Analysis of Interests of Civic Tech Communities in Japan towards Developing Civic Tech Community Recommendation System

Ryota Kozakai[1], Shun Shiramatsu[1, 2], Masato Kikuchi[1], Tadachika Ozono[1], Akira Kamiya[2], Tetsuji Ishii[3]
ryouta039@gmail.com, {siramatu,kikuchi.masato,ozono}@nitech.ac.jp, a.kamiya.208@srmtlab.org, ishii.gifu403@gmail.com

1 Graduate School of Engineering, Nagoya Institute of Technology

2 Code for Nagoya
Nagoya, Japan

3 Code for GIFU
GIFU, Japan,

***Abstract*— In Japan, although the participants of Slack workspace of "Code for Japan" are rapidly increasing, they tend not to participate in local civic tech communities because the interests of each civic tech community are difficult to grasp. Thus, we aim to develop a system for recommending local civic tech communities in Japan to enhance participation in regional civic tech activities. In 2020, we collected and analyzed data of interests of civic tech communities in Japan using a questionnaire to build a system to recommend appropriate civic tech communities for local issues and technologies. The recommendation system was then evaluated. Since we were able to receive evaluations from people who are actually familiar with certain civic tech communities, we conducted a questionnaire to check each recommendation, and the overall result was that the recommendations were appropriate, adequate, and/or relevant. In addition, the results showed that those who knew about the activities of the civic tech community rated the recommendations slightly higher than those who did not, and in this sense, we are looking forward to further developing the recommendation system.**

I. Introduction

Civic tech (civic technology) refers to public activities for addressing social issues using information technologies through collaboration and co-creation among citizens, engineers, government officials, and various experts. Recently, the civic tech activities have been socially focused on in Japan because there are a lot of activities to address problems related to the COVID-19 Pandemic. Since 2020, Japanese mass media have frequently reported the activities of "Code for Japan," a civic tech organization attempting to enhance civic tech activities in Japan. This promotion has rapidly increased the number of the participants of Slack workspace of Code for Japan. However, the participants tend not to participate in local civic tech communities because the interests of each civic tech community are hard to grasp.

To enhance participation in regional civic tech activities in Japan, we aimed to develop a web application for recommending local civic tech communities. We first conducted a questionnaire to collect data on each local civic tech community's interests in social issues and technologies before the Code for Japan Summit 2020 Online. In this paper, we analyze civic tech communities as a first step to create a recommendation system.

In this paper, we aim to develop questionnaire analysis methods satisfying three requirements.

1. Extract the types of clusters and their respective characteristics that are proximate to each civic tech community.
2. Visualize the main characteristics of each civic tech community.
3. Select recommended civic tech communities on the basis of technologies that can be used for local issues.

These efforts are intended to enable people (civic tech beginners and civic tech veterans) to understand the characteristics of civic tech communities and information on the local issues that they are currently addressing.

II. Related works

In the same vein of promoting civic tech activities, which is the purpose of this paper, Code for Nagoya applied "GoalShare," a web system for collecting open data to share hierarchical goals for solving social issues, to a civic hackathon held in 2014. The system was intended to encourage continuous development of prototypes that were created in the hackathon [1].

One study used pointwise mutual information (PMI) to detect stop words in order to exclude in advance those with low availability. In the content of this research, the method is the same because PMI is used to extract those with high relevance to local issues and technologies, and to avoid recommending those with low relevance [3]. We also use PMI for recommending civic tech communities related to a users' interests in this study.

Some studies have applied the weighted coefficients of the multi-class logistic regression to feature extraction, which is also used in this study. Some predict precipitation types by coefficients using multi-class logistic after predetermining the weather variables that can be used for weather forecasting [4]. For geochemical discrimination of monazite source rocks, there is an example of discrimination by multi-class logistic, and the resulting coefficients identified important discriminants reflecting metamorphism [5].

