一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
AI2017-33 (2018-02)

グラフに基づくエントロピーを用いた変化予兆検知

ーデータジャケットによるデータ市場が生んだアルゴリズム：その X ー

大澤 幸生

†東京大学工学系研究科システム創成学専攻 〒113-8653 東京都文京区本郷 7-3-1
E-mail: ohsawa@sys.t.u-tokyo.ac.jp

あらまし データ市場の目的は，単なる取引だけではない．取引を動機づけとしながら，データ駆動イノベーションの場となる潜在力を発揮することがデータ市場の持続性に繋がる．本論文では，データ市場モデルのひとつ Innovators Marketplace on Data Jackets から生まれた計算モデルの一つとしてグラフに基づくエントロピーを紹介し、その地震分析，オンラインQ＆A分析，POS データによるマーケティングに関する例を示す．このように基礎的な計算基盤モデルを生み出すことは，データ市場に本来備わっている，本質的な効果である．

Graph-based Entropy for Detecting Precursors of Changes

-- A product of innovators marketplace on data jackets --

Yukio Ohsawa

†Dept. Systems Innovation, School of Engineering,

The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
E-mail: ohsawa@sys.t.u-tokyo.ac.jp

Abstract. Innovators Marketplace on Data Jackets is a platform for creating not only use scenario of data, but also basic models for computation in data analysis/visualization. Graph-based entropy is an index representing the structural diversity of items, one of the computational models recently created in IMDJ. This proposed index is computed on the co-occurrence graph of items in the data, where the distribution of items to subgraphs is reflected. The variation of this index corresponds to the separation and the combination of events, regarded as effects of latent dynamics or external forces of the target system. In this paper, we show the application of GBE to the analysis of earthquakes, analysis of online Q&A, and marketing with POS data.

Keywords graph-based entropy, change detection, data jackets, innovators marketplace

1 Introduction

Innovators Marketplace on Data Jackets (IMDJ[1]) is a workplace for inventing use scenarios of data via the combination of data described on data jackets (DJ), that are digest information about existing or expected data. Requirements for data-driven innovation have been collected from the side of users of data and of data use scenarios, to which participants proposed solutions to combine technologies for data analysis (including tools with AI) and techniques for data collection.

In this paper, we focus on a solution to the following requirements in IMDJ so far: (1) collect credible and persuasive information, (2) detect items in retail stores that explain essential changes of customers' behaviors, and (3) recommend things to consume reflecting detected changes. The solutions eventually obtained, after action planning[2], were (1) to collect high impact information from text (2) to apply tangled string [3] and (3) to provide timely recommendation fitting consumers' new preference. For all these solutions, we have been developing tools for the detection of changes that may *not* be linked to any common topic(s) of the society [4]. This is because we had found on the way that users of information in (1) and consumers or customers in (2) and (3) do not necessarily act with sharing a common topic.

In our vision, creative activities can be classified into generative phase (making a plan, designing a product, cooking dinner, etc.) and exploratory phase (walking in the retail store, browsing for useful information, etc., for both

assigning a meaning to the generated product in the generative phase or for collecting elements for generating a new product) [5]. Decisions can be the output of either of the two phases. For example, when one buys food for cooking, one browses the variety of shelves in a retail store until finding interesting items, in the exploratory phase before deciding to buy [6]. Or, one may decide to sell what he generated in the generative phase. In the exploratory phase, one may be influenced by peripheral information such as famous sports players having used a certain item in the market [7], or externalize latent interests via resonance without response [8].

