Context-based Network Analysis of Structured Knowledge for Data Utilization

Teruaki Hayashi, Yukio Ohsawa
Department of Systems Innovation, School of Engineering, The University of Tokyo
hayashi@sys.t.u-tokyo.ac.jp, ohsawa@sys.t.u-tokyo.ac.jp

Abstract

In this study, we analyzed and discussed the network of structured knowledge for data utilization using data jackets (DJ). A DJ is a method used to describe the summary information related to data, and its purposes are to improve the readability of the data by the users and to promote data understanding. The knowledge for data utilization consists of three-partite graphs that correspond to the three tuples, including the requirements, solutions (proposals of data usage), and DJs. By combining these units of knowledge, we can describe complex knowledge based on data utilization. Application of the methods of network analysis to the knowledge base allows us to understand the latent connections between data and knowledge elements that cannot be understood in the conventional network of data.

Introduction

In view of the worldwide trend on big data and AI, cross-disciplinary data exchange and collaboration constitutes one of the essential social demands. Although platforms used to display data information and their exchange have been developed, a knowledge base intended to be used for a) the promotion and the discovery of valuable data and b) for the use of exchange methods for data utilization, has not been well developed yet. To create a collaborative knowledge base for data in this study, we have created a network of structured knowledge for data utilization and have discussed the results based on the analyses of its structure and features.

Knowledge Graph for Data Utilization

In this study, we used the results generated in the workshops of Innovators Marketplace on Data Jackets (IMDJ) as the knowledge elements for data utilization. The IMDJ was a workshop conducted to analyze and discuss cross-disciplinary data utilization using Data Jackets (DJs). A DJ is a method used to summarize information regarding data, and its purpose is to improve the readability of the data by the users, and to promote data understanding (Ohsawa et al. 2013). Data owners provide their datasets as DJs, and data users state their demands as requirements, while analysts create solutions to solve user requirements. To create the structured knowledge, we introduce two units of knowledge for data utilization with binary predicate logic (Hayashi and Ohsawa 2018a),

$$\textbf{satisfy}(solution, requirement) \quad (1)$$

$$\textbf{combine}(solution, DJ) \quad (2)$$

Equation (1) formulates the relationship such that a certain solution satisfies a stated requirement. Equation (2) indicates that a combination of the DJs generates a solution. Therefore, we can create a network by combining the units of knowledge. The nodes are requirements, solutions, and DJs, and the links of the three entities are the predicates (**satisfy** and **combine**). We conducted 19 IMDJ workshops and stored 291 requirements, 419 solutions, and 271 DJs, to create the network.

Results and Discussion

The network consisted of a three-partite graph. Figure 1 shows the largest component of the network. Hayashi and Ohsawa (2018b) used the terms in the data outlines in each