In this study, hierarchical clustering is used to classify the civic tech communities into several clusters on the basis of their features. One of the first papers to develop hierarchical clustering was published in 1967 [6], and since then it has been applied in various fields. Some papers describe efficient implementations that can be used in software [7]. The gap statistic specifies the optimal number of clusters by the hierarchical cluster method. An example of using this method is Habib et al. [8]. In their paper, to avoid sparsity since the data to be obtained is a categorical attribute, a similarity matrix is created from the binary data, and after dimensional compression, a continuous coordinate vector is constructed, and hierarchical clustering by a gap estimator is performed.

III. Collecting Data of Civic Tech Communities' Interests and Web Application Visualizing It

Since August 2020, we have conducted a questionnaire to collect data of interests of civic tech communities in Japan. This questionnaire includes each community's interests, activity period, activities, characteristics, and personnel composition.

As a preparation of Code for Japan Summit 2020 held in October 2020, we developed a web application, "Civic Tech Fukan Zukan (overviewing illustration)" [2], which visualizes the similarity among communities shown in Figure 1. Although this web application enabled users to grasp the similarities of interests of civic tech communities, it was not enough to support collaboration between communities.

For the analysis in this paper, we use 50 samples corresponding to the civic tech communities that responded to this survey. Five items for each sample are used in our analysis.

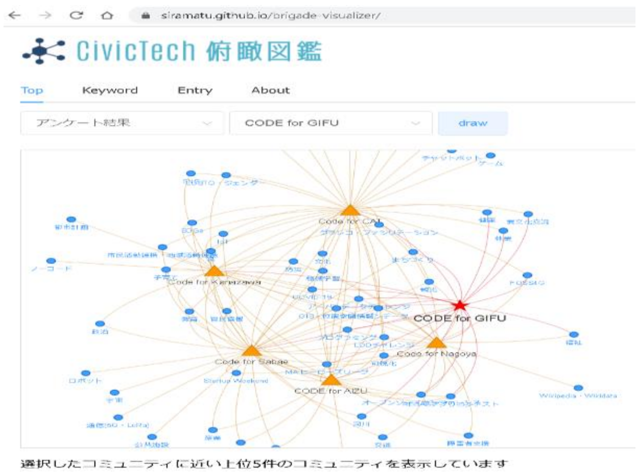

Fig. 1: Civic Tech Fukan Zukan (overviewing illustration)

- Local issues: Please select the local issue or area of interest you are working on.
- Technology: Please select the technology you are working on.
- Exhibit: Please select any competitions or exhibit events that you have applied for or considered participating in.
- Members: What are the attributes of the members?
- Characteristics: If you have any other keywords that describe your Civic Tech Community, please enter them here.

IV. Hierarchical Clustering of Communities

The first layer (classification of two clusters) is divided by the most significant factors. Among these, the most divergent feature was the Graphic recording and facilitation (technology), which had the lowest frequency in Cluster 1, followed by Researchers and faculty members (members). If the frequency in Cluster 1 is not low (20-40%), we find that the frequency of that part may increase when the number of clusters is further subdivided. Since the first layer is considered to include several potential clusters, the frequency may decrease. In the second layer (i.e., the classification of three clusters), there can be one cluster with only Code for Cat because this cluster contains results for "Cat Activities." Example include Protected Cat (local issues), Local Cat (local issues), My Number Card for Cat (local issues), Zero Killing (local issues), Cat Lovers (characteristics), and Cat Bowl (characteristics). It is characteristic that Cat Protection Activities and Cat PR Activities can only be seen in this group. In addition, Fig. 2 shows that Code for AICHI is a characteristic organization. In that part, WEB (technology), Regional Revitalization (characteristics), Creative (characteristics), Community Design (characteristics), Ideathon, and Hackathon (characteristics) are characteristics and do not apply to other organizations. Code for AICHI also featured the RESAS Regional Revitalization Idea Contest, Call for Code, GUGEN, and Maker Faire (exhibited). The regional economic analysis system (RESAS) aggregates and visualizes public and private big data such as vital statistics, industrial structure, and flow of people. The RESAS Regional Revitalization Idea Contest solicits policy ideas that will revitalize the region, on the basis of the analysis of regional

Table 1: Features that have over 40% frequency difference between the top-level two clusters.

