A support tool for composing questionnaires in social survey data archive SRDQ

Koichi Higuchi, Norihisa Komoda, Shingo Tamura, Yoshitomo Ikkai
Graduate School of Information Science and Technology
Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871,
Japan

Abstract: When new social survey questionnaires are composed, it is helpful to create a "summary of question items," a table in which similar question items used in past surveys are tabulated. It can take more than one week to manually prepare a summary of question items from three or four surveys. In this research, a support tool that creates a summary of question items accurately and automatically for questionnaires stored in the "SRDQ" social survey data archive is developed. We have also constructed a man-machine interface system that allows the user to edit the summary of question items created automatically with the proposed method. The system includes a function to help the user locate incorrect items by showing potentially incorrect items in a different color. Through the generation tests of "summary of question items," it is confirmed the creation time of a summary of question items with the assistance of the proposed tool has been reduced to approximately one tenth of the time required to manually create a summary.

Key-Words: Social survey data archive, SRDQ, Summary of question items, Jaccard Coefficient

1 Introduction

In recent years, data from social surveys is being collected and published by data archives for use in secondary analysis. When utilizing this type of data archive to compose new survey questionnaires, it is useful to create a summary of question items covering a specific group of surveys. Such summaries are frequently created by hand. However, it can take approximately one week to manually create a summary of question items that covers just a few surveys.

To reduce the effort of summary generation, we have developed a new tool to assist in organizing summaries of question items from questionnaires. This tool includes functions for automatically creating a summary of question items and an interface for manually editing the automatically created summary in order to produce a final, completed summary.

Various methods are proposed for calculating the level of similarity among corresponding documents. Those methods that are widely used include the Jaccard coefficient method and the cosine similarity method[1-2]. However, with the social survey questionnaires that are the subject of this study, the question items consist of a string of nouns. In some cases, the overall ratio of matching words in two similar questions may be low, resulting in an incorrect judgment of dissimilar. Likewise, two dissimilar question items may differ only in a few words that are key to the meaning of the questions, and be incorrectly judged as similar. Also, there are cases in which question items with the same meaning may not be correctly seen as similar due to differences in the wording or expressions used in the questions. Accordingly, it is not possible to correctly judge the similarity of question items in social survey questionnaires based solely on the judgment of the ratio of matching words in the question items.

Therefore, this study proposes a new method for judging similarity using new similarity indexes based on adjusted existing Jaccard coefficients. The adjustments will be applied when specific structural characteristic conditions are satisfied. A user interface is also constructed that allows the user to edit a summary of question items that has been automatically generated by the proposed method. This interface also displays potentially incorrect judgments in different colors so that the user can easily locate detection errors in the automatically generated summary.

2 Effective use of past social surveys

2.1 Social survey data archives

A data archive is an institution that specializes in the collection, editing and processing, storage, and dissemination of research data. A data archive collects research data when it is submitted for archiving or when research has been conducted by the data archive itself. Data archives must publish collected data in a format designed to enable secondary analysis, and all collected data is therefore edited and processed into a consistent format. Data archives store valuable data that can be used to perform secondary analyses and disseminate that data so that it can be used effectively[3].

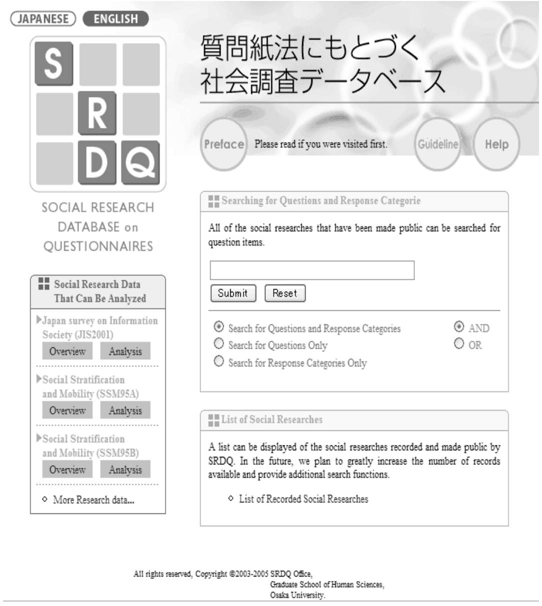
In the US, soon after World War II, the IBM card-format survey research data prepared by Elmo Roper was donated to Williams College and a data archive of public opinion surveys was created. This data archive was the predecessor to what is now the Roper Center for Public Opinion Research. Later, in 1960, Zentral archiv (ZA) was established in West Germany as a research institute of the University of Cologne. Since the mid-1960s, data archives have been established throughout North America and Europe[4].

