The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Entropy-based Knowledge Space Visualization for Data-driven Decision Support

Qi Wang Teruaki Hayashi Yukio Ohsawa

Department of Systems Innovation, School of Engineering, The University of Tokyo

This paper describes an entropy-based visualization for extracting utilization value of data in data-driven decision support. There is an increasing demand for data utilization from different domains to create new businesses or solve existing problems. Many researches focus on extracting similar data to create business value. However, data utilization is a more human-interactive field thus reusing data utilization knowledge is a more important task to define data relations in the field of solving multidisciplinary problems. That is to say, knowing the application fields of data is vital in promoting data utilization for problem-solving. Thus, this paper reuses the knowledge of data utilization to re-design data relations in visualization of knowledge space to assist data utilization recognition and decision-making processes. The Innovators' Marketplace on Data Jackets is employed to collect data utilization knowledge, and we use data co-occurrence entropy to reflect people's recognition process on data utilization because co-occurrence between data in utilization knowledge is an informative prior supporting for decision-making. The experimental results show the proposed entropy-based visualization method suppresses baseline methods in the data-utilization proposal numbers and qualities, which verifies proposed method assists people's recognition and decision-making processes in data utilization.

1. Introduction

1.1 Background

In recent years, data utilization is becoming an increasing demand to dig out the potential value of various data. However, the lack of data market to communicate and collaborate with data transaction stakeholders is the main obstacle of data utilization. In other words, the market of data needs a platform to facilitate data-driven decision-making to explore the utility value of data. The platform is supposed to help open and provide data information in the market to potential users while considering privacy and confidential policies; data users and data holders, as well as data analysts, can negotiate to discuss the potential use value of data; data transactions and collaborations can be conducted at reasonable conditions by negotiation. Under such data market platform, the value of data can be negotiated in a reasonable condition facilitating data utilization, thus to build a data-driven society. However, there is still a difficulty of data sharing policy in the first stage of the platform.

1.2 Literature Review

1.2.1 Data Jackets

On the one hand, data users hardly discover suitable data for their problem-solving without disclosure of datasets and utilization method; on the other hand, data providers such as individuals and companies are not willing to open and share their datasets because of privacy and security issues of sharing data. To overcome these problems, a novel conception of Data Jacket (hereafter DJ) is proposed to promote data exchanges and transactions [Ohsawa 13]. DJ, an analogue of CD jackets in CD shops, is the metadata of the dataset which illustrates the necessary information without disclosing data detailed contents. By this technique, stakeholders can share the information of data and discuss the potential use approaches of data without sharing data itself. Besides, DJ transfer unstructured datasets into structured data information. Thus, it is an ideal method for stakeholders to discuss the value of data while reducing the risk of security management cost and business loss, encouraging stakeholders to share information and combine latent related datasets from different domains to solve problems.

1.2.2 Innovators' Marketplace on Data Jackets

However, although the data sharing problem has been solved by adopting DJ, stakeholders still need a platform to negotiate and collaborate on data utilization. Inspired by the collaborative game of Innovators' Market Game [Ohsawa 13] and its extension [Wang 13], Innovators' Marketplace on Data Jackets (hereafter IMDJ) combining data mining and visualization is supposed to detect potential data utilization in one stage [Liu 13, Ohsawa 15]. The participants in the innovative game can communicate or think by themselves to explore potential connections of data through visualization graphs. They can recognize the utility of different domains of data and negotiate to exchange or collaborate with the data.

In the process of IMDJ, data users will elaborate their problems as requirements; analysts are expected to provide insights into data utilization for problem-solving by looking at a given DJ visualization graph which showing DJ relations and connections; data-utilization scenario proposals as solutions will be evaluated by others, and transactions can be negotiated in the final phase.