Features that have high frequency difference between the top-level two clusters	Cluster1 (frequency)	Cluster2 (frequency)	Frequency difference
GIS and geospatial information (Technology)	0.3437	0.875	0.5312
Researchers and Faculty (Member)	0.1562	0.6875	0.5312
Graphic recording and Facilitation (Technology)	0.0625	0.5625	0.500
Visualization (Technology)	0.2187	0.6875	0.4687
Urban Data Challenge (Exhibit)	0.4062	0.875	0.4687
Designer (Member)	0.2187	0.6875	0.4687
Students (Members)	0.2187	0.6875	0.4687
COVID-19 (Local Issues)	0.375	0.8125	0.4375
Administrative staff (members)	0.375	0.8125	0.4375

issues using RESAS. In addition, Call for Code unites hundreds of thousands of developers to create and deploy applications using opensource technologies to confront global social issues. GUGEN supports manufacturing activities through events such as hardware hackathons and contests that focus on problem solving. Maker Faire is a gathering of "Makers" who use new technologies that have become available to everyone in unique ways to create surprising and useful things that have never existed before and hold exhibitions and demonstrations.

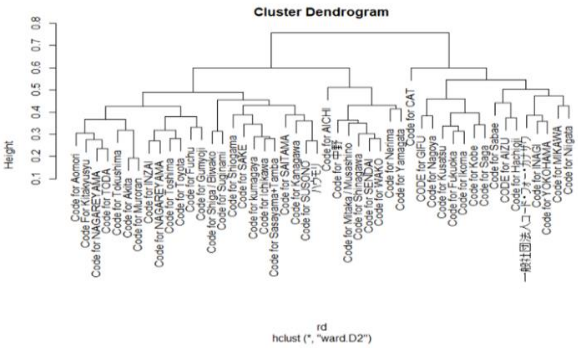

Fig. 2: Cluster Dendrogram of Civic Tech Community.

We also calculated the optimal number of clusters for each feature related to "technology," "local issues," "feature," "exhibit," and "member" using the gap statistic. The gap statistic is a measure of the difference in the degree of cohesion between clusters obtained by clustering data and clusters obtained by clustering random data. Gap statistic is expressed:

$$Gap(k) = E[\log(W_k)] - \log(W_k)$$

W_k is the pooled within-cluster sum of squares around the cluster means. From this, we calculated the optimal number of clusters (Fig. 3 and Table 2).

Table 2 represents the Gap statistics of several clusters starting from one cluster, and since there is no significant decrease in Gap statistic after Cluster 6, we set Cluster 6 as the optimum and confirmed the main activity features for each cluster. Fig. 2 shows that the clusters of the Cluster Dendrogram are divided into many clusters,

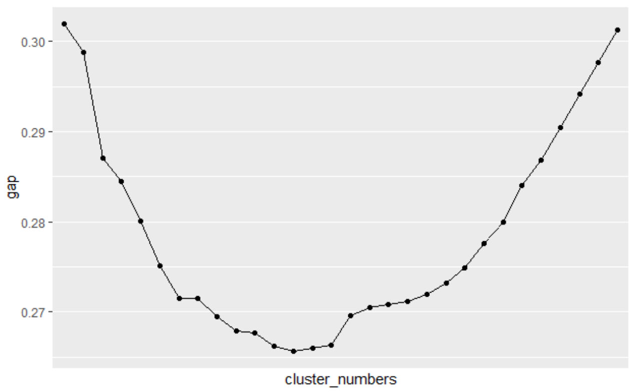

Fig. 3: Graph of gap statistics for each cluster.

indicating that the clusters are classified to some extent. For the actual clustered features, we calculated the weight coefficient matrix using multi-class logistic regression, ranked the weights by their magnitudes, and tabulated the results.