Data archives are important for research insofar as they allow data from past surveys stored in the archive to be effectively used in advancing research. At first, it helps maintaining the quality of social surveys. When composing questionnaires for new surveys, it is imperative to review question items and data set of existing surveys for maintaining the quality. Data archives will help researchers to review the past surveys. Second, large amounts of survey costs can be eliminated by the effective use of existing data. Also, as secondary analyses of past survey data is conducted, it reduces the need to conduct repetitious surveys for similar purposes, and makes it unnecessary for respondents to go through the trouble of responding to multiple surveys with similar questions.

Data archives are also important for education insofar as they make it possible to develop social survey methodology lessons using high quality data. Social survey method instructors often explain, based on inferential statistics, that “the null hypothesis will be rejected at a level of significance of 5%,” but this explanation has special significance when analyzing survey data based on random sampling. Without data archives, it is difficult for students to get and / or analyze such data[5].

2.2 SRDQ

"SRDQ" stands for the Social Research Database on Questionnaires, and is one of the most advanced social survey data archives in Japan that was developed by the Graduated School of Human Sciences of Osaka University in 2003[6]. SRDQ is publishing social survey data and related information through WWW. Fig.1 shows the top page of SRDQ[7].


JAPANESE
ENGLISH
SRDQ
質問紙法にもとづく
社会調査データベース
Preface
Guideline
Help
SOCIAL RESEARCH DATABASE on QUESTIONNAIRES
Social Research Data That Can Be Analyzed
Japan survey on Information Society (JIS2001)
Overview
Analysis
Social Stratification and Mobility (SSM95A)
Overview
Analysis
Social Stratification and Mobility (SSM95B)
Overview
Analysis
More Research data...
Searching for Questions and Response Categorie
Submit
Reset
List of Social Researches
List of Recorded Social Researches

Please read if you were visited first.

All of the social researches that have been made public can be searched for question items.

● Search for Questions and Response Categories ● AND
○ Search for Questions Only ○ OR
○ Search for Response Categories Only

A list can be displayed of the social researches recorded and made public by SRDQ. In the future, we plan to greatly increase the number of records available and provide additional search functions.

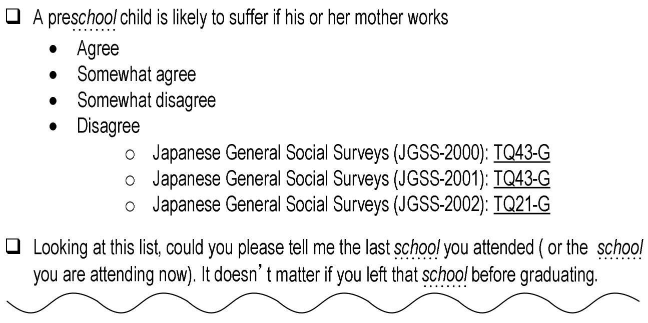
All rights reserved, Copyright ©2003-2005 SRDQ Office,
Graduate School of Human Sciences,
Osaka University.

Fig.1: The top page of SRDQ

Visiting SRDQ, it is possible to view questions and multiple-choice selection items that were used in various previous social surveys. Information about each survey such as subjects, sample designs, reports and papers are also available. Currently, 119 surveys are stored in SRDQ. Number of question items included in those surveys are 17,232. The only thing required to view these hierarchical textual data is a web browser.

With previous social survey data archives in Japan, searches were limited to brief, generalized information such as the research title and research subject. Using the string search function of SRDQ, the content of a social survey can be fully searched for question items and general information. Fig.2 is an example of search result.

Also, SRDQ allows the statistical analysis of survey data over the WWW. Since survey data can be directly analyzed, researchers can instantly obtain the analytical results they are seeking.



❑ A preschool child is likely to suffer if his or her mother works

- Agree
- Somewhat agree
- Somewhat disagree
- Disagree
    - Japanese General Social Surveys (JGSS-2000): TQ43-G
    - Japanese General Social Surveys (JGSS-2001): TQ43-G
    - Japanese General Social Surveys (JGSS-2002): TQ21-G

❑ Looking at this list, could you please tell me the last school you attended (or the school you are attending now). It doesn' t matter if you left that school before graduating.

