Study on Analysis of Questionnaire Data based on Interactive Clustering

Yosuke Watanabe
Nagoya University
yosuke@cmplx.cse.nagoya-u.ac.jp

Tomohiro Yoshikawa
Nagoya University

Takeshi Furuhashi
Nagoya University

Abstract— **Recently, several kinds of values have been employed with respect to the diversification of individuality in the market. Some of these values are currently supported by only a few people, who are referred to as a "minority group". However, there is the possibility that such groups will grow into majority groups with changes in historical background or people's sensitivity. It is both important and effective for market analysis to determine these minority groups at an early stage. Companies often employ questionnaires to develop marketing strategies or design new products, which offer a chance to determine these minority groups. With conventional methods, respondents to a questionnaire are classified based on such attributes as gender and age, and then the classified groups are analyzed or compared. Although conventional analysis is effective for grasping the overall tendency of the evaluation data, it is difficult to determine minority groups because of the diversity of individuality. On the other hand, we have proposed clustering methods based on the tendencies of the answers to the questionnaire. This paper proposes a new method for visualizing the evaluated data based on both the obtained values and their correlation with cluster respondents interactively in the visible space. This paper applies the proposed method to web questionnaire data and shows that an analysis of the results effectively assists us to determine minority groups.**

I. Introduction

Recently, several kinds of values have been employed with respect to the diversification of individuality in markets. Some of these values are currently supported by only a few people, who are referred to collectively as a "minority group". However, there is the possibility that such groups will grow into majority groups with changes in historical background or people's sensitivity. For example, portable game players used to be children's toys enjoyed largely by teenage boys, and thus the game software that was developed was designed primarily for children and the market was not very large. However, now a wide range of people from small children to adult women own and enjoy their own portable game players. The game software that has been developed has also diversified and now targets not only children but also a broad range of generations and both genders, and includes software covering a variety of topics including language learning and recipes. Advertisements and events promoting this software are also popular. Thus, portable game players have developed into a major market. If we are able to determine values early on that make it possible for small groups to grow into major groups and thus develop major future markets as seen with portable game players, it would be extremely beneficial for companies as they plan their marketing strategies. One type of data that can be used for quantifying people's sensitivity and values is that provided by questionnaires. The rating scale method is widely used in questionnaires designed to obtain impressions about such evaluation subjects as products, services and brands. This method requires multiple evaluation subjects and multiple questions, and respondents have to answer to each question by grading their impressions about the evaluation subjects from multiple grade scales while looking at each of the subjects. In this way, we can quantify people's impressions about evaluation subjects in the form of graded data.

The conventional method that companies use for questionnaire data analysis classifies the data based on such attributes as age and gender, and the consumption patterns of the respondents. The classified data are then subjected to multivariate analysis methods [1], such as principal component analysis and factor analysis, to determine the profiles of prospective purchasers and their impressions of the products. However, small groups with potential to grow and unique values are often buried under the majority and are hard to detect since they are small in number and their attributes are diverse.

In this report, we propose classification and analysis methods based on the trends of the answers given to questionnaires such as the dispersion and distribution of questionnaire data rather than classification based on attributes and consumption patterns [2][3][4].

We also propose a method of lowering (visualizing) the level of graded data based on the increase/decrease trends (relative evaluation) of the grades and the average grades and interactive clustering of respondents in the visible space. The proposed method provides users with the distribution status and the tendencies of the data and encourages them to make useful "findings" when performing data analysis through data visualization. Moreover, during the trial and error analysis process, the method helps users to grasp the characteristics of grade tendencies and obtain useful marketing knowledge by interactively clustering the data. For data visualization, Multi Dimensional Scaling (MDS) [5], which uses the difference between the size of the correlation coefficient (increase/decrease of grades) and the average grade as a non-resemblance is employed as a standard for the distance between the two. By focusing on the increase/decrease in grades, we can obtain information about the questions on which each respondent placed an emphasis when evaluating the evaluation subjects. At the same time, we can obtain knowledge about the overall evaluation level (high or low

978-1-4244-3597-5/09/$25.00 ©2009 IEEE

evaluation) of the evaluation subjects from the average grade.

