoning system such as CMS and AAR was shown. This is a keypooint of this paper. By this frmework, very frexible reasoning in IMDJ can be acheived.

This is a preliminary discussion of the possibility of introduction of ontology based data access to IMDJ for inferences. In the next paper, more detailed discussion can be performed.

REFERENCES

[1] Abe A.: Abductive Analogical Reasoning, *Systems and Computers in Japan, Vol. 31, No. 1, pp. 11–19* (2000)

[2] Abe A.: The Role of Abduction in IMDJ, *Proc. of IJCAI2015 International Workshop on Chance Discovery, Data Synthesis and Data Market, pp. 59–64* (2015)

[3] Anicic D., Fodor P., Rudolph S., and Stojanovic N.: EP-SPARQL: A Unified Language for Event Processing and Stream Reasoning, *Proc. of WWW 2011, pp. 635–644* (2011)

[4] Balduini M., Calbimonte, J-P., Corcho O., Dell'Aglio D., and Della Valle E.: *Tutorial on Stream Reasoning for Linked Data*, ISWC 2014, `http://streamreasoning.org/events/sr4ld2014/` (2014)

[5] Džeroski S. and Lavrač N. eds.: *Relational Data Mining*, Springer Verlag (2001)

[6] Gebser M., Grote T., Kaminski R., Obermeier P., Sabuncu O., and Schaub T.: Stream Reasoning with Answer Set Programming: Preliminary Report, *Proc. of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, pp. 613–617* (2012)

[7] Gentner D.: Structure-Mapping: A Theoretical Framework for Analogy, *COGNITIVE SCIENCE, Vol.7, pp.155–170* (1983)

[8] Hayashi T. and Ohsawa Y.: Knowledge Structuring and Reuse System Using RDF for Supporting Scenario Generation, *Proc. of 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2015), Vol.60, pp. 1281–1288* (2015)

[9] Kakas A. C., Kowalski R. A. and Toni F.: Abductive Logic Programming, *J. of Logic and Computation, Vol. 2, No. 6, pp. 719–770* (1992)

[10] McDermott D. and Doyle J.: Non-monotonic logic I, *Artif. Intell., Vol.13, pp.41–72* (1980)

[11] McDermott D.: Non Monotonic logic II : Non-monitonic modal logic, *J. of ACM, Vol.29-1, pp.33–57* (1982)

[12] Nagano S. and Kozaki K. eds.: Special issue "Linked Data and semantic technology", J. of the Japanese Society for Artificial Intelligence, , *Vol. 30, No.5* (2015) in Japanese

[13] Ng H. T. and Mooney R. J.: On the Role of Coherence in Abductive Explanation, *Proc. of AAAI90, pp. 337–342* (1990)

[14] Ohsawa Y.: *Innovators Marketplace on Data Jackets*, `http://www.panda.sys.t.u-tokyo.ac.jp/MoDAT/DJform.html`

[15] Ohsawa Y., Benson N. E. and Yachida M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, *Proc. Advanced Digital Library Conference (IEEE ADL'98), pp. 12–18* (1998)

[16] Ohsawa Y. and Nishihara Y. eds.: *Innovators' Marketplace*, Springer Verlag (2012)

[17] Ohsawa Y., Kido H., Hayashi T., and Liu C.: Data Jackets for Synthesizing Values in the Market of Data, *Proc. of KES2013*, pp. 013–709, Elsevier (2013)
[18] Okumura Y. : Open (Government) Data, Jurist, *#1464, pp. 45–52* (2014) in Japanese
[19] Peirce C. S.: Abduction and Induction, *chap. 11, Philosophical Writings of Peirce, Dover* (1955)
[20] Poole D., Goebel R. and Aleliunas R.: Theorist: A Logical Reasoning System for Defaults and Diagnosis, in *T*he Knowledge Frontier: Essays in the Representation of Knowledge (Cercone N.J., McCalla G. Eds.), pp. 331–352, Springer-Verlag, 1987.
[21] Reiter R. and de Kleer J.: Foundation of assumption-based truth maintenance systems: preliminary report, *Proc. of AAAI87, pp.183–188* (1987)
[22] Shimpo F.: The legal framework for the fundamental right to protection of personal data in EU, Jurist, *#1464, pp. 38–44* (2014) in Japanese
[23] Thagard P.: Explanatory coherence, *Behavioral and Brain Sciences, 12, pp. 435–502* (1989)

Study Chance Discovery in Temporal Linear Non-Transitive Logic with Agent's Knowledge

Vladimir V. Rybakov[1,2] [1]

Abstract. We consider CD interpreted in a framework of non-transitive temporal logic with elements of multi-agency. This approach is based on temporal linear intransitive models with elements of multi-agency, – agent's relation of the agent responsible for any given state. Non-transitive time imitates computational processes with incomplete information or parts of information lost while computational runs. We suggest a logical language and introduce a logic based at this language having the mentioned models as semantics. We develop a mathematical theory, give illustrating examples for CD based at these models and suggest an algorithm dealing with satisfiability problem.

Keywords: temporal logics, multi-agent logics, chance discovery, CD

1 Introduction

We aim to investigate elements of Chance Discovery (CD) interpreted in terms of temporal logic in case of non-transitive time. CD is a new active branch of computer science dealing with elements of uncertainty - in human reasoning or present data - a "chance" which is usually meant as a new event/situation that can be conceived either as an opportunity or as a risk in the future. CD, as a discipline (cf. Ohsawa and McBurney [23], Abe and Ohsawa [1]) initially appeared in Japanese school as a direction in Artificial Intelligence (AI) and Computer Science (CS). CD it works with construction and study various methods for discovering chance events (cf. e.g. papers [1, 2, 3, 4, 24, 25, 26, 13, 15, 26]).

Since a time ago (about 2007) some steps to develop a technique for study CD in terms of mathematical logic based on Kripke/Hintikka like models also were undertaken (cf. e.g. Rybakov [31]) Rybakov [28], [29, 30, 31, 32, 27]). In these papers interpretation of CD was primarily modeled via introductions of new logical operations based on possibility to happen (possibility to occur, etc).

In present time another elements from AI and logic were applied to CD area. For example, multi-agent systems (with autonomous or interacting, say competitive) were involved in the research. Technique and research outputs here are various, diverse and work well in many areas (cf. Nguyen et al [19, 20, 21], Arisha et al [5], Avouris [6], Hendler [14]). From area of mathematical logic, many techniques concerning multi-agency were applied (cf. [10, 11, 12]).

Multi-agent epistemic logics have found various applications in fields ranging from AI domains such as robotics, planning, and motivation analysis in natural language, to negotiation and game theory in economics, to distributed systems analysis and protocol authentication in computer security. These applications are based at the fact that intelligent agents must be able to reason about knowledge. Multi-agents' logics, in particular, appeared in the research about knowledge representation and reasoning about knowledge and beliefs (cf. for example, [10, 11, 12]). To mention some initial typical logical tools cf. ones from Brachman and Schmolze (1985, [8]), Moses and Shoham (1993, [16]), Nebel (1990, [18]), Quantz and Schmits (1994, [22]). Research of agent-decision oriented systems paid many attention to formal descriptions of the meaning of agents knowledge.