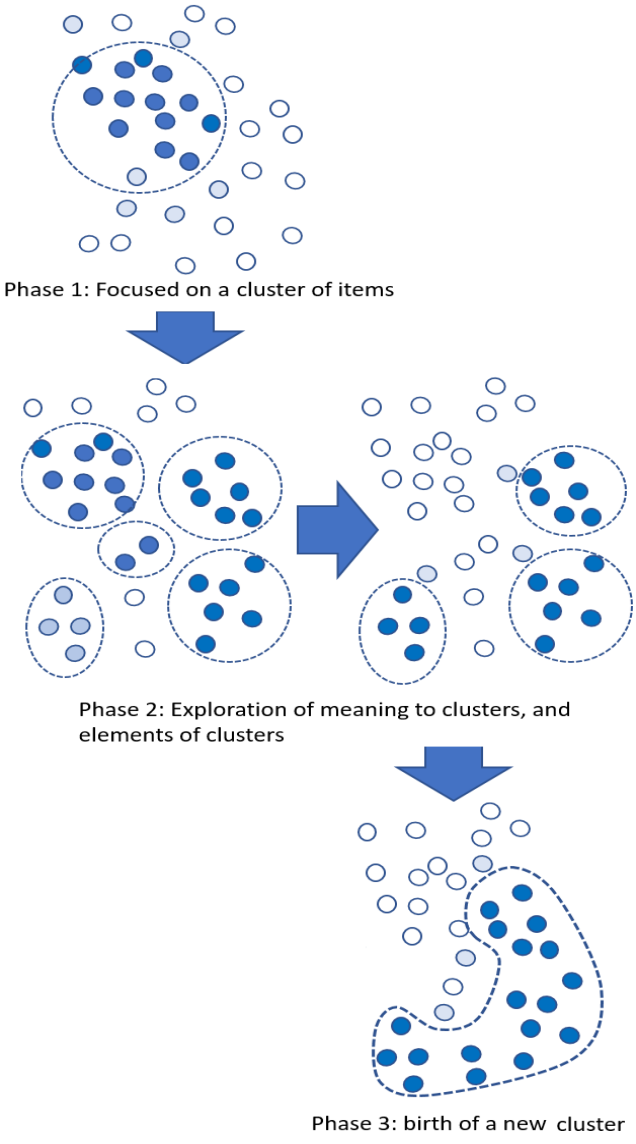

Fig. 1. Separation, movement, and fusion i.e., the combination of clusters, corresponding to the exploration phase and the generation phase.

All in all, external events and information from outside of given data affect consumer's behavior with providing a variety of contexts, in the exploratory phase, until the generative phase starts where the collected information is used. The exploratory phase may be called again, due to the requirement for additional elements or of the meaning of generated products.

2 A graph-based model of real-world dynamics

In Figure 1, each node represents an item (a product sold in the market, a word spoken or written, an active fault that quaked, etc), and the closeness of two nodes represents their tendency to co-occur in the data. Each area given by a closed curve shows a cluster of frequent items that co-occur in the data. A node colored the more densely shows the higher frequency of the corresponding item. On Fig.1, my model of dynamics in the target world is summarized as follows:

(Phase 1) Events occur in a certain focused part of the world, corresponding to a certain context.

(Phase 2) Small subgraphs (illustrated as clusters without edges in this figure) are created, corresponding to just-born contexts which may disappear soon, due to transient interaction with various items in exploration.

(Phase 3) A new and large subgraph is created, corresponding to the generative phase.

From the aspect of marketing science, variety seeking [6] roughly corresponds to exploration, and the choice can be positioned as a result of the decision that is an output of exploration or generation.

3 Graph-based entropy

In social sciences, entropy has been used as a measure of variety seeking tendency, uncertainty and variety in the behaviors of markets, societies, and organizations [9, 10, 11]. In digital image processing, entropy has been used in detecting of contours and abrupt changes [12]. Entropy has been also used in the analysis of accidents in traffics and computer networks [13, 14]. Below let us here model the process in the last section, as the changes in graph-based entropy defined below in Eq.(1).

$$\mathrm{Hg} = \sum_j p(\mathrm{subgraph}_j) \log p(\mathrm{subgraph}_j). \quad (1)$$

Here $p(\mathrm{subgraph}_j)$ is the probability of each item (or event) in data to belong to $\mathrm{subgraph}_j$, on a membership defined

specifically for the target problem. Ini Eq.(1), subgraph$_j$ is the j-th subgraph, that is typically a cluster i.e., a connected graph, where edges are defined by the distance between epicenters (for earthquakes) or the co-occurrence (for POS or documents) of items or words. And, p(subgraph$_j$) means the probability that an earthquake (basket/sentence) occurred in the area (interest/context) corresponding to the cluster of epicenters (items/words). Let us call Hg in Eq.(1) graph-based entropy (GBE).