Table 2: Gap statistics for each cluster.

logW	E.logW	gap	cluster_numbers
1.5335	1.8355	0.3020	1
1.4818	1.7806	0.2988	2
1.4462	1.7337	0.2876	3
1.4059	1.6904	0.2846	4
1.3702	1.6502	0.2800	5
1.3365	1.6119	0.2754	6
1.3025	1.5742	0.2717	7
1.2658	1.5371	0.2713	8

Table 3: Cluster details (shows the top 5 features except Code for AICHI and Code for CAT)

cluster	feature (1st)	feature (2nd)	feature (3rd)	feature (4th)	feature (5th)
1	Researcher/ Teacher (Member)	Politics (Local issues)	Food (Local issues)	Culture (Local issues)	Non-IT Engineer (Member)
2	Company employee/ Organization employee (other than the above) (Member)	Open data (Technology)	No code (Technology)	Tourism (Local issues)	City planning (Local issues)
3	Graphic recording and Facilitation (Technology)	Regional Archive (Local Issues)	Wikipedia/ Wikidata (Technology)	Diversity (Characteristics)	Mayan Calendar (Characteristics)
4	Programming (Technology)	Living (Local Issues)	Child Rearing (Local Issues)	Welfare (Local Issues)	Support for the Disabled (Local Issues)

Table 3 lists the 16 cluster features in descending order for each cluster. Table 3 shows the top five, excluding Code for AICHI and Code for CAT. As mentioned earlier, Code for AICHI and Code for CAT are not a cluster, because they constitute a cluster with one organization. If we check the features of each cluster except Code for AICHI and Code for CAT in descending order from the top, we can see that Cluster 2 has No code or Open data in terms of technology, and that Company employees and Organization employees are engaged in Tourism and City planning. Cluster 4 has programming as its most prominent feature, and although not included in the top five, IT engineers(member) is in sixth place. This cluster is very active in information technology. In addition, we can see the support activities in the community such as Living (local issues), Child rearing (local issues), Welfare (local issues), and Support for the disabled (local issues). In the first layer (classification of two clusters), graphic recording and facilitation (technology), and

Researchers and teachers (members) were cited as the major factors that divided the clusters, which are also characteristic of Clusters 1 and 3.

Table 4: Top 5 frequency of feature terms for each cluster (excluding Code for CAT and Code for AICHI)

cluster	feature (1st)	feature (2nd)	feature (3rd)	feature (4th)	feature (5th)
1	Public-Private Collaboration (Local Issues)	Open data (Technology)	IT Engineer (Member)	Education (Local Issues)	GIS & Geospatial Information (Technology)
2	Open data (Technology)	Company employee/ Organization employee (other than the above) (Member)	City planning (Local issues)	IT Engineer (Member)	Urban Data Challenge (Exhibitor)
3	Open data (Technology)	Urban Data Challenge (Exhibitor)	IT Engineer (Member)	Company employee /Organization employee (other than the above) (Member)	COVID-19 (Local Issues)
4	Cooperation in civic and community activities (local issues)	Programming (Technology)	IT Engineer (Member)	Education (Local Issues)	City planning (Local issues)

In Table 4, IT engineers (members) are included in all clusters, and Open data (technology) is also found in all clusters except Cluster 4. The reason it does not overlap with the characteristic activities is that it is not expected to be the main discriminating factor for the activities that are conducted in all clusters. In addition, the two clusters occur with different frequencies, so it is difficult to make a generalization.

V. Extracting Features of Communities Requirement

The features of each organization were listed from the first to the third place on the basis of the weight coefficient matrix of multi-class logistic regression. The data sample for each civic tech community was categorical and seemed insufficient to produce data trends, but someone familiar with the civic tech community told us that the results were good. The objective variable is dummy data with each civic tech community as a class, and the explanatory variable is the categorical data of the five specified features (technology, local issues, characteristics, exhibit, and member), modified into binary form (0 or 1). In a multi-class logistic regression, a judgment can be made on the basis of the weighted coefficients. The results were very interesting (Table 5). For example, we found several characteristic ones, such as Awa Odori in Code for Tokushima and Nomi Meeting in Code for Toshima. The Awa Odori dance is widely known as one of Japan's representative traditional performing arts.