In this report, we apply the proposed method to questionnaire data related to an outdoor product and cluster the respondents based on the resulting visualized data. We also calculate and present the margin of error values for the data of each respondent in the visible space based on the distance between each respondent's data in the grade space (called "original space") with the dimension of "the number of questions x the number of evaluation subjects." Respondents are classified based on the sizes of the margin of error values for further visualization and analysis to achieve more accurate clustering. By employing interactive clustering in the visible space, we proved that we can find minority groups with distinct values that are different from those of majority groups.

II. PROPOSED METHODS

This section describes a method for configuring correlation maps that visualize the grade data by using the MDS, which employs the difference or non-resemblance between the correlation coefficient and the average grade as a standard for distance. It also describes a method of classifying the data based on the margin of error values on the correlation map and reconfiguration of the correlation map.

A. Correlation map configuration

For the graded data obtained with the grade scaling method, the correlation coefficient of the graded data of respondents i and j is defined as r_{ij} and the average grade of respondent i is defined as X_i. The non-resemblance d_{ij} between the MDS of respondents i and j is defined as follows (however, ω is a weight for the correlation):

$$d_{ij} = |X_i - X_j| + \omega(1 - r_{ij}) \tag{1}$$

By defining the non-resemblance d_{ij} as shown in Formula (1), the distance between the respondents in the visualized data decreases as the correlation coefficient increases and the difference between the average values of the respondents' grades decreases. The correlation coefficient r_{ij}, as calculated from the graded data of the two respondents, takes a value in the -1 to 1 range, and the value becomes closer to 1 as the tendency of the grades to increase/decrease becomes more similar. Therefore, the second term in Formula (1) does not represent the value of the grades themselves but indicates the similarity between the grades of the respondents. In other words, it is an index designed to show which questions each respondent evaluated highly or poorly (good or poor impressions). In contrast, the first term in Formula (1) represents high or low grades for the questions as a whole. For example, if a questionnaire is designed in such a way that high grades are given to all the questions, the evaluation or impression regarding the evaluation subjects will also be high, since it is assumed to act as an index representing the similarity in the overall evaluation or the impression. The value of a weight ω on the correlation distance should be configured in such a way that the axis representing the increase/decrease direction for the average grade and the axis representing the change in the correlation distance on the configured map can be as orthogonal as possible. This is to make it easier to find groups of respondents with similar average grades but different evaluation tendencies, or vice versa.

B. Map reconfiguration based on margin of error values

If the similarity between respondents i and j in the original space is defined as d_{ij} and the distance on the correlation map configured in II-B above as d'_{ij}, the margin of error value Ei of respondent i in the visible space can be expressed as follows:

$$E_i = \frac{1}{N-1} \sum_{j=1}^{N} |d_{ij} - d'_{ij}| \quad (i \neq j) \tag{2}$$

(where N represents the number of respondents). The E_i value becomes smaller with the distance with other data as shown in Formula (1) in the original space and the distance on the correlation map become more similar, indicating that much more information is retained from the original space. If an area where respondent data with large margin of error values E_i is found on the resulting correlation map, such respondent data are indicative of the fact that the similarity with the other respondent data is not represented clearly on the map. Therefore, the respondents should be classified based on the size of the margin of error values E_i and the correlation map should be reconfigured using the classified groups of respondents in order to perform clustering and analysis with more information retained from the original space. This is also expected to help extract groups of respondents with unique graded data.

III. EXPERIMENT AND DISCUSSION

A. Experimental questionnaire

This experiment involved 707 respondents and 6 scenarios and using outdoor products α as evaluation subjects. The experiment employed the rating scale method and the respondents were asked to choose one of five grades 1,2,3,4,5 in response to each of 10 questions. In this survey, Grade 5 means "applicable" while Grade 1 means "not applicable." Table I shows the 6 scenarios (presented by videos during the questionnaire) used as evaluation subjects and Table II shows the 10 questions. As all of the questions asked for positive impressions about the evaluation subjects as shown in Table II, we assumed that a higher average grade indicated a better impression of the evaluation subject.