If we compare GBE with the dynamic topic models [15], the GBE model is freed both from the constraint in DTM that each of the latent topics at a time should succeed one of the topics at times before, by allowing clusters to be separated or combined to form new topics fitting the changing contexts in the external world. Thus, the meaning of temporal derivative of Hg differs from the changing of topics in DTM in that the latent dynamics in contextual restructuring can be reflected flexibly.

4 Results for three kinds of data

4.1 Analysis of the Earthquakes

The example in Fig.2 shows the variation of Hg(t) for the distribution of earthquakes, where t denotes each year in [1983,2015]. In this case, each of the 1600 square meshes of 0.1 [deg] (ca 11km) in the square of the diagonal (lat. N33.0, lng. E133.0)-(37.0,137.0), nearly corresponding to the Kansai area, is represented by a node (item) in a graph for computing Hg(t). That is, the overall graph nearly corresponds to the Kansai area. In this graph, the co-occurrence represented by an edge is defined by 1 if two items touch via a vertex or an edge (i.e. if they are neighbors), otherwise 0.

The result in Fig.2 (a) shows t's of local minima correspond with peaks i.e., local maxima, of the energy of earthquakes. As shown in (b), the concentration of epicenters to a large merged cluster (closely located dots) co-occur with the sudden decrease of gH in 1995.

4.2 Online answers to questions

Here, GBE is applied to answers to questions in of Yahoo! Chibukuro (2004-04-01 through 2009-04-07). This is a collection of on-line answers from people with knowledged viewpoints, to questions of people in the public. Fig.3 shows the variation of Hg(t), for t in [0,15] meaning the set of the 100(t−1)-th through the 100t−1-th messages, extracted on query “被災地” or “被爆” or “被ばく” or “被曝”, meaning the exposure to radiation and the exposed lands. 50 words relevant to lifestyles of people, living in districts with nuclear power plants, have been taken as items (nodes in the graph), to obtain a graph of co-occurrence of words, on which gH(t) has been obtained.

The curves in Fig.3 are the results for 2 of 5 randomly selected parts from the extracted answers. The details of analysis will be presented elsewhere, here finding a local maximum or minimum (mostly minimum) of GBE tends to correspond with a high-impact event relevant to the query.

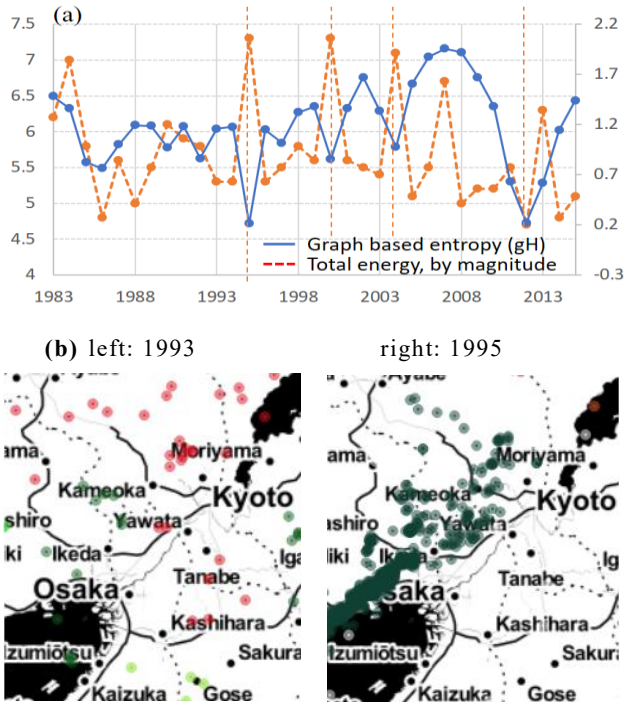

Fig. 2. The yearly variation of GBE in (a), computed on their distribution to clusters of epicenters plotted as in (b).