In present research we study ways of representation CD in the framework of non-transitive temporal logic with elements of multi-agency. The latter one will be represented as possession in any state its own accessibility relation - agent's relation of the agent responsible for this state. Non-transitive time imitates computational processes with incomplete information or parts of information lost while computational runs. We construct mathematical models and suggest an algorithm dealing with satisfiability problem. In the beginning we give some basic terminology and notation for reading the represented results.

2 Preliminaries

Our aim is to study the ways to interpret CD in the case when run of time (in applications - e.g. - computational threads, paths of information transition) is intransitive. We absorb and use traditional linear temporal logic as a departure point. Recall that the language of the Linear Temporal Logic ($\mathcal{LTL}$ in the sequel) extends the language of Boolean logic by operations $\mathbf{N}$ (next) and $\mathbf{U}$ (until). The formulas of $\mathcal{LTL}$ are built up from a set $Prop$ of atomic propositions (synonymously - propositional letters) and are closed under applications of Boolean operations, the unary operation $\mathbf{N}$ (next) and the binary operation $\mathbf{U}$ (until). The formula $\mathbf{N}\varphi$ has meaning: the statement φ holds in the next time point (state); the formula $\varphi\mathbf{U}\psi$ means: φ holds until ψ will be true.

Semantics for $\mathcal{LTL}$ consists of *infinite transition systems*

[1] School of Computing, Mathematics and Digital Technologies, Manchester Metropolitan University(1), Manchester M1 5GD, UK, and – part time: Institute of Mathematics, Siberian Federal University(2), Krasnoyarsk, Russia, e-mail: V.Rybakov@mmu.ac.uk

(runs, computations); formally they are represented as linear Kripke structures based on natural numbers. The infinite linear Kripke structure is a quadruple $\mathcal{M} := \langle \mathcal{N}, \leq, \text{Next}, V\rangle$, where $\mathcal{N}$ is the set of all natural numbers (for some extended versions of $\mathcal{LTL}$ – the set of all integer numbers Z); $\leq$ is the standard order on $\mathcal{N}$, Next is the binary relation, where a Next b means b is the number next to a. V is a valuation of a subset S of *Prop*. Hence the valuation V assigns truth values to elements of S. So, for any $p \in S$, $V(p) \subseteq \mathcal{N}$, $V(p)$ is the set of all n from $\mathcal{N}$ where p is true (w.r.t. the valuation V).

All elements of $\mathcal{N}$ are possible *states* (worlds), $\leq$ is the *transition* relation (which is linear in our case), and V can be interpreted as *labeling* of the states with atomic propositions. The triple $\langle \mathcal{N}, \leq, \text{Next}\rangle$ is a Kripke frame which we will denote for short by $\mathcal{N}$.

The truth values in any Kripke structure $\mathcal{M}$, can be extended from propositions of S to arbitrary formulas constructed from these propositions as follows:

$\forall p \in S\ (\mathcal{M}, a) \mid\!\vdash_V p \Leftrightarrow a \in \mathcal{N} \wedge\ a \in V(p)$;

$(\mathcal{M}, a) \mid\!\vdash_V (\varphi \wedge \psi) \Leftrightarrow (\mathcal{M}, a) \mid\!\vdash_V \varphi \wedge (\mathcal{M}, a) \mid\!\vdash_V \psi$;

$(\mathcal{M}, a) \mid\!\vdash_V \neg\varphi \Leftrightarrow not[(\mathcal{M}, a) \mid\!\vdash_V \varphi]$;

$(\mathcal{M}, a) \mid\!\vdash_V \mathbf{N}\varphi \Leftrightarrow [[(a \text{ Next } b) \Rightarrow (\mathcal{M}, b) \mid\!\vdash_V \varphi]]$;

$(\mathcal{M}, a) \mid\!\vdash_V (\varphi \mathbf{U} \psi) \Leftrightarrow \exists b[(a \leq b) \wedge ((\mathcal{M}, b) \mid\!\vdash_V \psi) \wedge$

$\forall c[(a \leq c < b) \Rightarrow (\mathcal{M}, c) \mid\!\vdash_V \varphi]]$;

For a Kripke structure $\mathcal{M} := \langle \mathcal{N}, \leq, \text{Next}, V\rangle$ and a formula φ with letters from the domain of V, we say φ is valid in $\mathcal{M}$ (denotation – $\mathcal{M} \Vdash \varphi$) if, for any b of $\mathcal{M}$ ($b \in \mathcal{N}$), the formula φ is true at b (denotation: $(\mathcal{M}, b) \Vdash_V \varphi$).

The linear temporal logic $\mathcal{LTL}$ is the set of all formulas which are valid in all infinite temporal linear Kripke structures $\mathcal{M}$ based on $\mathcal{N}$ with standard $\leq$ and Next.

3 CD in Linear Temporal Logic Based at Non-transitive Time

Our new non-transitive linear temporal logic is based at the following frames and models. The base set of these frames is N - the set of all natural numbers.

Definition 1. *A linear non-transitive frame is* $\mathcal{F} := \langle N, \leq, \text{Next}, (\bigcup_{i\in N}\{R_i\})\rangle$, *where for some fixed* $X \subset N$, $N = \bigcup_{i\in X, m_i\in X, i<m_i}[i, m_i)$, *and for all different* $i \in X$ *intervals* $[i, m_i)$ *have empty intersections. For any* i *from* X *and* j *from* $[i, m_i)$, R_j *is the standard linear order on* $[j, m_i]$. *For any* i *from* X, $t(i) = m_i$ *and* $Node(\mathcal{F}) := X$ *(nodes of* $\mathcal{F}$*).*

Notice that we consider $(\bigcup_{i\in N}\{R_i\})$ as not a binary relation, but as the infinite countable set of finite binary relations R_i. As earlier, we may define a model $\mathcal{M}$ on $\mathcal{F}$ by introduction a valuation V on $\mathcal{F}$ and extend it on all formulas as earlier, but for formulas of sort $\varphi\mathbf{U}\psi$ we define the truth value as follows:

Definition 2. *For any* $a \in N$*:*

$(\mathcal{M}, a) \mid\!\vdash_V (\varphi\ \mathbf{U}\ \psi) \quad \Leftrightarrow$

$\exists b[(aR_ab) \wedge ((\mathcal{M}, b) \mid\!\vdash_V \psi) \wedge \forall c[(a \leq c < b) \Rightarrow (\mathcal{M}, c) \mid\!\vdash_V \varphi]]$;

$(\mathcal{M}, a) \mid\!\vdash_V \mathbf{N}\varphi \quad \Leftrightarrow \quad [(a\ Next\ b) \Rightarrow (\mathcal{M}, b) \mid\!\vdash_V \varphi]$.