TABLE I
EVALUATION SUBJECTS

Scenario1	Operating a projector
Scenario2	Coffee maker and refrigerator
Scenario3	PC, blog
Scenario4	Shower and dryer
Scenario5	Electric thruster
Scenario6	Pure water

TABLE II
STATEMENT

Question 1	It will make me feel superior to those around me.
Question 2	Maybe I can perform outdoor activities cleanly.
Question 3	I want to perform such activities outside.
Question 4	It may be useful in emergencies such as disasters.
Question 5	It looks easy to carry.
Question 6	It looks easy to assemble and set up.
Question 7	It will make my friends and family happy.
Question 8	I will enjoy such activities outdoors.
Question 9	Maybe, I cannot enjoy these activities without the product α.
Question 10	It will make my outdoor leisure activities more pleasant.

B. Correlation map configuration

Figure 1 shows a correlation map (ω=6) configured using the proposed method. Each dot in Figure 1 represents respondent data, and closer plotted dots indicate a greater similarity between the evaluations and impressions of the respondents. Each of the colored symbols (1 to 4) represents the average grade (rounded up to a whole number)) for each respondent. Figure 1 shows that the respondents who were positive about the product α with high average grades are concentrated at the top of the map while those negative about the product with low average grades are concentrated at the bottom of the map.

We then calculated the margin of error value in the visible space for each respondent based on Formula (2). Figure 2 shows the results we obtained when we compared the size of the margin of error values to the correlation map. The color intensity in Figure 2 represents the size of the margin of error value at each dot, namely a darker color indicates a smaller margin of error value. This reveals that the respondent data retain more information from the original space. The average margin of error value for all the respondents was 1.11 and the distribution was 0.06. Figure 2 shows that the respondent data with small margin of error values are concentrated in the circled area.

As the MDS tends to depend on the default values, the relative positions among respondents change slightly every time the map is reconfigured, which disperses the margin of error value of each respondent. However, when the margin of error values were calculated by reconfiguring the map several times in this experiment, the value for each respondent did not change significantly and the values for certain respondents tended always to be smaller. This is probably because when a map is configured, the majority group of respondents with similar grade tendencies and similar average grades is left as it is while retaining information from the original space. And then, multiple small groups of respondents with distinguished grade tendencies are positioned around the majority group.

C. Classifying groups of respondents based on margin of error values

The respondents were then classified into three groups: Group A (351 respondents) with a below average margin of error value (1.11), Group B (185 respondents) with a higher than average margin of error value and relatively high average grades are distributed in the top left of Figure2 and Group C (171 respondents) with a higher than average margin of error value and relatively low average grades are distributed in the bottom right of Figure 2.

The correlation map was reconfigured for each of the three groups. Figure 3 (a) and (b) show reconfigured correlation maps for Groups A and B, respectively. Each of the dots (1-4) in Figure 3 represents the average grade of each respondent (rounded up to a whole number)) as in Figure 1.

Table III shows the average margin of error value for each group in the Figure 1 "Correlation map" and the average margin of error value for each group in the reconfigured correlation map. When these margin of error values were measured by one side t of the level of significance at 5%, a statistically significant difference was confirmed for all the groups on the correlation map in Figure 1 and on the reconfigured map.

Table III shows that Group A with low margin of error

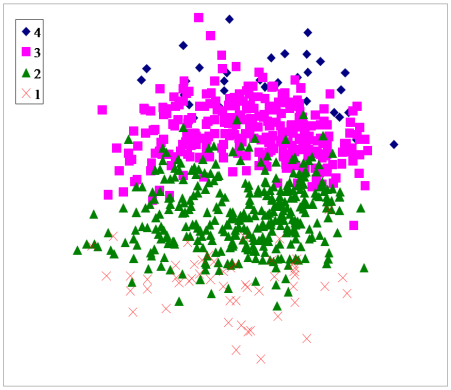


Fig. 1. Correlation map

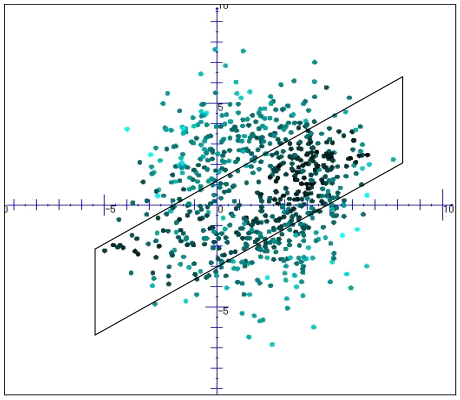


Fig. 2. Comparison with margin of error values

TABLE III
AVERAGE MARGIN OF ERROR VALUES

	Correlation map	Reconfigured map
Group A	0.92	0.75
Group B	1.31	1.46
Group C	1.31	1.39