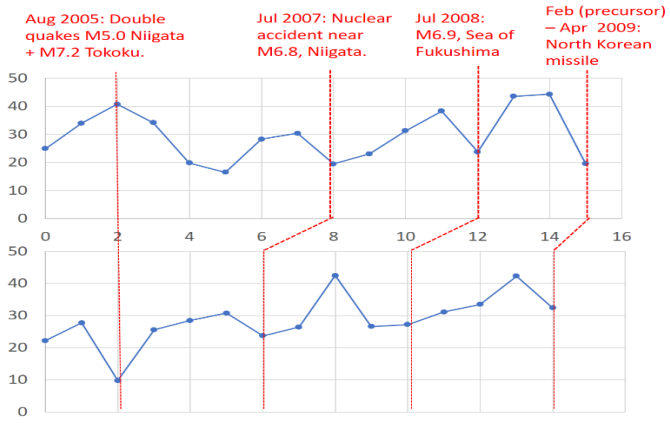

Fig. 3. Hg(t) of 1500 messages retrieved for a query about nuclear hazards, from two parts of Yahoo Chiebukuro.

4.3 Analysis of the variation of consumers' motivation from POS data

The target in this case is the purchase history in a supermarket. Fig.4 shows Hg(t), for t denoting each of 52 weeks in one year from 2014/7/1 through 2015/6/30, where the items of breads or buns are taken as nodes in the graph for computing GBE. The co-occurrence is counted if two items co-occur in the same basket of customers. The compared line "change of order" shows the sum of changes in the rank of top 10 items for two serial weeks. For example, if a croissant is the 5th in a week and this rank is switched next week with raisins role of the 7th, the change of order is 4 because the ranks of two items changed by 2.

A cluster in the left of panels for the 34th through 36th weeks in Fig.5 has got merged with other items in the 37th week, and made a popular cluster by the 40th week. This change corresponds to the change in the graph-based entropy that peaks in the 37th week as in Fig.4. Here again, we put the details of computation and comparison with baselines out of scope in this paper.

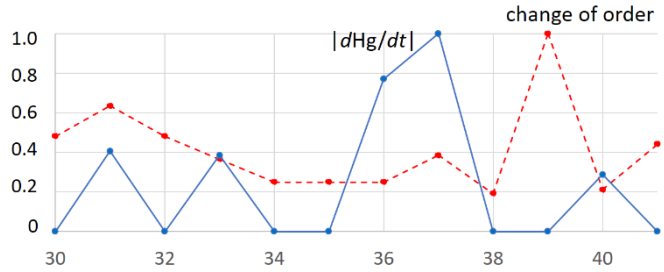

Fig. 4. The variation in GBE: the change peaks in the 37th week, whereas the change in the order of item sales frequency peaks two weeks later.

5 Discussions

All the results above contribute to the understanding of the dynamics in the target real world, by visualizing to someone interested in the requirements shown in the introduction i.e., (1) collect credible and persuasive information, (2) detect items in the retail store that explain essential changes of customers' behaviors, and (3) recommend things to consume, reflecting detected changes.

As mentioned in 4.1, if GBE (i.e., Hg) is applied to earthquakes, the concentration of epicenters to a merged cluster tends to co-occur with the sudden decrease of Hg. Furthermore, as the year 2012 as $\min_{t \text{ in } [1983,\ 2015]}$ Hg(t), the change in the value of Hg sometimes precede the real activation of earthquakes. These tendencies match with an expectation that an earthquake occurs in a region where multiple quaking parts of the lithosphere starts to interact, due to latent dynamics that may trigger a future earthquake. Such a precursor with explanation is credible for a user, and forms a piece of persuasive information required in (1). An improvement of GBE, where the noise of background earthquakes is omitted and the convenience for analysis is reinforced, is to be published elsewhere.

The result in 4.2 can be interpreted that peoples' interests converge into the topic about a new event if the impact of the event is high, but will be released from the convergence or be involved in the flood of information from various people interested in the event. An exception such as Aug 2005 in the upper of Fig.3 can be interpreted as that the flood of information appeared at the same time as the event because the log of quick (i.e., small data on the) convergence has been absorbed in other four of the five parts of the data. Thus, GBE is an essential tool for detecting the flood of information, in order to reduce the risk to be overwhelmed by a flood composed of untrustworthy information --- fitting the requirement of (1).