1.2.3 Entropy Distance

Information entropy was introduced by Claude Shannon to measure the average amount of information value produced by a stochastic source of events [Shannon 48]. The measurement of information entropy [Arndt 04] described as S, the sum of negative logarithm of the probability mass function multiples corresponding possible data value, which formulates as follows:

Contact: Qi Wang, The University of Tokyo, Hongo7-3-1, Bunkyo, Tokyo, Japan, 07044242665, wangqi940519@gmail.com

$$S = -\sum_{i} P_i ln P_i \tag{1}$$

Where P_i represents the probability of ith event occurs [Pathria 11]. The base of the logarithm is usually taken 2 in Shannon entropy as a measurement in bits.

The basic idea of entropy is described as a measurement of the unpredictability of the event, or in other words, the average information content the event conveys. When the event occurs with low-probability, it conveys more information than when the data has a higher-probability value [Martin 11]. The information entropy is the expected value of the random variable which is the amount of information carried by each event [Stone 14].

Recent years, the concept of information entropy has been extended to many research fields and used in many applications. For instance, in the field of graph theory [Tsai 08, Dehmer 08] and probability theory such as application in the Markov process [McCallum 00]. A similar application like this work, Liu discussed the relations of entropy, distance measure, and similarity measure with the fuzzy sets [Xuecheng 92]. [Shi 16] measures distance applied information entropy, and [Parker 16] uses entropy information to measure distance compared with Jaccard distance.

2. Method Proposal

2.1 Problem Statement

In the previous researches, few works focus on visualization techniques supporting data-driven decision-making processes, though many related works concentrate on data mining algorithms or retrieval systems. This paper will contribute to the visualization algorithm based on the literature review of IMDJ platform to support data utilization activities. As mentioned above, from the previous experiments, participants pay attention to the distance of DJs to define their relations. As a result, there is a demand to re-define the distance of DJs to further assist the data-driven decision-making process. The restriction of distance computation is supposed to enable subjects to understand the relations of DJs better and discover the hidden value of so-called data solutions by combining useful data quickly and feasibly. Information hidden in the distances between DJs reveals the latent relations of data.

2.2 Hypothesis

From the perspective of attention mechanism, human beings focus on only parts of the whole work at once. Thus, the short distance between DJs will enable participants to think their relations at first. This paper utilizes information theory integrating visual analytics to construct a knowledge space to support users recognition processes and data-driven decision-making processes. First of all, the information entropy proposed by Shannon is defined as the average amount of information produced by a stochastic source of data, and it can measure the uncertainty of an event or the diversity. The higher value of entropy is, the more unpredictable the event is. On the contrary, low entropy means a certain probability of the occurrence of an event. In this case, the event is the usage of DJ in solutions. The prior probability of co-occurrence of two DJs in solutions can compute their entropy which defines their relations. The high entropy reveals the independence of two DJs which means the uncertainty of data usage. On the other hand, low entropy represents certain relations of two DJs whether high dependence or low-frequent co-occurrence. In either case, the certainty of co-occurrence will assist users to think data relations easier and quickly exclude useless data from solutions.

DJs with high entropy mean they have complicated relations and their co-occurrences in solutions are unpredictable and need more information make decisions. That is to say, the information entropy value of DJs represents relations of data revealed in the visualization graph. The low entropy implicit little amount of information which is needed to decide whether the DJ is useful in the solution or not. Contrarily, high entropy means it is so complicated that needs more information to make decisions. Thus, the DJs with small entropy should be grouped to be a cognition cluster. By this mechanism, participants firstly consider simple things which are easy to think out solutions for data utilization. And then they can extend to the DJs which are far away from it to deeply consider the utilization method of DJs. In our hypothesis, the knowledge space constructed is supposed to support the decision-making process and enable to build a data-driven society exploring the true value of data.