Definition 3. *The logic* $\mathcal{LTL}_{NT}$ *is the set of all formulas which are valid in any model* $\mathcal{M}$ *with any valuation.*

It is easy to see that the relation $\langle\bigcup_{i\in N}\{R_i\}\rangle$ is (generally speaking) non-transitive, though any R_i is linear and transitive.

Indeed, for example, let the model taken is when always $m_i := i + 3$. Then we have $1R_13$ and $3R_36$, but none of $1R_16$, $1R_36$ or $1\langle\bigcup_{i\in N}\{R_i\}\rangle 6$ holds.

The action of the relation $\langle\bigcup_{i\in N}\{R_i\}\rangle$ may be interpreted as follows. The whole interval $[i, m_i]$ itself) may be seen as the complete interval of time which agent i remember. In any time point $i + k$ it might be the new agent, or the same as the old (previous) one - just those who inspect, but its name is referred to $i + k$. Relations R_j for $j \in [i, m_i]$ inherit accessible from i (interpretation is: j's remember to past exactly the same time as i itself, it was earlier but remember not bigger as the agent in point i, so the one in i is most knowledgeable.

Thus, we may interpret this approach as $i \in N$ is a time point and $[i, m_i]$ is the time interval available for the agent responsible for verification/reasoning in the time point i. The name of the agent in this case is anonymous - those one who is responsible for i. The CD framework may work in this formalization with reasonable preciseness – because it handles loosing of information – via non-transitivity of time.

Being based at this interpretation, we may consider *interpretations of various aspects of knowledge* E.g.

Examples :

$$(\mathcal{M}, a) \mid\!\vdash_V K\varphi \Leftrightarrow (\mathcal{M}, a) \mid\!\vdash_V [\mathbf{N}^m\varphi] \wedge [\mathbf{N}^{m+1}\neg\varphi] \wedge [\varphi\ \mathbf{U}\ \neg\varphi].$$

Here K acts to say that the knowledge coded by φ has been achieved m 'years' ago: φ held in observable past in m years from now, but it did not hold in $m + 1$'s year from now, and φ held true since m years ago until now.

$$(\mathcal{M}, a) \mid\!\vdash_V K_1\varphi \Leftrightarrow (\mathcal{M}, a) \mid\!\vdash_V \Box\neg\varphi\ \wedge \Diamond(\neg\varphi \wedge \mathbf{N}(K\varphi)).$$

Now K_1 determines that φ was wrong in all observable time in the past, but before it has been time interval of length m, when φ was true (so to say it was a local temporal knowledge).

$$(\mathcal{M}, a) \mid\!\vdash_V K_2\varphi \Leftrightarrow (\mathcal{M}, a) \mid\!\vdash_V \bigwedge_{i\leq k} \Box^i\neg\varphi\ \wedge \Diamond^k(\neg\varphi \wedge \mathbf{N}(K\varphi)).$$

Here K_2 says that φ was wrong in subsequent k 'remembered' intervals in time, but before it has been in the past a local knowledge for a time interval of length m. It is easy to imagine (even being based at these simple examples) the wide possibilities for the expression properties of knowledge in time perspective (which might be achieved via the assumption that time could be non-transitive).

CD may be illustrated in the suggested models as the possibility to find a state, where a statement φ is true. It is clear that this may be described by formula $CD\varphi := \top\mathbf{U}\varphi$.

In order to reason about agent's knowledge and CD in such framework we need tools, computational tools, for determination valid statements. In mathematical terms this means solution of decision problem for our logic. We solved this problem:

Theorem 1. *Logic $\mathcal{LTL}_{NT}$ is decidable; the satisfiability problem for $\mathcal{LTL}_{NT}$ is decidable: for any formula we can compute if it is satisfiable and if yes to compute a valuation satisfying this formula in a finite model of kind $\mathcal{F}(N(r,g))$.*

Thus, this theorem gives us a computational algorithm allowing to verify satisfiability of statements, or, vise versa, to verify that a statement is a law - the one true/valid in all models. This may be applied in CD framework for investigation if there is a model (if there is a chance to find this model) satisfying given specification described (formalized) by formula. That is a typical task for knowledge representation and data verification.

REFERENCES

[1] Abe A. and Ohsawa Y. Editors: Readings in Chance Discovery, International Series on Advanced Intelligence, 2005.

[2] Abe A. and Kogure K. (2006) *E-Nightingale: Crisis Detection in Nursing Activities.* In: Chance Discoveries in Real World Decision Making, pp. 357–371.

[3] Abe A. and Ohsawa Y. (2007). *Special issue on chance discovery. KES Journal*, 11(5), pp. 255-257.

[4] Abe A., Hagita N., Furutani M., Furutani Y. and Matsuoka R. (2008) *Exceptions as Chance for Computational Chance Discovery*, KES2008, pp. 750–757.

[5] K. Arisha, F. Ozcan, R. Ross, V.S. Subrahmanian, T. Eiter and S. Kraus. *Impact: A platform for collaborating agents*, IEEE Intelligent Systems 14, 1999 (2), pp. 64-72.

[6] N.M. Avouris. *Co-operation knowledge-based systems for environmental decision-support*, Knowledge-Based Systems 8 (1995), (1), pp. 39-53.

[7] Barwise J. *Three Views of Common Knowledge.* - In Vardi (Ed.). *Proc. Second Confeuence on Theoretical Aspects of Reasoning about Knowledge* (1988), San Francisco. Califounia, Morgan Kaufmann, 365 - 379.

[8] Brachman R.J., Schmolze J. G., *An overview on the KL-ONE knowledge representation system.* - Cognitive Science, 9(2), 1985, 179 -226.

[9] Dwork C., and Moses Y. *Knowledge and Common Knowledge in a Byzantine Environment: Crash Failures .* - Information and Computation, Vol. 68 (1990), No. 2, 156 - 183.

[10] Fagin R., Halpern J., Moses Y., Vardi M. *Reasoning About Knowledge* - Book, The MNT Press, Cambridge, Massachusetts, London, England, 1995, 410 pp.

[11] Kifer M, Lozinski L., *A Logic for Reasoning with Inconsistency.* - J. Automated Deduction, Vol 9 (1992), 171 - 115.

[12] Kraus S., Lehmann D.L. *Knowledge, Belief, and Time.*- Theoretical Computer Science, Vol. 98 (1988), 143 - 174.

[13] Hahum K. S. (2000) *The Window of Opportunity: Logic and Chance in Becquerel's Discovery of Radioactivity*, Physics in Perspective (PIP), Birkhäuser Basel, Volume 2, Number 1, pp. 63 – 99.

[14] J. Hendler. *Agents and the semantic web*, IEEE Intelligent Systems, 16 (2001) (2), pp. 30-37.

[15] Magnani L. (2008) *Abduction and chance discovery in science*, International J. of Knowledge-Based and Intelligent Engineering Systems, (Volume 12).

[16] Moses Y, Shoham Y. *Belief and Defeasible Knowledge.* - Artificial Intelligence, Vol. 64 (1993), No. 2, 609 - 322.