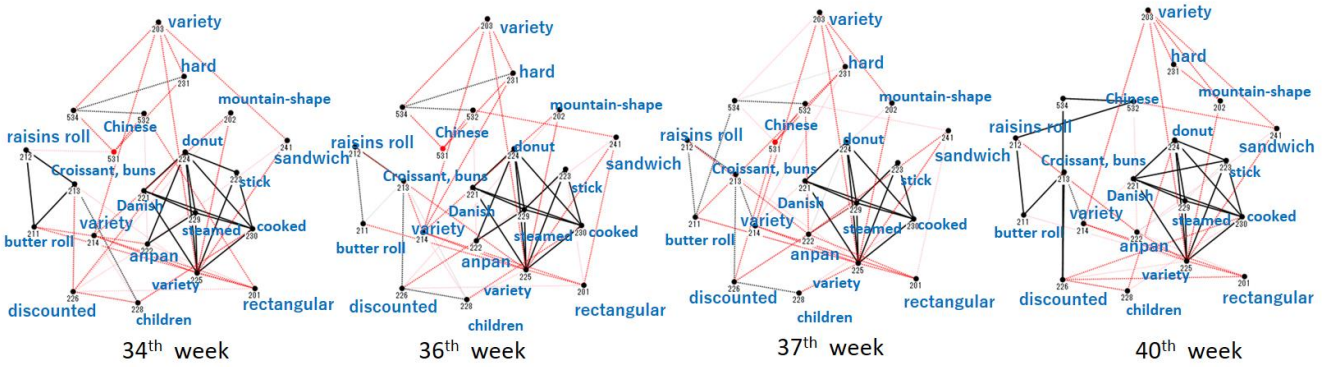

Fig. 5. The graphs corresponding to weeks from the 34th till the 40th for the category "bread and buns" in a retail store of supermarket.

In the result of 4.3, the 37th week corresponds to 17th through 23rd of March 2015, that means the end of one academic year in Japan. According to two supermarket managers, this change is interpreted to reflect the season of parties in Japan, such as for cherry blossoms under which Japanese people come together to eat and drink. The detection of the contribution of Danish and croissant to such a structural change, obviously relevant to requirement (2), was an "unexpected success." Danish and croissants are coming to be popularly recommended items in parties as we find in, for example, https://trends.google.co.jp/trends/explore?q=%22croissant%20party%22, fitting the requirement of (3). Such a seasonal trend may be learned if POS data for a sufficient number of years were available, but here note that we used only one-year data.

6 Future work as Conclusions

The definition of GBE is quite general and abstract in this paper, and the real setting of algorithms and parameters in each application are intentionally pended because such details are in the scopes of forthcoming papers showing technical aspects including comparison with other methods for change detection and/or for modeling the latent dynamics. Instead of showing these details, here I showed the changing of this index corresponds to the separation and the combination of items and/or events, regarded as pieces of evidence of unknown latent dynamics or external forces of the target system. This principle is widely applicable.

The quite abstract level of GBE is, however, useful in planning the next step. For example, suppose we have multiple worlds, such as domains of sciences and businesses. Each domain has a value of GBE, corresponding to the variety of beliefs of belonging experts. Let us assume a network of connections, corresponding to interdisciplinary communication among these domains as in Fig.6. In each domain shown as a cell in a dotted closed curve, the variety of contexts (interests if each node represents a word or a product, areas such as active faults if each node represents an epicenter, etc.) is quantified as