2.3 Data Knowledge Structure

Based on our hypothesis above, the whole procedures can be described as follows. First of all, we define the data-utilization knowledge structure. Data can be combined to create new value to solve social or personal problems which are the proposed requirements by data users. The process of problem-solving in IMDJ is called data-utilization scenario proposals or so-called solutions. The data analysts can discuss the value of data by negotiating with the solution providers to find satisfactory solutions and conclude deals in a reasonable condition. The knowledge of data-utilization can be defined as the follows [Hayashi 15]:

$$\cup_{1,\cdots k} \text{Data Jackets} \xRightarrow{combined} \text{solutions} \xRightarrow{satisfy} \text{requirements}$$

From the previous workshop experiments, the data-utilization knowledge has been collected and stored in Data Jacket Store, and we can use this knowledge to construct a knowledge space for data-utilization. There are several steps: first of all, extracting the knowledge structure needed to construct a knowledge matrix. The matrix is a binary matrix with rows of solutions and columns of DJs. The entry is 1 when the target DJ is used in the corresponding solution. Otherwise, the entry is 0.

2.4 Relation Matrix Generation

After obtaining the relation matrix of DJs with solutions, a distance matrix between DJs can be computed by conditional information entropy of pairwise DJs. To be specific, for shortening thinking process of subjects thus accelerating the making decisions process in IMDJ, it plays crucial roles in the circumstances of both the interdependence relations between DJs as well as the low-frequency co-existence of DJs representing hints for data utilization chances. On the other hand, DJs with uncertain applications or usages in different fields of problem-solving can be regarded as noninformative prior thus the distance between them should be far from certain ones. Under the hypothesis, we design the distance calculation method as:

$$R(X,Y) = \left(-P(X|Y)\log_2 P(X|Y) - (1-P(X|Y))\log_2(1-P(X|Y))\right) + \left(-P(Y|X)\log_2 P(Y|X) - (1-P(Y|X))\log_2(1-P(Y|X))\right) \quad (2)$$

$$P(X|Y) = \frac{|X \cap Y|}{|Y|} \quad (3)$$

$$P(Y|X) = \frac{|X \cap Y|}{|X|} \quad (4)$$

Where X, Y are the binary sets whether DJ_x and DJ_y are included in the solution respectively. And P is the conditional probability of pairwise DJs co-existence in the same solutions under the condition of the frequency of single DJ existence, or the complement of the possibility vice versa.

2.5 Visualization Configuration

Given that the computed distance matrix between DJs, a dimensionality reduction method of multidimensional scaling (MDS) is adopted. Since MDS is an algorithm to reduce dimensions preserving the relations of data points in the original data space, it applies to many other works such as psychology to explicate the disparities of cognitions in the map [Leeuw 88]. In this paper, we use MDS to distribute the DJs in the 2-dimensional Euclidean space to easily captured by subjects and enhance the insights cognition of them. As for our distance matrix does not meet the Euclidean space, a majorization algorithm of Scaling by MAjorizing a COmplicated Function (SMACOF) is employed to reach a minimum of the stress loss function [Leeuw 11]. Thus the configuration of coordinates of DJs can be used to plot the visualization map.

3. Experiments and Results

3.1 Experiment Procedures

Based on the hypothesis, we conduct experiments to evaluate the distance visualization method implementing on DJs proposed above, to inspect and verify whether the distance computation method based on information entropy in data-utilization knowledge space is feasible for guiding the recognition process and play a role of decision-making assistance in IMDJ.

We expect to gather experimenters to conduct revised edition of single person IMDJ comparing with other visualization maps to see whether there is the difference between them. For the baseline distances calculation methods to draw the visualization maps, we choose two other methods. One is Jaccard distance which also a binary distance calculation method. The relation matrix of DJs can be the same one with our approach, but the computation steps are different corresponding to each method. Another distance calculation method is based on word vector space, say, the word2vec method in this case because there is a related word showing the advantages of word2vec distances between DJs during IMDJ.

The purpose of this experiment is to compare the performance of these three distance calculation methods of DJs revealing in the visualization graph on the results of solutions in personal IMDJ. Experiment subjects are separated into 3 groups, each of subjects conduct single experiment with different themes and methods for 3 times, each theme including 10 DJs. They are encouraged to raise proposals as much as possible in a limited 10 minutes.