[17] Neiger G., Tuttle M.R. *Common knowledge and consistent simultaneous coordination.* - Distributed Computing, Vol 5 (1993), No. 3, 334 - 352.

[18] Nebel B., *Reasoning and Revision in Hybrid Representation Systems.* - Lecture Notes in Computer Science, vol. 322 (1940) Springer Cerlag.

[19] Nguyen N.T. et al. (2008, Eds): Agent and Multi-agent Systems: Technologies and Applications. Proceedings of KES-AMSTA 2008. Lecture Notes in Artificial Intelligence 4953. Springer-Verlag.

[20] Nguyen N.T, Huang D.S. (2009): Knowledge Management for Autonomous Systems and Computational Intelligence. Journal of Universal Computer Science 15(4).

[21] Nguyen N.T., Katarzyniak R. (2009): Actions and Social Interactions in Multi-agent Systems. Special issue for International Journal of Knowledge and Information Systems 18(2).

[22] Quantz J., Schmitz B., *Knowledge-based disambiguation of machine translation.* - Minds and Machines (9996), 9: 99 - 97.

[23] Ohsawa Y. and McBurney P., Editors, *Chance Discovery (Advanced Information Processing),* Springer Verlag, 2003.

[24] Ohsawa Y. (2004) *Chance Discovery with Emergence of Future Scenarios.* KES 2004, pp. 11–12.

[25] Ohsawa Y. (2006) *Chance Discovery, Data-based Decision for Systems Design*, ISDA .

[26] Ohsawa Y. and Ishii M. (2008) *Gap between advertisers and designers: Results of visualizing messages*, International J. of Knowledge-Based and Intelligent Engineering Systems, Volume 12.

[27] Rybakov V.V. Linear Temporal Logic with Until and Before on Integer Numbers, Deciding Algorithms.- Computer Science - Theory and Applications, *Lecture Notes in Computer Science, Springer*, Vol. 3967, 2006, pp. 322 – 334.

[28] Rybakov V. Until-Since Temporal Logic Baed on Parallel Time with Common Past. Deciding Algorithms. In. Eds. S.Artemov, A.Nerode. Logical Foundations of Computer Science, LFCS 2007, Ney York, USA, *Lecture Notes in Computer Science*, 4514, 2007, pp. 486 – 497.

[29] Sergey Babenyshev, Vladimir V. Rybakov. *Describing Evolutions of Multi-Agent Systems.* – KES (1) 2009, Lecture Notes in Computer Science, 5711, Springer, 2009, pp. 38-45.

[30] Vladimir V. Rybakov *Linear Temporal Logic LTK_K extended by Multi-Agent Logic K_n with Interacting Agents.* – J. Log. Comput., Oxford Press, 19(6), (2009), pp. 989-1017.

[31] Vladimir V. Rybakov. *Algorithm for Decision Procedure in Temporal Logic Treating Uncertainty, Plausibility, Knowledge and Interacting Agents.* – IJIIT 6(1), (2010), pp. 31-45.

[32] Vladimir V. Rybakov. *Interpretation of Chance Discovery in Temporal Logic, Admissible Inference Rules.* – KES (3) 2010, Lecture Notes in Computer Science, 6278, Springer, 2010, pp. 323-330.

[33] Vladimir V. Rybakov. *Algorithm for Decision Procedure in Temporal Logic Treating Uncertainty, Plausibility, Knowledge and Interacting Agents.* – IJIIT, 6(1), 2010, pp. 31 - 45.

[34] A.T. Steinberg, A.T. A chance for possibility: an investigation into the grounds of modality. Doctoral thesis, UCL (University College London), 2011.

Vladimir V. Rybakov: Algorithm for Decision Procedure in Temporal Logic Treating Uncertainty, Plausibility, Knowledge and Interacting Agents. IJIIT 6(1): 31-45 (2010).

Preliminary Case Study about Analysis Scenarios and Actual Data Analysis in the Market of Data

Teruaki Hayashi[1] and **Yukio Ohsawa**[2]

Abstract. Decision making by using results of data analysis may reduce risks of actions and lead good performance in the field of real business. However, datasets are not obtained for free, and every action, such as analyses and acquisitions of data, requests the expenses. Therefore, it is important to form the hypotheses and make logical scenarios in advance from the information about datasets and analysis tools before the real actions. In this paper, based on the precondition, we conduct a workshop for observing a series of data utilization process, focusing on a gap between analysis scenarios and the actual analyses. Observing the actions of participants (discussion of data utilization and analysis, acquisition of data, or actual data analysis), we discuss the features of the gap between the scenarios and the real actions based on scenarios, and obtain suggestions for supporting data utilization.

1 INTRODUCTION

Data analysis has been conducted in order to reduce the risk of the real action, and to support meaningful decision making. In recent years, with the wide spread of personal devices and the development of sensors, data such as consumer purchasing history or personal moving records have been able to be acquired, which used to be difficult to obtain. Therefore, there are the growing demands for creating new products or services by collecting and taking advantage of data. However, datasets are not obtained for free, and every action, such as analyses and acquisitions of data, requests the expenses. Data acquired by putting a lot of time and effort might be traded at a high price, and data containing a large amount of personal information might be evaluated higher. In addition, the analysis results of complex data may be necessary to be considered from various angles and interpreted considering several analytical techniques. Moreover, in creating new businesses based on the results obtained from data analysis, it is necessary to consider the stakeholders and required resources, which requires further costs. In other words, it costs a lot to obtain a series of relevant information for decision making. Therefore, it is considered important to create feasible scenarios in advance by discussing what hypotheses should be tested for achieving goals, what kinds of datasets are necessary to obtain for testing hypotheses, or what kinds of tools should be applied to data.

However, there is a gap between analysis scenarios and the actual analyses. For example, when you start to test a hypothesis with data based on the analysis scenario and find the data unavailable, you have to achieve the goal in alternative ways, e.g., changing data for testing the hypothesis, or modifying the hypothesis based on available data. Such reworking of the product design and data analysis process has been recognized as a serious risk. An improper design process, such as not taking into account the cost and technical risk in the early stages of the product design, may cause the contradictions and conflicts in the latter stage of product design [1]. Thus, reworking in the design process is an important issue to be solved, and it is considered that the same problems may also occur in the planning process of data analysis.

In this paper, we conduct a workshop for observing a series of data utilization process, focusing on a gap between analysis scenarios and the actual analyses. Observing the actions of participants (discussion of data utilization and analysis, acquisition of data, or real data analysis), we discuss the difficult parts in the planning and the implementing process of data analysis.

The remainder of this paper is organized as follows. In Section 2, we show the methods for activating data utilization, Innovators Marketplace on Data Jackets and Action Planning. Section 3 describes experimental details of the workshop. Section 4 shows the results of actual data analyses in the workshop and the discussion about participants' data analysis processes. Finally, Section 5 concludes the paper with a brief review.