Table 5: Three main features in each civic tech community (Code for Tokushima, Toshima, SAITAMA, SAKE, Nagoya)

Team	Feature (1st)	Feature (2nd)	Feature (3rd)
Code for Tokushima	Festival (Local issues)	Awa Odori (Characteristics)	MA Heroes League (Exhibit)
Code for Toshima	Nomi Meeting (Characteristics)	No application experience (Exhibit)	Artist (Member)
Code for SAITAMA	Remote Sensing (Characteristics)	Point cloud data (Characteristics)	AI (Characteristics)
Code for SAKE	Japanese Sake (Local issues)	Japanese Sake (Characteristics)	Health (Local issues)
Code for Nagoya	System the representative changes monthly (Characteristics)	Support for the Disabled (Local Issues)	Welfare (Local issues)

VI. Prototype Recommendation Method

We created a model that provides appropriate civic tech communities and related technologies for local issues that users are interested in. Local issues and technologies related to them were selected on the basis of the PMI. From the weight coefficients of the multi-class logistic regression, civic tech communities with particularly strong characteristics of the local issues and technologies are selected on the basis of the PMI. For example, in the area of traffic (local issues), Code for Muroran, Code for Akita, and Code for Tokushima were notable.

Table 6: Examples of communities recommended for "traffic" (local issues)

Fields of interest of expected users	Technology that can be used in fields of interest	PMI	Recommended communities (in order from first to third)
Traffic (Local issues)	GIS & Geospatial Information (Technology)	0.6523	Code for Muroran, Code for Akita, Code for Tokushima
Traffic (Local issues)	Open data (Technology)	0.2076	Code for Muroran, Code for Akita, Code for Tokushima

In addition, Table 7 shows examples of recommended civic tech communities when a user inputs "disaster prevention" as a local issue that she/he is interested in. This indicates that the organizations are involved in activities related to disaster prevention, along with the technologies involved.

Table 7: Examples of communities recommended for "disaster prevention" (local issues)

Fields of interest of expected users	Technology that can be used in fields of interest	PMI	Recommended communities (in order from first to third)
Disaster prevention (Local issues)	Visualization (Technology)	0.4700	Code for Ichikawa , Code for Saga, Code for INAGI
Disaster prevention (Local issues)	SNS (Technology)	0.4159	Code for Kumagaya, Code for Sasayama+Tamba, Code for Ichikawa
Disaster prevention (Local issues)	GIS & Geospatial Information (Technology)	0.3421	Code for Kumagaya, Code for TODA, Code for SUSONO
Disaster prevention (Local issues)	Programming (Technology)	0.1698	Code for TODA, Code for SUSONO, Code for Saga
Disaster prevention (Local issues)	Open data (Technology)	0.0645	Code for Ichikawa, Code for Sasayama+Tamba, Code for TODA

VII. Evaluation Experiment and Discussion

The overall evaluation of the selection of each community and the PMI was done on a scale of 1 to 7. These evaluation scores were collected on Lancers, a Japanese crowdsourcing website. There were 141 cloud workers in total. Each question on the selection of communities was

- Does the following list of features (clusters) include any features that you think are appropriate for your selected civic tech community?
- Please rate the appropriateness of the following three characteristics for your selected civic tech community.
- Suppose you are interested in a certain community issue and input it to the recommendation system. Assume that the system makes the following recommendation. There is a possibility that you can work with the community "Code for ~" using the technology "~". How appropriate is this recommendation?

Other questions at the beginning of the survey asked about programming experience and perceptions of examples of civic tech community activities, and questions at the end of the survey asked about confidence in the answers. The question on PMI was: "Please rate the relevance of the following IT technologies in developing services that address the following social issues." Two-hundred crowd workers responded to the PMI questionnaire. The beginning of the questionnaire asked about the respondent's number of years of experience in IT. The results of the community selection (overall evaluation) for each question were as follows. We asked respondents to research the civic tech community of their choice before answering.