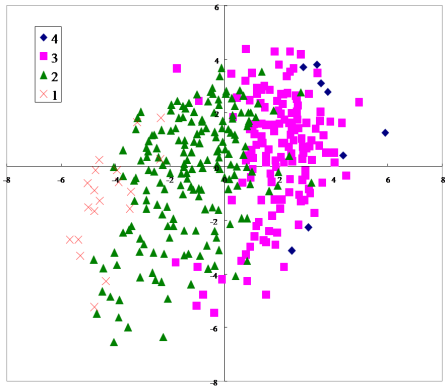
values decreased the margin of error values and retained more information from the original space in the visible space after the map was reconfigured, while the values of Groups B and C with high margin of error values increased as the maps were reconfigured only for those groups. We believe there are many respondents in the center of the correlation map because Group A almost corresponds to the circled area in Figure 2, and the direction of the circled line is fairly similar to the direction that indicates the difference in average grade in Figure 1. Moreover, we consider that the respondents are distributed in a band on the correlation map as the grade tendencies are relatively similar for this group, and groups of respondents with different average grades exist continuously. In this report, we call such groups of respondents the "majority group" and it consists of people with relatively similar grade tendencies.

Meanwhile, for Groups B and C, we believe that many small groups with different average grades and grade tendencies exist discontinuously. The similarity between these groups is slight. However, we believe that by emphasizing their similarity to the majority group, the margin of error value increased as a result of placing the respondents who were supposed to be positioned apart from each other closer together. To confirm this outcome, we used Formula (2) to calculate the margin of error value using only the distance between the respondents in Group B, and the average margin of error value was 2.16. The average margin of error value calculated based on the distance between respondents in Group A was 0.98. We believe that these small groups are "minority groups" whose grade tendencies are significantly different from those of the majority group.

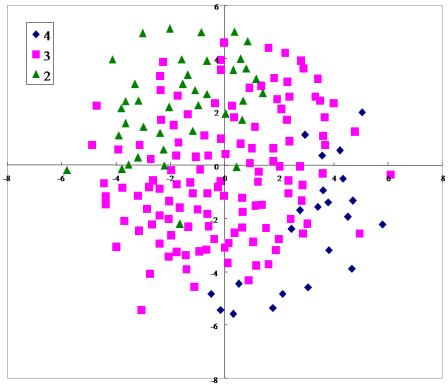
D. Analyzing minority groups

In this section, we extract and analyze a minority group from Group B, which will be useful for marketing the product α. Figure 3(b) shows that the average grades in the top left to bottom right in the figure tend to increase, indicating that the respondents in the bottom right are more positive about the product α. We extracted multiple numbers of small and large clusters mainly from respondents in the bottom right. Figure 4 shows the most distinguished Cluster A (28 respondents), Cluster B (28 respondents) and Cluster C (28 respondents). Figure 5 shows the average grade per evaluation subject and question for each group.

Figure 5(a) shows that Cluster A evaluated the "projector" and "electric thruster" relatively poorly while giving a high evaluation to the other evaluation subjects. Cluster A gave a particularly high evaluation to Question 3 "I want to perform



(a) Group A



(b) Group B

Fig. 3. Re-configured correlation map

such activities outside" and Question 8 "I will enjoy such activities outdoors", which the majority group did not. Thus, Cluster A evaluated the product α highly for outdoor use. Cluster A also gave a high evaluation to Questions 5 and 6, indicating that Cluster A does not have a poor impression as regards portability and ease-of-assembly unlike the majority Positive Group. For these reasons, we believe that Cluster A constitutes prospective purchasers of the product α for outdoor use.