2 TECHNIQUES FOR SUPPORTING DATA UTILIZATION AND EXCHANGE

2.1 Data Jacket and Tool Jacket

Data Jacket (hereafter DJ) is a summary of datasets, that is, meta-data. We can understand the outline, format, or variables of data referring to meta-data on DJs, even if data itself is not open. DJ has been developed as a technique for sharing information of data and for considering the potential value of datasets, allowing data itself hidden [2, 3]. The idea of DJ is to share "a summary of data" as meta-data without sharing data itself, which enables stakeholders of data utilization to understand the contents of data and discuss the possible combinations of data, reducing the risk of data management cost and privacy. Moreover, the description rule of DJs enables datasets stored in different domains with various formats to be managed in a unified way. For example, when a pair of DJs shares the same variable labels, which is the name/meaning of a variable in a dataset, the datasets are possible to be combined through the shared variable labels, which makes it easy for decision makers to think of the combinations of datasets (Fig.1).

[1] Department of Systems Innovation, School of Engineering, The University of Tokyo, Japan, email: teruaki.hayashi@panda.sys.t.u-tokyo.ac.jp
[2] Department of Systems Innovation, School of Engineering, The University of Tokyo, Japan, email: ohsawa@sys.t.u-tokyo.ac.jp

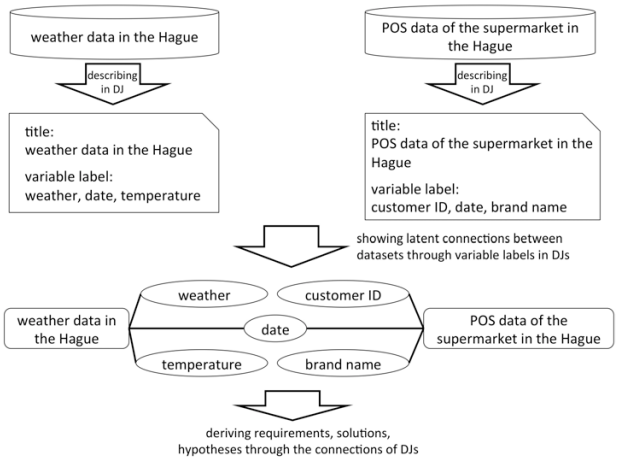

Figure 1. The flowchart of describing datasets as DJs and the process of deriving values from the combination of datasets

As there are different types of data in the world, there are various kinds of analysis tools. Tool Jacket (TJ) is the meta-data of analysis tools. TJ consists of a set of "the explanation of function", "input variable labels" and "output variable labels", which is the smallest unit for describing the attributes of analysis tools [4]. While DJ is the method for considering the usage and the potential value of datasets allowing data itself hidden for data owners, TJ is the method for evaluating the usage and the value of data analysis tools for data analysts (or for the inventors of data analysis tools).

Reading DJs and TJs, stakeholders can understand the contents of data and analysis tools. In addition, by introducing a visualization method such as KeyGraph [5] to DJs and TJs, it is possible to combine data and tools through the input or output variable labels, which reduces the cognitive load of the participants and supports to find the potential combinations of data and tools. Currently, we published the portal site for DJs and TJs [6], and collected about 1000 DJs and 100 TJs from data analysts and researchers (June 2016).

2.2 Innovators Marketplace on Data Jackets

Although there are the expectations to create value from the combination of data, Bollier mentions that combining data from multiple sources does not offer valuable insights, and even it makes the objective interpretation difficult [7]. Boyd and Crawford criticize that the size of the datasets is meaningless without taking into account the sample of each dataset, and suggest the importance of understanding the values of small data stored in different domains [8]. Therefore, it is suggested that the setting of hypotheses derived from the understanding and appropriate combination of dataset or tool is important in the discussion about data utilization. However, the number of combinations of data and tools is innumerable, and the number of hypotheses may increase exponentially according to the growth of combinations of data and tools. It is difficult for humans to take into account all the combinations and extract necessary hypotheses for solving problems from various pieces of data, because of the cognitive limitations of the rationality of individuals [9].

In order to support the human creativity of data utilization, Innovators Marketplace on Data Jackets (IMDJ) is proposed, which is the gamified workshop for discussing the data utilization. In the process of IMDJ, we introduce tools for data visualization, e.g., KeyGraph [5]. Creating a map on which shows possible combinations of DJs by connecting DJs through variables labels or feature words included in DJs, it reduces the cognitive load of participants and supports them to find potential combinations of datasets. Data owners provide their datasets as DJs, and participants of IMDJ (data owners, data users, and data analysts) create solutions for solving data users' problems stated as requirements. Through the communication among participants, data owners are expected to learn how to use their own data from a possible combination of DJs proposed by data analysts by reading a visualized map. Users are expected to learn how their requirements can be satisfied with proposed solutions. Through the process of IMDJ, participants start to negotiate for data exchange or buying/selling to create new businesses.

2.3 Action Planning

A scenario is a series of information derived from data or related knowledge. Human decision makers read, interpret, and perform actions according to the scenarios. Action Planning (AP) is a workshop method for creating scenarios, which is designed for formulating a discussion and leading atypical viewpoints and knowledge [10, 11]. Participants resolve conflicts and reduce risks of taking actions, through the process of externalizing and serializing related elements. AP has the following three phases for refining solutions proposed in IMDJ into scenarios.

(1) Requirement Analysis phase: Requirement Analysis means the phase of acquiring covert requirements which targets do not recognize yet, through the discussion. Starting from targets' overt requirements, participants find the background factors of overt requirements from objective data or suppositions with logical thinking, and clarify the covert requirements, potential stakeholders, and hypotheses to be tested.

(2) Element Externalization phase: Externalizing related knowledge or information for creating solutions from the requirements or solutions clarified in Requirement Analysis phase. Externalized knowledge includes cost, time to realize solutions, resources (technologies, equipment, budget, and data/tools), and stakeholders (targets, supporters, dissidents).

(3) Element Serialization phase: Serializing the knowledge and information considering relations among elements. Serialization means to find relationships among elements, and connect them to form a scenario by following particular rule sets (time management, business modeling, pseudo code, etc.).

The scenario generation of AP proceeds by filling sheets. By planning actions resolving conflicts among participants with various viewpoints, created scenarios help participants perform and reduce risks of taking actions, i.e., actual data analysis or creating businesses.

3 PRELIMINARY CASE STUDY

3.1 Overview

The purpose of this paper is to understand the features of the gap between the scenarios and the real actions based on scenarios, and to obtain suggestions for supporting data utilization. The essential precondition is that datasets are not obtained for free, and every action, such as analyses and acquisitions of data, requests the

expenses in the market of data. Therefore, we have to design a workshop to make participants consider analysis scenarios in advance from the information about datasets and analysis tools before the actual data analysis. In this study, we conduct a workshop as a preliminary case study assembling 18 researchers, businesspersons, and students in the department of engineering. We obtain and analyze the data about the parts where conflicts occur or where participants think difficult in the process of creating scenarios or analyzing data, by acquiring handwriting data using digital pens, observing the behaviors of participants, and the results of questionnaires after the workshop.