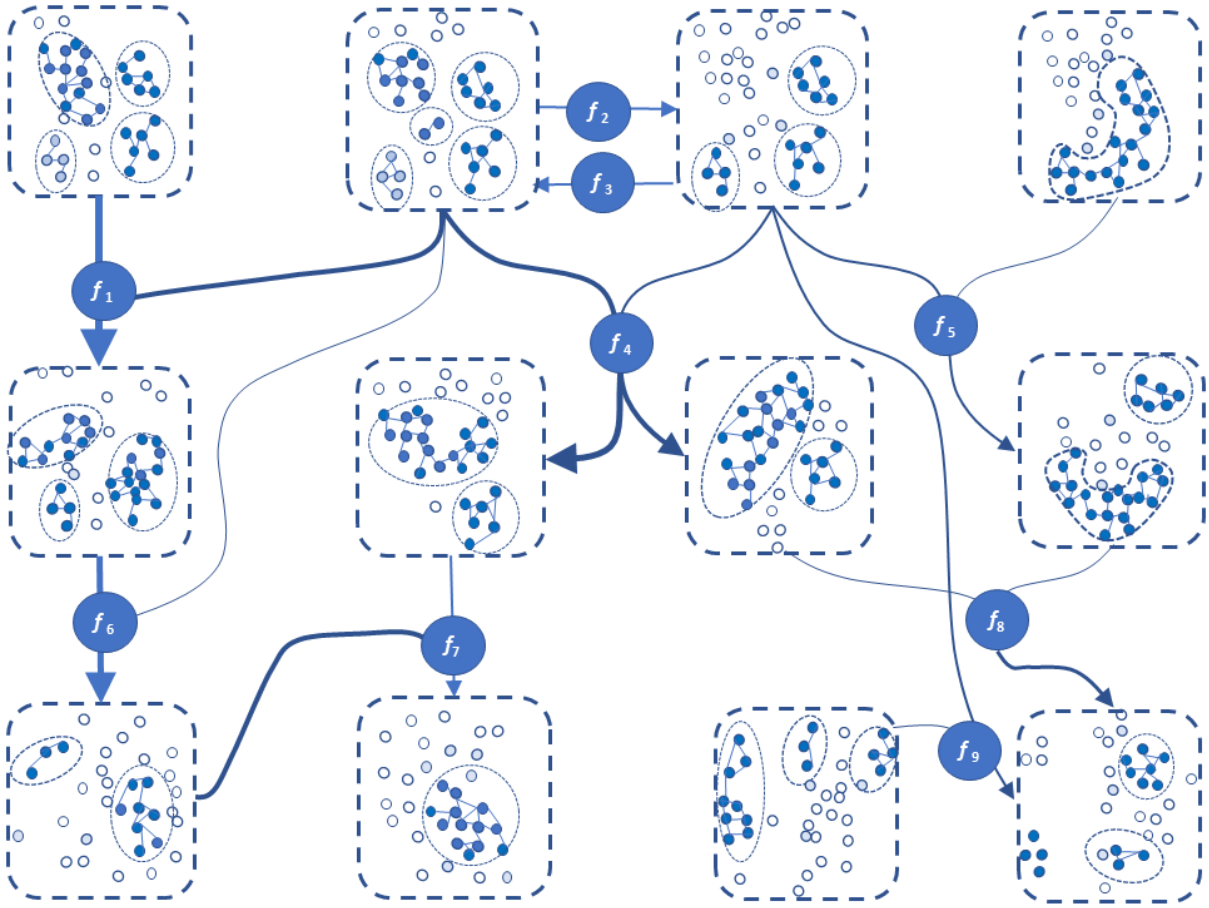

Fig. 6. Network of Demons (NODEM): a neural network model of inter-domain interactions, where activation functions (f_i) and weights are to be learned from the values of Hg(t) for t's. Figure source from http://www.panda.sys.t.u-tokyo.ac.jp/ohsawa/.

a value of GBE. Via the connections, the changes in GBE propagate and the domains in the network change the state of belief mutually. Thus, by learning the weight of each connection represented by an arrow, we can expect to simulate future trends of sciences and businesses by detecting the outstanding changes simulated. The transformation of a set of inputs into an output, given by *fi* for each integer *i*, can be an activation function used in a neural network. However, the special feature of this network is that the emergence of a high-level concept or dynamics is represented by the construction of a complex structure, that is a connection of multiple subgraphs, quantified by a reduced value of GBE. If this concept affects other domains, the GBE of the affected domains is expected to be reduced. Thus, we call this a Network of Demons (NODEM), borrowing from the Maxwell's demon who reduces the entropy of a system it manages. Because a system should increase entropy if left without external forces, the existence of external forces can be modeled in NODEM on the observed values of GBE. By this, NODEM is expected to be used for designing a market, for predicting peoples' social behaviors, or for simulating earthquakes. For the next step, we are calling for collaborators with us, both from businesses and sciences (including students).