Figure 5(b) shows that Cluster B evaluated Question 8 poorly in relation to the "projector" and "PC" but evaluated the product α highly with respect to the other evaluation subjects. It is noticeable that Cluster B gave a lower evaluation to Question 5 than to Question 6 for all the evaluation subjects while the other groups gave similar evaluations for both questions. This indicates that Cluster B is not very concerned about assembling the product but stresses its poor portability. These results suggest that Cluster B can also be prospective purchasers of the product α and the number of purchasers can be increased further if the poor impression about portability can be overcome.

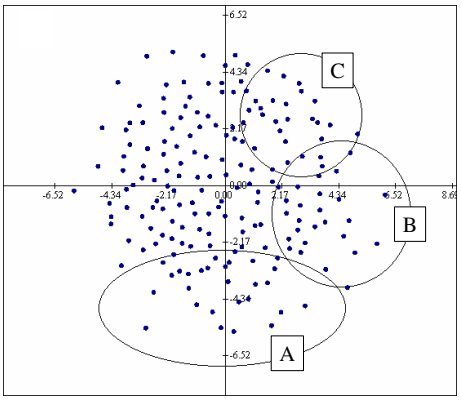


Fig. 4. Cluster formed on minority map

TABLE IV
CLUSTERS OBTAINED BY FCM

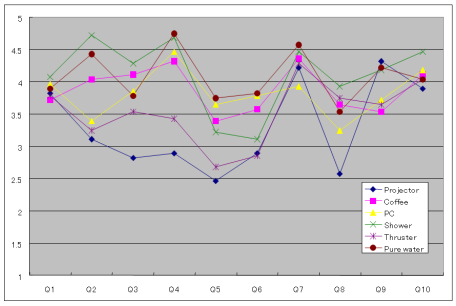
	Number of people	Average grade
Cluster 1	46	2.38
Cluster 2	84	3.21
Cluster 3	65	4.00
Cluster 4	49	1.67
Cluster 5	97	2.25
Cluster 6	63	2.76
Cluster 7	65	2.51
Cluster 8	72	3.50
Cluster 9	75	3.30
Cluster 10	91	2.88

Figure 5(c) shows that Cluster C provided an extremely poor evaluation for only Questions 3 and 8 as regards all the evaluation subjects. Cluster C evaluated the product α relatively highly, but it belongs to the Non-positive Group when it comes to using the product outdoors. For these reasons, we believe it would be effective to recommend the product α for emergency use rather than outdoor use when marketing it to Cluster C.

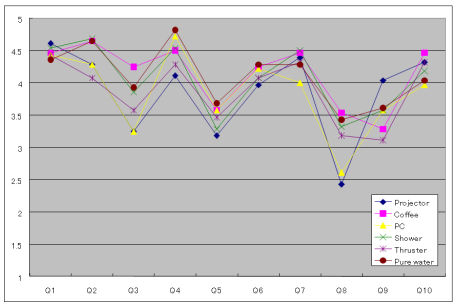
E. Comparison with other methods

To provide a comparison with other methods of extracting minority groups, FCM [5] was applied to the graded data and the respondent data were segmented into 10 clusters. Table IV shows the number of people and the average grade for each cluster. Figure 6, which was created based on Table IV, shows the average grade per evaluation subject and question for Clusters 3 and 8, which are considered to have relatively high average grades and are positive about the product. The comparison with Figure 5 shows that Cluster 3 has a distinct feature that is in the middle of Clusters A and B of the minority groups. These results suggest that increasing the number of clusters and applying FCM and segmenting data could extract clusters similar to those extracted with the proposed method.