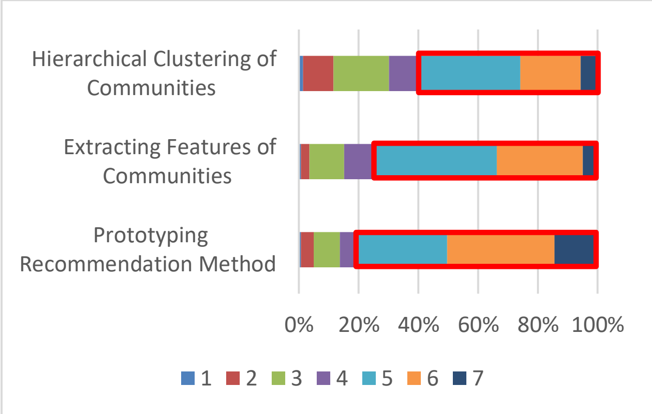

Fig. 4: Questionnaire survey results (1: Low rating to 7: High rating)

The medians of the hierarchical cluster and extracting features of communities using multiple logistic regression (=5.0) were the same, but the mean value of extracting features (=4.489) was lower than that of the hierarchical cluster (=4.935) . One reason for this is that multiple civic tech communities are included in a cluster, making it difficult to select features for the entire cluster. Since it is difficult to select by frequency of occurrence, we used the weight coefficient matrix of multi-class logistic regression to examine the trend. As for the selection of communities recommended by PMI for technologies related to issues of interest (see also Prototype Recommendation Method), the median value (=6.0) is high, but we want to raise the mean value (=5.258) slightly more.

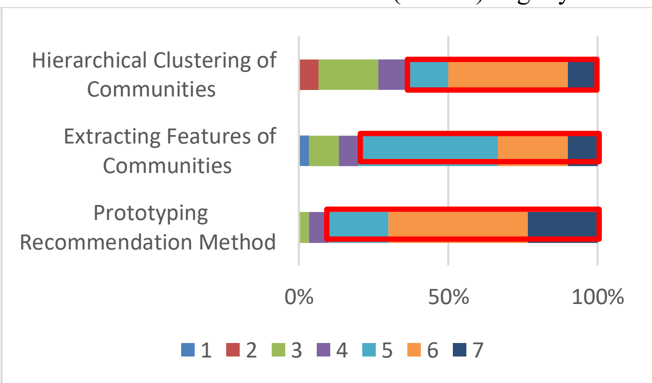

Fig. 5: Questionnaire survey results (Crowd workers who knew about civic tech communities, 1: Low rating to 7: High rating)

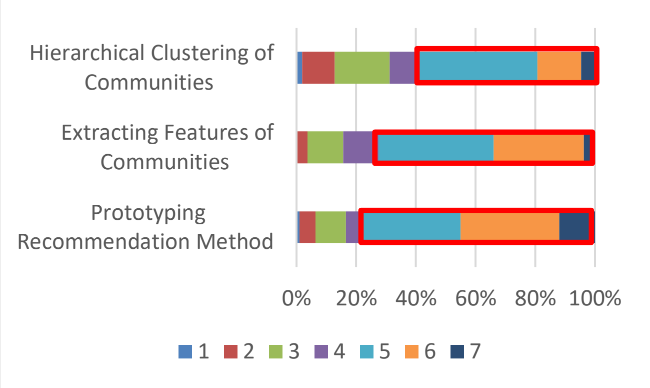

Fig. 6: Questionnaire survey results (Crowd workers who did not know about civic tech communities, 1: Low rating to 7: High rating)

The results were also compared between crowd workers who knew examples of civic tech community activities (familiar crowd workers) and those who did not (unfamiliar crowd workers). For the hierarchical clusters, familiar crowd workers gave higher median and mean values (5.5 and 4.9) than unfamiliar ones (5.0 and 4.376). In addition, among familiar crowd workers, the mean and median ratings for extracting features of communities using multi-class logistic regression were 5.000 and 5.033, respectively, which were lower than the results for hierarchical clusters. Among unfamiliar crowd workers, the mean and median values were 5.0 and 4.908, respectively. These results show that unfamiliar crowd workers gave the same median values for the hierarchical cluster, extracting features and prototyping recommendation method (5.0), but a lower mean value for the hierarchical cluster (4.376 vs. 4.908).