Fig.2: Result of string search for “school”

3 Purpose of the study

To make SRDQ more useful, we planed to add a new function to help researchers in composing new questionnaires. Fig.3 shows typical procedures to compose a new questionnaire. As Fig.3 indicates, when composing new survey questionnaires, it is often necessary to ascertain how question items have differed among other surveys, so summaries of existing question items covering a specific survey group are frequently created by hand.

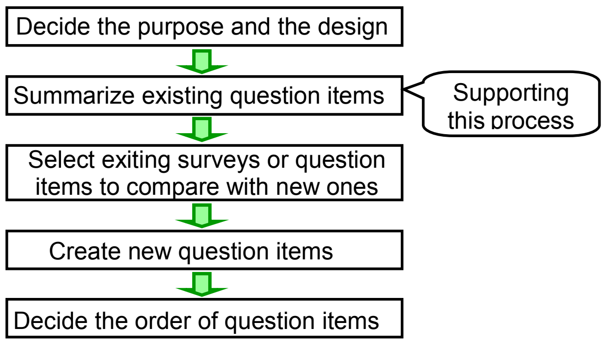
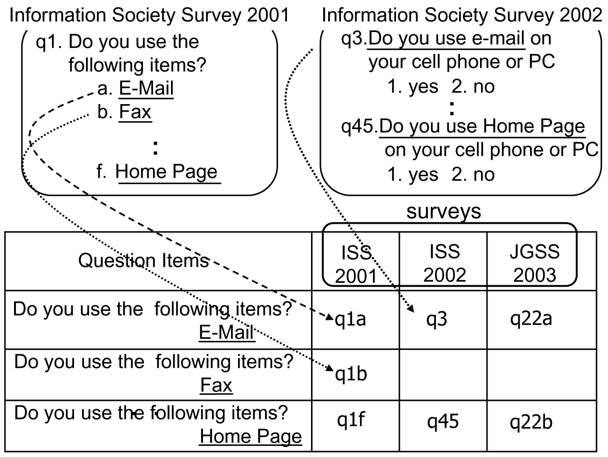

Decide the purpose and the design
Summarize existing question items
Supporting this process
Select exiting surveys or question items to compare with new ones
Create new question items
Decide the order of question items

Fig.3: Procedures to compose a new questionnaire

However, it can take approximately one week to manually create a summary of question items that covers just a few surveys. To reduce the effort of summary generation, we have developed a new tool in this study.

The bottom half of Fig.4 shows one part of a summary of question items. The first column is a list of the actual question items, that is, the question numbers (q1a, etc.) correspond to the question items on the left in the table. If a question item that is similar to the question item shown in the first column is included in one of the corresponding surveys, the question number for the similar question is shown in column 2 or subsequent columns. This table makes it possible to assess the differences in the question items of each survey, and is extremely useful in helping to determine question items for new surveys. For example, the “Do you use faxes?” question in line 3 is the only question item used in just one survey, from which we can understand that a change has taken place - this question is not being included in recent surveys.


Information Society Survey 2001
q1. Do you use the following items?
a. E-Mail
b. Fax
:
f. Home Page
Information Society Survey 2002
q3.Do you use e-mail on your cell phone or PC
1. yes 2. no
:
q45.Do you use Home Page on your cell phone or PC
1. yes 2. no
surveys

Question Items	ISS 2001	ISS 2002	JGSS 2003
Do you use the following items? E-Mail	q1a	q3	q22a
Do you use the following items? Fax	q1b		
Do you use the following items? Home Page	q1f	q45	q22b

Fig.4: Summarizing question items

With a summary of question items, similarity judgments are made for the minimum units of each question item, and if an item is judged as being similar, that item’s question number is shown in the corresponding row of the table. The upper half of Fig.4 shows the portions of the question items that are the minimum units used for similarity judgment. The minimum units are indicated by under lines in the figure.

When using a computer to create a summary of question items that will be exactly the same as a summary that is manually created by a social survey specialist, the following two user requirements must be satisfied.

- The automatic creation of the summary that is sufficiently accurate to meet the demands of social survey specialists.
- And, the provision of the editing interface to correct the errors and to produce a final, completed summary in less time.

Taking these guidelines into consideration, we developed functions for automatically creating a summary of question items and for editing the generated summary.