We set the main theme of this workshop as "creating solutions for detecting human errors caused by fatigue of organizations or individuals", which works as a common constraint to all participants of IMDJ and Action Planning. Three phases constitute the workshop.

First, participants discuss and consider requirements or solutions according to the given theme in IMDJ. In the phase of IMDJ, participants discuss the requirements which are the sources of hypotheses, and the solutions which are the possible combinations of data and tools for testing hypotheses.

Second, participants practice AP, and create analysis scenarios based on the requirements and solutions proposed in IMDJ. Participants make 6 teams (each team consists of 3 participants). Each team chooses highly evaluated solutions with requirements proposed in IMDJ. In the process of creating analysis scenarios in AP, each team hypothesizes the requirements according to the frameworks and items on the AP sheets. Describing the procedure of analyzing data in order to test the hypothesis using variable labels and pseudo code description, participants consider the possible set of variable labels as datasets and tools referring the combinations of data or tools proposed in IMDJ.

Finally, participants practice the data analysis phase based on the analysis scenarios created in the phase of AP. Given the actual data corresponding to the DJs, participants analyze data for testing their hypotheses based on analysis scenarios.

In the next section, we explain each phase in detail.

3.2 Process of IMDJ Phase

First of all, we explain the theme and the schedule of the workshop to the participants. After the brief lecture, participants practice IMDJ for 60 minutes, stating requirements or creating solutions based on the theme by combining DJs or TJs. In our previous experiments of IMDJ, because approximately 10 participants are appropriate for discussing the data utilization, we divide the 18 participants into two groups at random. IMDJ in this workshop is conducted as follows.

1. Preparing 39 DJs related to the theme in advance, and creating the scenario map by connecting the words shared among DJs, using KeyGraph [5]. Because we use KeyGraph only for reducing the cognitive load of participants by showing the possible connections of DJs, we do not discuss or compare the effect of the visualization tool in this paper. Fig.2 shows the scenario map. The titles of visualized 39 DJs and their IDs are presented in Appendix (Table 3).
2. Participants state their requirements from the standpoints of data users (about 15 minutes).
3. Participants create solutions by combining DJs on the map for satisfying related requirements, from the standpoints of data analysts (about 45 minutes). When DJs on the map are insufficient to create solutions, participants are allowed to add DJs/TJs found from Data Jacket Store [12], which is the retrieval system of DJs/TJs, and use them for creating solutions.
4. When data users' requirements are satisfied with data analysts' solutions, data users can pay toy money to data analysts for evaluating solutions.
5. After repeating from step 2 to step 4 for 45 minutes, the highly evaluated solutions are chosen as the candidates for AP.

In this workshop, 3 solutions were selected from each group (6 solutions in total), and were refined as analysis scenarios in the process of AP.

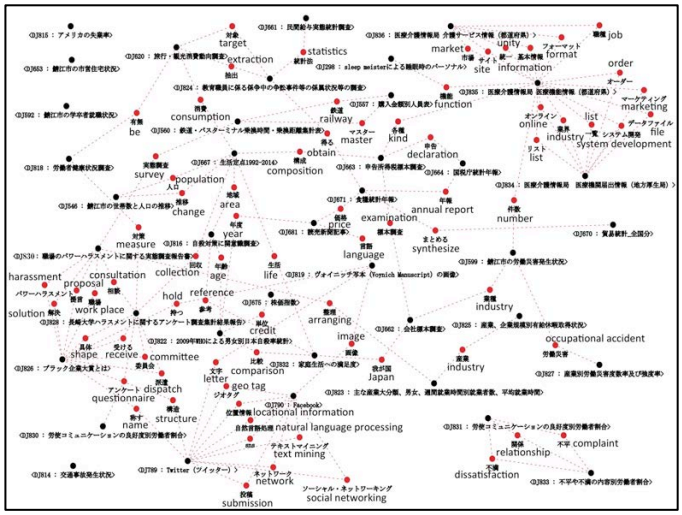

Figure 2. The scenario map using KeyGraph (black nodes represent DJs, and red nodes show the keywords included in DJs. Each DJ is interlinked with related Data Jackets via correlated words shown as red nodes, e.g., the words in outlines of data, or in variable labels).

3.3 Process of AP Phase

We introduce the frameworks for supporting to generate analysis scenarios in AP. AP in this workshop consists of 3 parts, Requirement Hypothesization Part, Pseudo Code Description Part, and Data/Tool Externalization Part (Fig.3).

Requirement Hypothesization means to refine a requirement to a description which is possible to discuss using data. For example, "We want to ease traffic congestions in the Olympics in Tokyo" is the requirement, which is difficult to discuss with data. When we refine this requirement to the hypothesis, e.g., "Interesting places may be different from the nationality of tourists", it comes to be easy to consider how to test this hypothesis with data. In this case, it may be tested by using "tourist data including the places for sightseeing and their nationalities". In other words, Requirement Hypothesization is the process of deriving a testable hypothesis from an ambiguous requirement, by considering the backgrounds, situations, or stakeholders of requirements according to the frameworks and items on the AP sheet. Pseudo Code Description is the part to consider the logical combinations of variable labels and analysis tools. Pseudo code is originally a description of the operating principle of an algorithm in natural language, which is often used for sketching out the structure of the program before the

actual coding takes place. By introducing Pseudo Code Description in the process of generating analysis scenarios, it comes to be possible to discuss the logic for obtaining expected results from the combinations of input/output variable labels and analysis tools, without actual datasets. Data/Tool Externalization is the process to describe the data or analysis tools required for data analysis, which is written when participants notice the data or tools through the process of hypothesizing or pseudo coding.

In this workshop, in order to obtain the data about the parts where conflicts occur or where participants think difficult in the process of creating scenarios or analyzing data, we acquire handwriting data using digital pens device (Anoto Digital Pen DP-201 by Hitachi Maxell Ltd.). The time stamps of handwriting can be obtained with digital pens, and we compare the thinking time in creating analysis scenarios, the results of data analysis, and the evaluation values of each team.

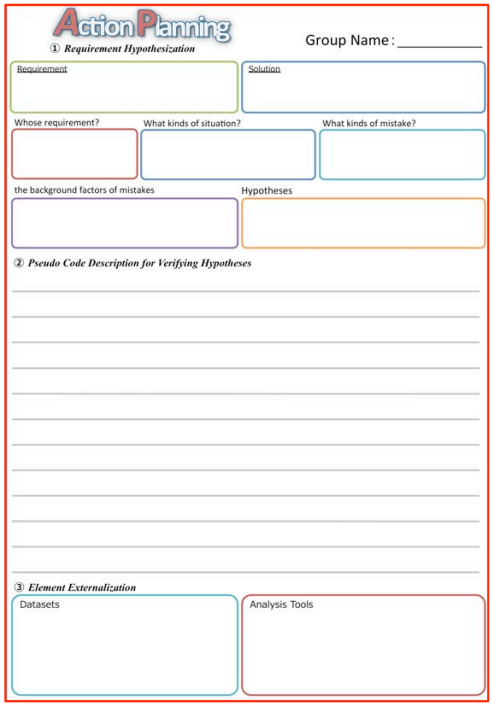

Figure 3. Action Planning sheet consists of three parts, Requirement Hypothesization Part, Pseudo Code Part, Data/Tool Externalization Part.