Table 1. Illustration of Experiment Groups

	Dataset1: Olympics Preparation	Dataset2: Town Development	Dataset3: Regional Revitalization
Method1: Entropy Distance	Group1	Group3	Group2
Method2: Word Vector Distance	Group2	Group1	Group3
Method3: Jaccard Distance	Group3	Group2	Group1

Visualization graphs given to subjects only show the DJ numbers as figure1 in order to eliminate bias. DJ details can be checked in the appendix by participants.

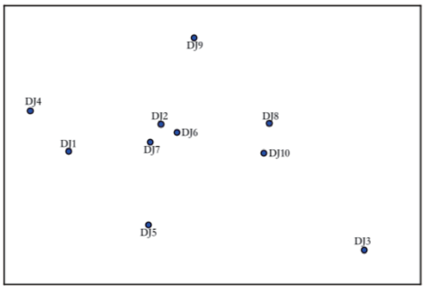


Figure 1. Example of Visualization Graph

3.2 Quantitative Evaluation

As a result, 375 pieces of data have been collected from the experiments. The data is described as the knowledge flow as defined above: data-utilization scenario proposals drive from combinations of DJs.

Table 2. Results of Proposal Numbers

# of proposals	Method1	Method2	Method3	Total
Theme1	51	40	37	128
Theme2	50	38	36	124
Theme3	47	37	39	123
Total	148	115	112	375

*Method1: Entropy-based distance: Method2: Word vector distance: Method3: Jaccard distance

From the results we can see there is no statistic significant difference between method2 and method3 using Kruskal-Wallis test, whereas method1 differs from the method2 and other two methods. The total proposal number of method1 is greater than the other two methods. It verifies the effectiveness of the decision support function of the proposed method. To draw the conclusion, the proposed method of distance between DJs outperforms the other two methods on the number of proposals. That is to say, the hints of distance between DJs revealed in the visualization map assist subjects' cognition process, which is a decision making of whether pairwise DJs have a possibility to be combined for solving problems.

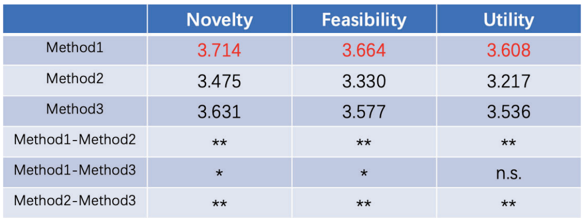
3.3 Qualitative Evaluation

To evaluate the quality of proposals, a third-party has been involved to rate on a five-point scale: 5 to 1 represents for excellent, good, average, below average and poor respectively. The three criteria are as follows:

(1 Novelty: do you think the proposal is a new idea or an existing one? (2 Feasibility: do you think the proposal can be realized by real actions or not? (3 Utility: do you think the proposal is useful or valuable if it been conducted/analyzed?

The results are shown in table3. From the statistic results, we can see there is a significant difference between method 1 and method2 in novelty, feasibility and utility, and method 1 differs from method3 in novelty and feasibility at the significance level of 0.05.

Table 3. Results of Significance Analyses



	Novelty	Feasibility	Utility
Method1	3.714	3.664	3.608
Method2	3.475	3.330	3.217
Method3	3.631	3.577	3.536
Method1-Method2	**	**	**
Method1-Method3	*	*	n.s.
Method2-Method3	**	**	**

n.s.: non significance *: P<0.05, **: P<0.01

The entropy-based distance algorithm can help trigger more novel and feasible insights for ideas. We can conclude that proposed data knowledge space influences the outcomes of data-utilization proposals. From the questionnaire analyses, we conclude that low co-occurrence DJs clusters will trigger more novel ideas because participants can notice rare but significant relations from the visualization distance. The emphasis on infrequent pair-wise DJs is the reason for high quality and novelty proposals.

4. Conclusion and Future Work

In this work, we construct the data knowledge space and re-define the data relations revealed by distances of DJs in the visualization graph. To design distance algorithm assisting thinking and decision-making processes, information entropy is adopted. The distance between two DJs is defined as their addition of conditional entropy of their co-occurrence in solutions. The experimental results show outperformance of proposed method in quantitative and qualitative evaluation. Thus we conclude that proposed entropy-based visualization of knowledge space can support for the data-driven decision-making in the market of data.