4 Automatic creation of the summary

The overview of the automatic summarization system developed in this study is shown in Fig.5. First, the user must specify several surveys. And if the user wants to focus on specific topics, keywords are entered. Then, specific question items are extracted from the specified surveys, and words are extracted from the question items using morphological analysis tool “ChaSen” [8-9].

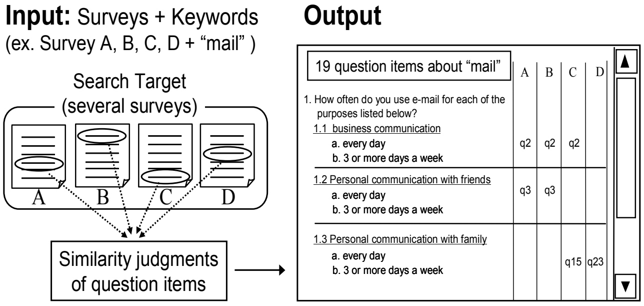

Input: Surveys + Keywords
(ex. Survey A, B, C, D + “mail”)
Search Target
(several surveys)
A
B
C
D
Similarity judgments of question items
Output
19 question items about “mail”
A
B
C
D
1. How often do you use e-mail for each of the purposes listed below?
1.1 business communication
a. every day
b. 3 or more days a week
q2
q2
q2
1.2 Personal communication with friends
a. every day
b. 3 or more days a week
q3
q3
1.3 Personal communication with family
a. every day
b. 3 or more days a week
q15
q23
▲
▼

Fig.5: Overview of the automatic summarization tool

Similarity judgments are made for question items from different surveys, and the level of similarity between questions from the same survey are not used to make similarity judgments. For similarity judgments, the similarity level values calculated between question items from all of the corresponding surveys are used. Then, the pair which has maximum similarity value will be judged as “similar”. The last step will be repeated while similarity values are higher than the set threshold value.

However, the following factors make this type of judgment difficult. First, if just one “core word” in a question item differs, the intended purpose of the question item can change completely. For example, only one word differs in the following two question items, but the intended purposes of the questions are different:

“How often do you use e-mail for personal communication with *friends*?”

“How often do you use e-mail for personal communication with *family*?”

The second factor is just the opposite - different words might be used to ask the same thing. For example, the following two question items are phrased very differently, but ask the same thing and are similar question items.

“Do you perform following actions in your daily life? Reuse bathwater for laundering to conserve water”

“Do you try to do things in this list? Saving resources such as water”

A proposed method for judging similarity is based on the existing Jaccard coefficient. The similarity judgment method using the Jaccard coefficient is a typical method for judgment of similarity between two sentences. The Jaccard coefficient is a percentage of the number of common words by the number of total words in two sentences. For pairs of relatively similar question items within one questionnaire, if only a few words differ between the two items, that words are recognized as an “core words”. If there is any discrepancy between these core words, a penalty is applied to the level of similarity. This adjustment is based on a premise that a questionnaire dose not contain multiple questions which ask the same thing. There is also significance to the order of the question items in a questionnaire, and question items having the same meaning tend to be arranged in the same order in most questionnaires. Accordingly, if the question items coming before and after a question item pair under consideration are similar between the compared questionnaires, a bonus is applied to the level of similarity assessed for the corresponding question item pair.

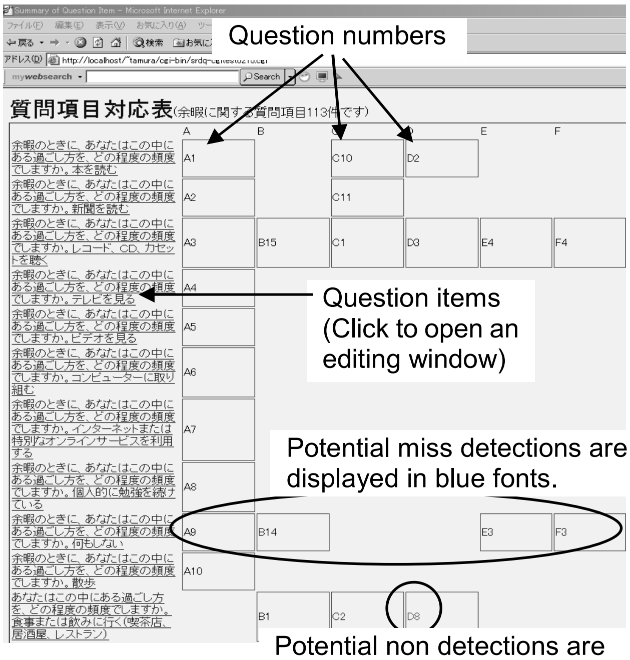
5 Editing interface

A man-machine interface system is provided with which the user can edit a summary of question items that was automatically generated by this tool in order to create a final, completed summary. The prototype system is build as CGI script using Perl. And this system features the following two functions.