Acknowledgement This study has been supported technically by JST CREST, JSPS Kakenh JP 16H01836, JP 16K1242 and Kozo Keikaku Engineering Inc. The earthquake data for 3.1 has been provided by the open data site of Japanese Meteorological Agency http://www.data.jma.go.jp/svd/eqev/data/bulletin/hypo.html), the text data for 3.2 in Yahoo' Chiebukuro provided by the National Institute of Informatics in Japan, POS data for 3.3 has been provided by Kasumi Inc. Also, we appreciate participants of Data Jacket Promotion WG, who share and promote the methodology in the book and articles [16,17,18] (in Japanese).

References

[1] Ohsawa, Y., Kido, H., Hayashi, T., and Liu, C., Data Jackets for Synthesizing Values in the Market of Data *Procedia Computer Science* 22, pp. 709-716 (2013)

[2] Hayashi, T. and Ohsawa, Y., Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game plus Action Planning, *Knowledge and Systems Science* (2013)

[3] Ohsawa, Y., and Hayashi, T., Tangled string for sequence visualization as fruit of ideas in innovators marketplace on data jackets, *Intelligent Decision Technologies*, DOI: 10.3233/IDT-150251 (2016)

[4] Kasuga, A., Ohsawa, Y., Non-Conformity Detection in High-Dimensional Time Series of Stock Market Data, *The 29th Int'l Conf. IEA/AIE*, Morioka, Japan (2016)

[5] R. A. Finke, T. B. Ward and S. M. Smith, *Creative Cognition: Theory, Research, and Applications*, A Bradford Book (1996)

[6] Kahn, B.K., Consumer variety seeking among goods and service, *J. Retailing and Consumer Services* 2, p.139-148 (1995)

[7] Petty, R.E., and Cacioppo, J.T., and Schumann, D., title *J. Consumer Research* 10, pp.135-145 (1983)

[8] Ohsawa, Y., Matsumura, N., and Takahashi, N., Resonance without Response: The Way of Topic Growth in Communication, *Studies in Computational Intelligence* (SCI) 30, 155-165 (2006)

[9] Saviotti, P.P., Information, variety and entropy in technoeconomic development, *Research Policy* 17 (2), pp.89-103 (1988)

[10] Alexander, P.J., Product variety and market structure: A new measure and a simple test, *J. Economic Behavior & Organization* 32 (2) pp. 207-214 (1997)

[11] Grigoriev, A.V., The Evaluation of Variety of Market Structure Using the Entropy Indicator, *J. Siberian Federal University. Humanities & Social Sciences* 4 (2012 5) 521-527

[12] Fuchs,M., Homann, R., and Schwonke, F., Comparison of images taken by different sensors, *Geoinformatics FCE CTU* 3, 28-28 (2008)

[13] Nychis, G., Sekar, V., Andersen, D.G., Kim, H., Zhang, H.: An Empirical Evaluation of Entropy-based Traffic Anomaly Detection. *Proc. of the 8th ACM SIGCOMM Conference on Internet Measurement.* 151 – 156, New York, NY, USA (2008)

[14] Winter, P., Lampesberger, H., Zeilinger, M., & Hermann, E.. On detecting abrupt changes in network entropy time series. *IFIP International Conference on Communications and Multimedia Security,* 194-205, Springer Heidelberg (2011)

[15] Blei, D., and Lafferty, J.D., Dynamic Topic Models, *Proc. International Conference on Machine Learning* 23, pp.113-120 (2006)

[16] 大澤幸生（編著），早矢仕晃章，秋元正博（著），久代，中村，寺本（著）：データ市場：データを活かすイノベーションゲーム，近代科学社（2017）

[17] .大澤・早矢仕：データ市場を活かす国家的戦略を （政策シンクネット Evidence Vol.003） http://thinknet.org/evidence/2017092712.html

[18] 早矢仕晃章，大澤幸生，データジャケットを用いた異分野データ連携，人工知能学会誌 3月号掲載予定（2018）