For the market of data, our method is supposed to facilitate data stakeholders' recognition process that can correctly grasp the value of data consisting of utilization approaches to using it. Furthermore, the data-driven society can be built as we reduce the uncertainty of business loss and waste opportunity by digging out the value of data.

This study is not finished by this thesis; there are still improvements remained for further research. As mentioned above, we will continue to elucidate the visual influences on data utilization efficiency, excavate the use value of data, and extend the method to other fields as support for visual analytics decision making.

Acknowledgement

This work was funded by JSPS KAKENHI JP16H01836, JP16K12428, and industrial collaborators.

References

[Ohsawa 13] Ohsawa Y, Kido H, Hayashi T, et al. Data jackets for synthesizing values in the market of data. Procedia Computer Science, 2013, 22: 709-716.

[Hayashi 13] Hayashi T, Ohsawa Y. Processing combinatorial thinking: Innovators marketplace as role-based game plus action planning. International Journal of Knowledge and Systems Science (IJKSS), 2013, 4(3): 14-38.

[Wang 13] Wang H, Ohsawa Y. Idea discovery: A scenario-based systematic approach for decision making in market innovation. Expert Systems with Applications, 2013, 40(2): 429-438.

[Liu 13] Liu C, Ohsawa Y, Suda Y. Valuation of data through use-scenarios in innovators' marketplace on data jackets. Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013: 694-701.

[Ohsawa 15] Ohsawa Y, Kido H, Hayashi T, et al. Innovators marketplace on data jackets, for valuating, sharing, and synthesizing data. Knowledge-Based Information Systems in Practice. Springer, Cham, 2015: 83-97.

[Shannon 48] Shannon C E. A mathematical theory of communication. Bell system technical journal, 1948, 27(3): 379-423.

[Arndt 04] Arndt, C. (2004), Information Measures: Information and its Description in Science and Engineering, Springer, ISBN 978-3-540-40855-0

[Pathria 11] Pathria, R. K.; Beale, Paul (2011). Statistical Mechanics (Third Edition). Academic Press. p. 51. ISBN 978-0123821881.

[Martin 11] Martin, Nathaniel F.G. & England, James W. (2011). Mathematical Theory of Entropy. Cambridge University Press. ISBN 978-0-521-17738-2.

[Stone 14] Stone, J. V. (2014), Chapter 1 of Information Theory: A Tutorial Introduction, University of Sheffield, England. ISBN 978-0956372857.

[Tsai 08] Tsai D Y, Lee Y, Matsuyama E. Information entropy measure for evaluation of image quality. Journal of digital imaging, 2008, 21(3): 338-347

[Dehmer 08] Dehmer M. Information processing in complex networks: Graph entropy and information functionals. Applied Mathematics and Computation, 2008, 201(1-2): 82-94.

[McCallum 00] McCallum A, Freitag D, Pereira F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation. Icml. 2000, 17(2000): 591-598.

[Xuecheng 92] Xuecheng L. Entropy, distance measure and similarity measure of fuzzy sets and their relations. Fuzzy sets and systems, 1992, 52(3): 305-318.

[Parker 16] Parker A J, Yancey K B, Yancey M P. Regular Language Distance and Entropy. arXiv preprint arXiv:1602.07715, 2016.

[Hayashi 15] Hayashi, T., Ohsawa, Y.: 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.

[Leeuw 88] De Leeuw J. Convergence of the majorization method for multidimensional scaling. Journal of classification, 1988, 5(2): 163-180.

[Leeuw 11] De Leeuw J, Mair P. Multidimensional scaling using majorization: SMACOF in R. 2011

[Shi 16] Shi Q, Chen Z, Fang C, et al. Measuring the diversity of a test set with distance entropy. IEEE Transactions on Reliability, 2016, 65(1): 19-27.