As for the selection of communities recommended by PMI for technologies related to the issues of interest, familiar crowd workers gave higher median and mean scores (6.0 and 5.8) than unfamiliar ones (5.0 and 5.11). One reason for the higher overall rating seems to be the influence of the familiar crowd workers. When we examined PMI and user ratings later, the correlation coefficient value was 0.21, indicating a weak positive correlation. We also felt that there was a possibility that the ratings would differ depending on the number of years of experience in IT, but no particular differences were found.

VIII. CONCLUSION

As basic research in the development of a system for recommending civic tech communities in Japan, we tried three analysis methods: (1) hierarchical clustering using a similarity matrix calculated from the categorical data, and the number of clusters is determined by the gap statistics, (2) extracting features of communities using multi-class logistic regression, and (3) recommending communities and technologies from issues that a user is interested in by using pointwise mutual information (PMI). These analyses were conducted as part of the visualization of the activities and features of civic tech communities and as part of the analysis for providing useful information to users. Moreover, we conducted an evaluation experiment and found that although the overall evaluation of the hierarchical clusters was low, the rating was high for crowd workers who knew examples of civic tech community activities and low for the those who did not. Otherwise, in the evaluation of those who did not know examples of civic tech community activities, the median value remained the same for (1), (2), and (3), and the mean value is only slightly lower that in the hierarchical clustering section (1). This suggests that we need to take into account the validity of the evaluator's assessment. In addition, (1) and (3) were highly rated by those who were familiar with the examples of civic tech community activities.

As future work, we are currently planning to reexamine (2), which was relatively lowly evaluated, because it received better qualitative evaluations from an interviewee who is familiar with the civic tech communities. Furthermore, we need to develop a function for recommending civic tech communities on the basis of our proposed methods and append it to our web application “Civic Tech Fukan Zukan.”

ACKNOWLEDGEMENT

This work was supported by JST CREST (JPMJCR15E1, JPMJCR20D1) and NEDO (JPNP20006). The authors would like to greatly thank Japanese civic tech communities and their members who cooperated with this research.

REFERENCES

[1] Shun Shiramatsu,Teemu Tossavainen, Tadachika Ozono, Toramatsu Shintani. Towards Continuous Collaboration on Civic Tech Projects: Use Cases of a Goal Sharing System Based on Linked Open Data. in *Electronic Participation - 7th IFIP WG 8.5 International Conference, ePart 2015*, Springer LNCS, Vol. 9249,2015

[2] Civic Tech Fukan Zukan, https://siramatu.github.io/brigade-visualizer/, 2020 (accessed on Aug.8, 2021)

[3] Jung-Been LEE, Taek LEE, and Hoh Peter, Automatic Stop Word Generation for Mining Software Artifact Using Topic Model with Pointwise Mutual Information. *IEICE TRANS. INF. & SYST., VOL.E102–D, NO.9*,2019.

[4] Seung-HyunMoon and Yong-HyukKim, An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression. *Atmospheric Research Volume 240*,2020.

[5] Keita Itano,Kenta Ueki,Tsuyoshi Iizuka and Tatsu Kuwatani. Geochemical Discrimination of Monazite Source Rock Based on Machine Learning Techniques and Multinomial Logistic Regression Analysis, *Geosciences*, 2020.

[6] STEPHEN C. JOHNSON, HIERACHICAL CLUSTERING SCHEME. *PSYCHOMETRIKA--VOL 32~ NO. 3*, pp.241–254, 1967

[7] Fionn Murtagh and Pedro Contreras, Algorithms for hierarchical clustering: an overview. *WIREs Data Mining Knowl Discov*, pp.86–97, 2012.

[8] Usman Habib, Khizar Hayat and Gerhard Zucker, Complex building’s energy system opera-tion patterns analysis using bag of words representation with hierarchical cluster-ing. *Complex Adaptive Systems Modeling*, 2016