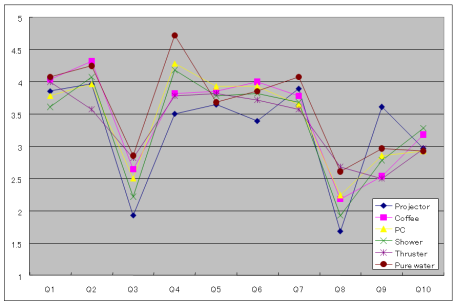
Although we were able to extract clusters by using the average grades as an index in this experiment based on the result described in III-D above, it is necessary to understand



(a) Cluster A



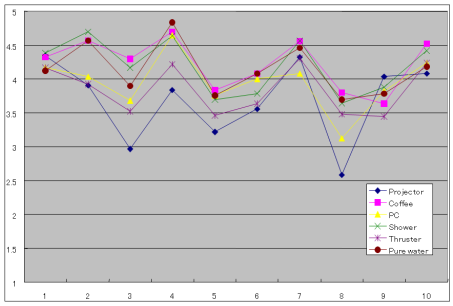
(b) Cluster B



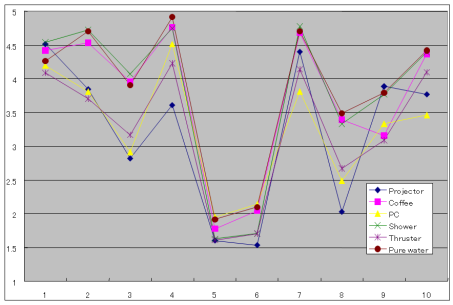
(c) Cluster C

Fig. 5. Average grade (minority group)

the distinct feature of each cluster for analysis with this method. 20-30 clusters are required if we are to extract a cluster of about 30 people as shown in III-D above. Given the number of clusters needed, the analysis will not be easy. A possible alternative method would be to extract clusters with distinguished features while changing the number of clusters by using the statistical significance as a standard. However, it is often impossible to obtain a statistically significant volume of questionnaire data, and segmenting the data reduces the number of people in each classified cluster, making it difficult to verify the statistical significance. Clustering while understanding the entire distict features by employing the visualization will be helpful in finding and analyzing majority and minority groups.



(a) Cluster 3



(b) Cluster 8

Fig. 6. Average grade (Cluster obtained through FCM)

IV. CONCLUSION

In this report, we proposed a method of visualizing questionnaire data based on average grades and a method for interactively clustering respondents on the visible space. In this experiment, we employed the proposed method for actual questionnaire data related to a product for outdoor use and clustered the respondents based on the visualized results. Extracting majority and minority groups was made possible by classifying and analyzing the respondents based on the margin of error values in the visible space. We proved that the overall impression of the product α can be grasped, and that minority groups with distinct values that could develop into prospective purchasers of the product can be found by interactively clustering respondents with high margin of error values in the visible space. Future challenges include establishing a standard for classifying respondents based on margin of error values in the visible space, verifying the relationship between margin of error values and minorities, introducing fuzzy theory [7] to interactive clustering, understanding the characteristics of the detected minority groups and comparing the results with those of respondent data classified by using, for example, rough sets [8].

REFERENCES

[1] Katsuo Inoue: How to Use Multivariate Analysis (Tsukuba Press, 2002)
[2] Noboru Yamada, Yasutaka Yamamoto, Tomohiro Yoshikawa, Takeshi Furuhashi: Classification of SD Evaluation Data by Clustering Based on Data Distribution Structure of the Data (Vol. 7 No. 2, Japan Society of Kansei Engineering Thesis, pp. 381-390, 2007)
[3] Hidetoshi Tatematsu, Tomohiro Yoshikawa, Takeshi Furuhashi, Hirohito Iguchi, Eiji Hirao: Study on Interest in Evaluation Subjects, 23rd Fuzzy System Symposium Speeches, 2007
[4] Toshikazu Fukami, Tomohiro Yoshikawa, Takeshi Furuhashi, Ioki Hara, Takuya Mochizuki: Analysis of Questionnaire Data based on Individuality of Frequency of Using Evaluation Values, International Conference on Kansei Engineering and Emotion Reserch 2007, 2007
[5] Takayuki Saito, Hiroshi Yadohisa: How to Analyze Relevant Data (Kyoritsu Press, 2006)
[6] Sadaaki Miyamoto: Cluster Analysis (Morikita Press, 1999)
[7] Michio Sugano: Developing Fuzzy Theory (Science Press, 1989)
[8] Norihiko Mori, Hideo Tanaka, Katsuo Inoue: Rough Set and Sensitivity (Kaibundo Press, 2006)