3.4 Process of Data Acquisition and Analysis Phase

In the phase of data analysis, participants analyze data and try to test hypotheses based on the analysis scenarios created in the phase of AP for 60 minutes. We hand out the actual data corresponding to the DJs and the laptop computers to each team. When participants notice that they need additional data or tools for testing their hypotheses during the analysis, we allow them to access and use the data or tools available on the Web. When participants have to change their hypotheses in the process of data analysis, we allow them to modify their hypotheses.

After the practice, each team presents their results orally by showing their outcomes of analysis for 5 minutes, and has question-and-answer sessions after their presentations for 5 minutes. After all the presentations, we have the mutual evaluations by voting. Each participant votes once for a team which carried out the best data analysis. In this workshop, we got 19 votes from 18 participants and one organizer.

4 RESULT AND DISCUSSION

In this section, we show and discuss the results obtained by observing the participants' performances in the workshop. In IMDJ phase, one group created 18 requirements and 12 solutions, the other created 22 requirements and 18 solutions, and there was no great difference between groups. Table 1 shows the 6 highly evaluated solutions with requirements created in IMDJ, which was refined as analysis scenarios in AP phase.

In AP phase, we divided participants into 6 teams, and got 6 analysis scenarios. Through the process of AP, the analysis scenarios of 5 teams (A, B, C, D, F) changed from the solutions in IMDJ. For example, the solution of team B (Table 1) changed into "Calculation of the mismatch rate and dissatisfaction analysis from the number of care workers in order to reduce the number of human errors in the field of nursing care". On the other hand, the solution of team E did not change. The previous studies of AP mention that the ideas created in IMDJ (or Innovators Market Game, which is the former of IMDJ) may change through the process of AP. It can be said that the result in this workshop fits the fact of previous studies [10, 11].

Table 1. Highly evaluated solutions, requirements, and combined DJs (DJs corresponding to ID number is shown in the Appendix).

Team	Requirement	Solution	ID number of Combined DJs
A	We want to reduce the number of suicides.	Preventing the suicides surveying the suicide rate statistics separated by gender.	822
B	We want to reduce stresses by making the processes and the tasks clear in the early part of projects.	The system to distribute the tasks appropriately among project members.	813, 832
C	We want to know which is the most stressful business field.	Comparing the labors' health, living areas, the number of suicides with the companies that exploit its employees.	818, 822, 835, 827
D	-	Extracting the keywords of harassment.	828, 830
E	-	Examining the correlation between stock prices and paid vacations.	675, 815
F	We want to learn the habitual patterns of businesspersons who have less stress.	Examining the relationship between stresses and daily life log.	298, 667

We obtained handwriting data of teams using digital pens, and visualized the tracks of handwritings after the workshop. Fig.4 is an example of handwriting track of Action Planning in team B. As is shown in Fig.4, there are some parts where tracing back occurs from Externalization of Data/Tool to Pseudo Code Description. The transition from Externalization of Data/Tool to Pseudo Code Description indicates that a member of team B noticed required variable labels in discussing the part of Externalization of Data/Tool, and added the related information to the prior parts.

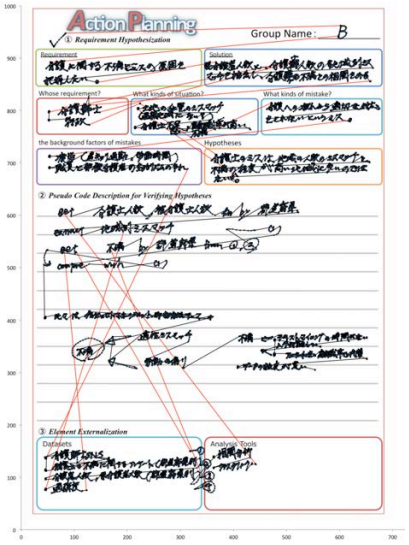

Figure 4. One example of a visualized writing track of Action Planning in team B

Table 2 shows the transitions on the AP sheets of each team and the evaluation values of the results of the actual data analysis based on analysis scenarios. The evaluation values of team B and F are relatively lower than other teams. As shown in the transition on the sheet, team B frequently moves from Pseudo Code Part and Data/Tool Externalization Part. In other words, team B frequently noticed the inconsistencies in their scenario, and resolved the conflicts. According to the previous study, allowing the cost of time and reworking, the quality of scenarios may develop [13]. However, because the time was limited to 60 minutes for creating analysis scenarios in this workshop, team B may not have enough time to improve their scenario. On the other hand, because team F took too long time in Requirement Hypothesization Part, it can be said that they may not have enough time to discuss Pseudo Code Part and got the low evaluation.

Table 2. The transitions on the Action Planning sheet and evaluation values of the results of data analysis

Team	Sequence	Evaluation Value
A	1 → 2	6
B	1 → 3 → 2 → 3 → 2 → 3 → 2	0
C	1 → 2 → 3	4
D	1 → 3 → 2	1
E	1 → 2	8
F	1	0

1: Requirement Hypothesization Part
2: Pseudo Code Part
3: Data/Tool Externalization Part

Moreover, we compare how much time each team considered the part of Pseudo Code, which is the most important part of actual data analysis, with evaluation values. We adopt an experimental method of Ikegami & Ohsawa [14] and Hayashi & Ohsawa [13] for calculating thinking time using a digital pen (TT). They define thinking time as a period that participants stop their handwriting for over 5 seconds (1). t means a time stamp obtained by a digital pen.

$$TT = \sum (t_i - t_{i-1}) \ (if \ t_i - t_{i-1} > 5 \ seconds) \ (i \in \mathbb{N}) \quad (1)$$

As a result, we found the positive correlation between thinking time in Pseudo Code Part in AP and evaluation values of the results of data analysis (Fig.5), where the correlation coefficient is 0.84. This result suggests that a team which spends enough time of discussion in considering pseudo code for actual data analysis, may obtain a result which is evaluated higher.

In addition, in the questionnaires after the workshop, most of the participants answered that Pseudo Code Part is hard to consider, and the technique for supporting the description is required. According to the examination of the results of this workshop and the answers of participants, it is necessary to propose the technique for supporting Pseudo Code Description in the future.