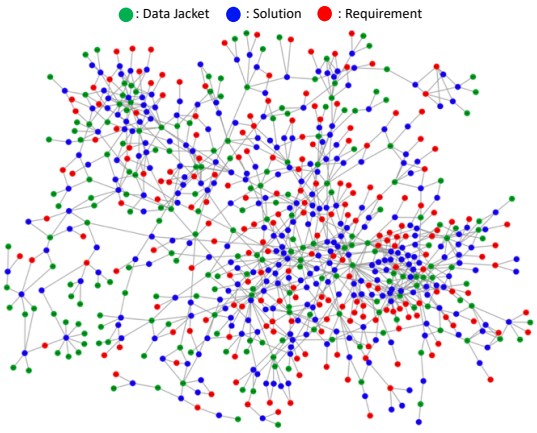

Figure 1: Network of knowledge for data utilization

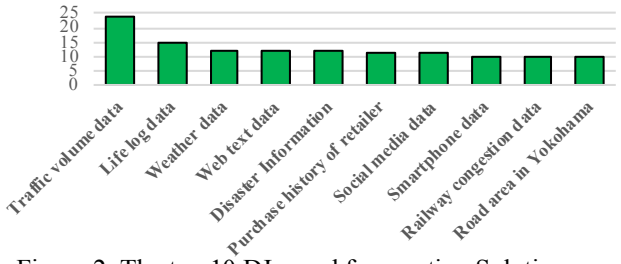

Figure 2: The top 10 DJs used for creating Solutions

DJ as the contexts to create the context-based data network. Although the outlines included the intentions of the data acquisition and backgrounds of data usage, the context defined by their research was the context within the existing data. For cross-disciplinary data exchange and collaboration, it is necessary to consider the newly given contexts to combine data from different areas. Therefore, in this study, we defined the solutions and requirements that corresponded to the upper layer as the contexts, and not the terms in the DJ. Figure 2 shows the top 10 DJs used to creating solutions equivalent to the degrees of DJs in the network. Traffic volume data is most frequently used. Accordingly, three out of the top 10 DJs are related to traffic. The result suggests that data related to traffic are the hubs in the network. Furthermore, data derived from personal information—life log data, social media, and smartphones—are also frequently used (three out of the top 10 DJs).

Table 1 lists the basic features of the largest component of the network. Given that it is a three-partite graph and there is no triad in the network, the average degree is relatively low, the density is extremely low, and the clustering coefficient is zero, compared to a random network. Assortativity is shown to be low but positive, and it has a structure that is close to that of a human relationship. Compared to the data network (Hayashi and Ohsawa 2018b), the diameter is larger, and the average path length is longer so that the network of knowledge is sparse both globally and locally.

Second, we compared the degree and the betweenness centralities of each knowledge element and found that DJs in Fig. 1 have increased degrees and betweenness centrality values. Traffic volume data frequently used for creating solutions have the highest degrees of centrality (0.033) and the highest betweenness centrality value (0.244). Additionally, the life log data has the second highest degree of centrality (0.021) and the second highest betweenness centrality (0.204) values. Based on these results, the data related to traffic and personal information act as the hub nodes in the network and appear positions that bridge different areas.

The solution with the highest degree of centrality "identified the traffic volume at the time of disaster and assessed the risks based on past disaster information, traffic volume at major points, and weak transportations, by adding traffic accident statistics." This was accomplished based on the combination of seven traffic datasets. Conversely, the solution with the highest betweenness centrality "improved the allocation efficiency by combining weather conditions, major regions, personal behaviors, traveling data, and taxi usage conditions." This was accomplished based on the combination of five datasets from different domains. We found that the solutions identified with the use of data from the different areas have increased betweenness centrality values, while the solutions that combine data in the same areas have a higher degree of centrality than other solutions. The results seem quite natural, but we did not use any textual information in the description of the solutions in the analyses. It is interesting to note that the results obtained by analyzing the network of structured knowledge and their connections were in agreement with intuitive understanding based on the use of linguistic information.

In this study, we analyzed and discussed the features of the network of knowledge for data utilization. By analyzing the network of data with the contexts (solutions and requirements), we can understand the latent connections which could not be understood only with the use of data networks. In the future, we will consider the texts in knowledge elements, and we will analyze the deeper connections between data and knowledge.

Table 1: Characteristics of the network

	Values
Number of nodes	719
Number of links	1009
Average degree	2.81
Link density	0.004
Clustering coefficient	0.00
Assortativity	0.022
Average shortest path	8.90
Diameter	28
Radius	14

Acknowledgments

This study was partially supported by JST-CREST Grant Number JPMJCR1304, and JSPS KAKENHI Grant Numbers JP16H01836 and JP16K12428.

References

Ohsawa, Y.; Kido, H.; Hayashi, T.; and Liu, C. 2013. Data jackets for synthesizing values in the market of data. *Procedia Computer Science,* 22:709-716.

Hayashi, T., and Ohsawa, Y. 2018a. Retrieval System for Data Utilization Knowledge Integrating Stakeholders' Interests. *AAAI Spring symposium 2018 Beyond Machine Intelligence: Understanding Cognitive Bias and Humanity for Well-being AI,* 2018.

Hayashi, T., and Ohsawa, Y. 2018b. The Difference between Variable-based and Context-based Networks of Data Using Data Jackets, *Procedia Computer Science,* 126: 1740-1747.