5.1 Displaying a summary of question items

As shown in Fig.6, a summary of question items based on the proposed method is displayed.

This system has functions for displaying information that allows the user to easily locate potential detection errors in the automatically generated summary. If the value of adjusted Jaccard coefficient exceeds a specific threshold value, the item is judged as being similar. However, if an item exceeds this threshold but the value is close to the threshold value, it is possible that the judgment is a “miss detection” in which an item that is not actually similar was incorrectly judged as being similar. Conversely, if an item does not exceed the threshold but is close to the threshold value, it is possible that the corresponding judgment is a “non-detection” in which an item that is actually similar could not be judged as similar. Items that could possibly be a miss detection or non-detection are displayed in different colors that represent each.


Summary of Question Item - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツー
戻る 検索 お気に
アドレス(D) http://localhost/~tamura/cgi-bin/srdq-cgitest0210.cgi
mywebsearch Search

質問項目対応表(余暇に関する質問項目113件です)

	A	B	C	D	E	F
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。本を読む	A1		C10	D2		
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。新聞を読む	A2		C11			
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。レコード、CD、カセットを聴く	A3	B15	C1	D3	E4	F4
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。テレビを見る	A4					
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。ビデオを見る	A5					
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。コンピューターに取り組む	A6					
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。インターネットまたは特別なオンラインサービスを利用する	A7					
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。個人的に勉強を続けている	A8					
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。何もしない	A9	B14			E3	F3
余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。散歩	A10					
あなたはこの中にある過ごし方を、どの程度の頻度でしますか。食事または飲みに行く(喫茶店、居酒屋、レストラン)		B1	C2	D8		

Question numbers

Question items (Click to open an editing window)

Potential miss detections are displayed in blue fonts.

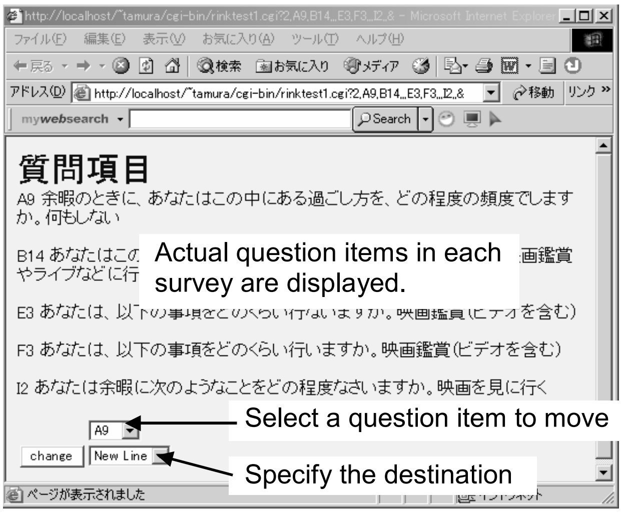
Potential non detections are displayed in red fonts.

Fig.6: Display of a summary of question items

5.2 Reorganizing incorrectly judged items

With the summary display functions, potential non-detections and potential miss detections are each shown in different colors, but there may also be other incorrectly judged items that are not included in these potential incorrect judgments. Therefore, the system is equipped with a function that can be used to reorganize any question item that the user has determined to be an incorrect judgment. This function can be applied to all of the question items in the summary.

By clicking a question items in the summary, the editing window shown in Fig.7 will open. In this window, all actual question items placed in same row of the summary will be displayed so that the user can easily check whether there are detection errors or not. If there is a detection error, the user can reorganize the summary by moving a question item to the other row or to a new row.


http://localhost/~tamura/cgi-bin/rinktest1.cgi?2,A9,B14,,,E3,F3,,,I2,,& - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り メディア
アドレス(D) http://localhost/~tamura/cgi-bin/rinktest1.cgi?2,A9,B14,,,E3,F3,,,I2,,&
移動 リンク
mywebsearch Search
ページが表示されました

質問項目

A9 余暇のときに、あなたはこの中にある過ごし方を、どの程度の頻度でしますか。何もしない

B14 あなたはこの 画鑑賞やライブなどに行

E3 あなたは、以下の事項をどのくらい行ないますか。映画鑑賞(ビデオを含む)

F3 あなたは、以下の事項をどのくらい行いますか。映画鑑賞(ビデオを含む)

I2 あなたは余暇に次のようなことをどの程度なさいますか。映画を見に行く

A9

change New Line

Actual question items in each survey are displayed.