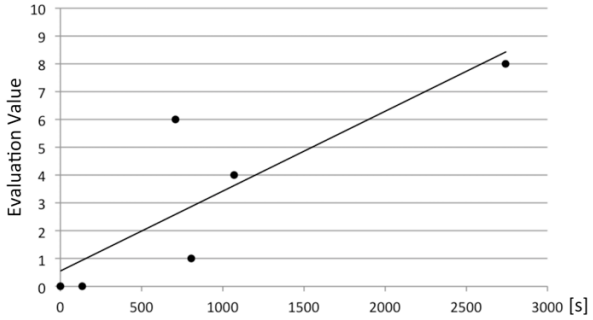

Figure 5. The correlation between thinking time in Pseudo Code Part in AP and evaluation values of the results of data analysis

5 CONCLUSION

In this paper, we held a workshop as the preliminary case study for observing a series of data utilization process, and discussed the difficult parts in the planning and the implementing process of data analysis. The essential precondition of this workshop is that datasets are not obtained for free, and every action, such as analyses and acquisitions of data, requests the expenses. Therefore, we design the workshop to make participants consider analysis scenarios first, i.e., considering the hypotheses and making logical scenarios in advance from the information about datasets and analysis tools before the actual data analysis. Through the observation, we found that there was the correlation between the thinking time of Pseudo Code Part and the evaluation values of the results of actual data analysis. This result suggests that if Pseudo Code Part (the structure and the process of data analysis) is discussed sufficiently before the actual coding takes place, a result of data analysis may work well and be evaluated higher.

In addition, some participants answered that the most difficult part in the phase of creating data analysis scenarios was Pseudo Code Part in the questionnaire. In our future work, it is necessary to support participants to describe pseudo code.

APPENDIX

Following table is the list of Data Jackets used for creating the scenario map using KeyGraph, which are visualized on the map of Fig.2 (translated in English).

Table 3. The list of Data Jacket introduced in the workshop

DJ ID	DJ Title
298	Personal sleep data using Sleep Meister
546	The number of households and population in Sabae City
557	Transportation fee of census
560	Rail and bus terminal transfer time and transfer distance t
592	Graduates employment situation of Sabae City
599	Occupational Accidents in Sabae City
620	Survey data of travel and tourism consumption
653	Municipal housing situation of Sabae City
661	Survey data of salary statistics
662	Survey data of private companies
663	Survey data of declared income tax
664	Annual report of National Tax Agency
667	Survey data of daily life from 1992 to 2014
670	Trade statistics
671	Annual report of food statistics
675	Stock index
681	Articles of Yomiuri News Paper
789	twitter text data
790	facebook text data
814	Data of traffic accidents
815	Unemployment rate in the US
816	Survey on suicides
818	Survey data of labors' health
819	The images of Voynich Manuscript
822	WHO survey on suicide rate statistics separated by gender in Japan in 2009
823	Weekly average working hours separated by gender and categories of business
824	Investigation of pending status of disputes cases relating to education staff
825	The number of taken paid holidays by categories and sizes of business
826	Black companies (the companies that exploit its employees) Prize
827	Frequency rate and severity rate of labor accidents separated by categories of business
828	Survey data of harassments in Nagasaki University
829	Salary and years of service separated by gender and categories of business
830	The report on workplace harassments
831	Data of communications between management and workers
832	Satisfaction with family life
833	Complaint and dissatisfaction data of labors
834	Administration data of medical institutions separated by local Bureau of Health and Welfare
835	Functional medical information separated by prefectures
836	Care service information separated by prefectures

ACKNOWLEDGEMENTS

This study was partially supported by JST-CREST, and JSPS KAKENHI Grant Number 16J06450. Also, we would like to thank all the staff members of KKE (Kozo Keikaku Engineering Inc.) for supporting our research.

The present research was partially supported by the Leading Graduates Schools Program, “Global Leader Program for Social Design and Management,” by the Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

[1] T. Koga and K. Aoyama, “Product Behavior and Topological Structure Design System by Step-by-step Decomposition,” ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, pp.425-437, 2004.

[2] Y. Ohsawa, C. Liu, T. Hayashi, and H. Kido, “Data Jackets for Externalizing Use Value of Hidden Datasets,” 18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, Vol.35, pp.946-953, 2014.

[3] Y. Ohsawa, H. Kido, T. Hayashi, C. Liu, and K. Komoda, “Innovators Marketplace on Data Jackets, for Valuating, Sharing, and Synthesizing Data,” Knowledge-based Information Systems in Practice, Smart Innovation, Systems and Technologies, Tweedale, W.J., Jain, C.L., Watada, J., and Howlett, R. (eds), Springer International Publishing, Vol.30, pp.83-97, 2015.

[4] T. Hayashi and Y. Ohsawa, “Meta-data Generation of Analysis Tools and Connection with Structured Meta-data of Datasets,” 3rd International Conference on Signal Processing and Integrated Networks, 2016.

[5] Y. Ohsawa, N.E. Benson, and M. Yachida, “KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor,” Proceedings of Advanced Digital Library Conference, pp.12-18, 1998.

[6] Data Jackets Site, [Online]. Available from: <https://sites.google.com/site/datajackets/>, [Last access 17th June 2016].

[7] D. Bollier, “The promise and peril of big data,” Communications and Society Program, The Aspen Institute, Washington, DC, 2010.

[8] D. Boyd and K. Crawford, “Critical Questions for Big Data,” Information, Communication & Society, Vol.15, No.5, pp.662-679, 2012.

[9] H.A. Simon, “A Behavioral Model of Rational Choice,” The Quarterly Journal of Economics, vol. 69, pp. 99-118, 1955.

[10] T. Hayashi, and Y. Ohsawa, “Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game plus Action Planning,” International Journal of Knowledge and Systems Science, Vol.4, No.3, pp.14-38, 2013.

[11] T. Hayashi and Y. Ohsawa, “Relationship between externalized knowledge and evaluation in the process of creating strategic scenarios,” Open Journal of Information Systems, Vol.2, No.1, pp.29-40, 2015.

[12] T. Hayashi, and Y. Ohsawa, “Knowledge Structuring and Reuse System Design Using RDF for Creating a Market of Data,” 2nd International Conference on Signal Processing and Integrated Networks, pp.566-571, 2015.

[13] T. Hayashi, Y. Ohsawa, “Comparison of Conflict Resolution Behavior and Scenario Generating Process in Group and Individual by Handwriting Process Analysis,” Intelligent Decision Technologies, pp.1-9, 2016.

[14] K. Ikegami and Y. Ohsawa, “Modeling of Writing and Thinking Process in Handwriting by Digital Pen Analysis,” International Conference on Data Mining 2014, The Workshop of Designing the Market of Data, pp.447-454, 2014.

Mini session on Data Curation in the Market of Data

organised by Teruaki Hayashi
Department of Systems Innovation, School of Engineering, The University of Tokyo, Japan

Abstract

The usage of data and the exchange of data have come to be discussed and active in the Market of Data. However, a method to discover data related to his/her actions from the accumulation of massive small data in the world has not been established. Therefore, there is the gap between users who want to discover data related to their interests and data holders who want to provide their data. In this special session of EWCDDS16, we discuss "Data Curation in the Market of Data", which is the management of data collected from various sources including not only public data but also private data stored by private companies or individuals. In the first half of this session, we show some practical examples of the gaps between data users and data holders, and discuss the necessity of Data Curation. In the second half, we conduct the brief workshop to consider of a method for selecting valuable data by extracting latent requirements from the data users' requests. **This session is open to all the participants in ECAI2016.**