Select a question item to move

Specify the destination

Fig.7: Editing window

5.3 Evaluation test

A summary of question items was created using the proposed tool applied to question items extracted manually from the data in the SRDQ social survey data archive. First, tests were conducted to verify if a summary of question items generated automatically based on the proposed similarity judgment method would be accurate enough to satisfy the user. Then, the amount of time required for the user to edit and complete the summary generated by the proposed method was compared with the amount of time required to manually create the same summary.

Based on the results of interviews with social survey specialists, we have determined that the proposed method will be suitably practical if the number of incorrect rows is kept to a maximum of 10% of the total number of rows in the summary. And proposed similarity judgment method has satisfied this number.

For time comparison, 113 question items from 10 surveys related to leisure were used. Time required to manually create a summary of question items was 3 hours. And time required to create a summary of question items using the proposed tool was 20 minutes.

Two types of work are involved in the creation of a summary table using the proposed tool - viewing question items in order to locate incorrectly judged items, and reorganizing the incorrectly judged items. In this study, the viewing process took 15 minutes, and the reorganization work took five minutes, a great reduction in the amount of time required to create a summary of question items. The summary of question items created using the proposed method contained three rows of incorrect judgments, and 10 incorrectly judged items in these three rows were reorganized. With the data used in this test, there were six potential non-detections and 22 potential miss detections. Both actual non-detections and actual miss detections were found in these potential incorrect judgments, and this is thought to be a factor in reducing the overall amount of time required to locate incorrect judgments.

6 Conclusions

In this study, a tool was developed for automatically summarizing and organizing similar question items included in social survey questionnaires. This tool is also equipped with the interface system for editing the generated summary in order to produce a final, completed summary. For the similarity judgment of question items, a method of similarity judgment has been proposed that is based on a new method of calculating similarity in which the structural characteristics of social survey questionnaires are used, and adjustments are made to existing Jaccard coefficients if specific conditions are met.

Evaluation tests were performed to determine if the proposed similarity judgment method could be used to create a summary of question items with the level of accuracy required by social survey specialists. Comparing the proposed method with the method of calculating similarity using only Jaccard coefficients, the results with the proposed method had far fewer incorrectly judged items. The number of incorrect judgment rows was also kept at a level capable of satisfying the requirements of the user, and the effectiveness of the proposed method could be verified. Furthermore, the amount of time required to create a completed summary with this tool was far less than the time required to manually create a summary, and the practicality and usefulness of the proposed tool have been verified.

Acknowledgments:

We would like to express our deep appreciation to Professor Atushi Naoi of the Graduate School of Human Sciences, as well as to the members of the SRDQ Office, for offering their opinions on methods for utilizing the SRDQ and for providing data used in this study.

References:

[1] M. R. Anderberg; *Cluster Analysis for Applications*, Academic Press 1973.

[2] D. Sullvian, *Document Warehousing and Text Minig*, John Wiley & Sons, 2001.

[3] A. Bryman; *Social Research Methods*, Oxford Univ. Press 2001.

[4] J. M. Barry & B. W. Roper, The Roper Center: the World's Largest Archive of Survey Data, *Reference Services Review*, 16(1), pp. 41-50, 1988.

[5] K. J. Kiecolt & L. E. Nathan, *Secondary Analysis of Survey Data*, Sage 1986.

[6] K. Higuchi & A. Kawabata, SRDQ (Social Research Database on Questionnaires) Data Archive, in *Abstracts of 36th World Congress of the International Institute of Sociology*, pp.584-585, 2004.

[7] http://srdq.hus.osaka-u.ac.jp/en/, 2003.

[8] http://chasen.naist.jp/, 2000.

[9] M. Asahara, Y. Matsumoto, Extended Models and Tools for High-performance Part-of-Speech Tagger, in *Proceedings of COLING 2000*, 2000.