Proceedings of International Conference on Technology and Social Science 2018 (ICTSS 2018)
Invited Paper

Optimization of Attributes Elimination Order in a Privacy-Enhanced Survey System

Atsushi Iwai

Faculty of Social and Information Studies, Gunma University, 4-2 Aramaki-mach,
Maebashi City, Gunma Prefecture, 371-8510, Japan
iwai@gunma-u.ac.jp

Keywords: attributes elimination, privacy preservation, social survey, anonymity

Abstract. This study presents a design for the optimization of the attributes elimination order in a privacy-enhanced survey system. The target survey system comprises a framework that analyzes the input data to find elements that can cause an information leakage and a mechanism to correct such flaws by modifying the questionnaire design in the database. The original framework employs the method of eliminating attributes in the modification of the questionnaire design, and the order of the elimination must be predetermined by researchers. This study presents a new design for automatizing the selection of the best elimination order.

1. Introduction

This study presents a design for the optimization of the attributes elimination order in a privacy-enhanced survey system. By expanding on survey systems that were designed in previous studies, a formal description of the new enhanced survey system design is to be demonstrated.

The target survey system comprises a framework that analyzes the input data to find elements that could cause an information leakage and a mechanism to correct such flaws by modifying the questionnaire design in the database. Technical tools for enhancing privacy such as k-anonymity by Sweeny ([6]) and l-diversity by Machanavajjhala et al. ([5]) are widely known. However, these frameworks do not address the problem of protecting personal information from the survey assessors. In several survey projects that focus on evaluation by users, such as class evaluation or hospital evaluation, the privacy from organization staff is important for preventing the deterioration of the quality of the obtained data. The target system is expected to be advantageous in the case of this type of surveys.

The original framework, however, employs the eliminating attributes method in the modification of the questionnaire design, and the order of the elimination should be predetermined by researchers. This study presents a new design for automating the selection of the best elimination order.

2. Previous Studies

This study focuses on the typical survey system design presented in the previous studies [1], [3], [4]. The basic design approach is exemplified as follows for the purpose of illustrating the inherent issues: A course evaluation was conducted in a small class comprising 3 male and 15 female students with a single question sheet that contained a question concerning gender and other questions regarding the course evaluation. This could be potentially harmful to the male students' privacy and could result in a deterioration in the quality of the obtained data. However, if the question sheet was divided into two parts, with one part including only the gender question and the other part only the course evaluation questions, then no privacy problem would arise to compromise the quality of the students' answers. The target system processes this division operation after all the students have finished responding to the questionnaire and when it finds problematic questions that can lead to information leakage. The division process is realized as a database operation for modifying the table structure related to the questionnaire design. As the computational process is triggered automatically and is perfectly completed before a lecturer obtains the output of the system, no information leakage is possible.

This system design is based on the hypothesis that all the questions in a question sheet can be divided into two categories, X and Y. X is defined as a category comprising individual attributes, such as gender or age. Y is defined as a category comprising individual attitudes such as the course evaluation. For each Y category question, a cross tabulation of several X-category questions is likely to yield special cells wherein only a small number of respondents exist, and these cells are likely to cause some unintended information leakage. In surveys with multiple X-category questions, the question sheet is divided by considering each X item one by one, i.e., the process of protecting privacy takes the form of attributes elimination.

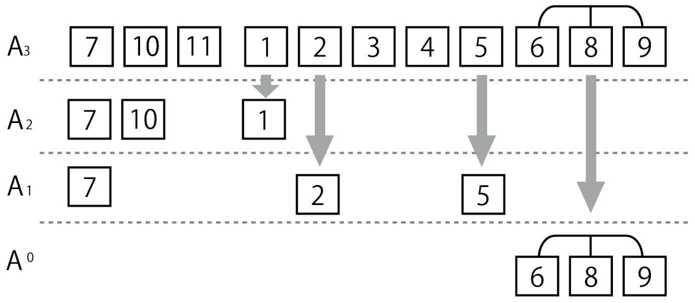

Figure 1 An Example of Questionnaire Design Modification Process ([1])

Figure 1 illustrates an example of a questionnaire design modification comprising the process of the attributes elimination ([1]). In this example, the question sheet has three X items: q7, q10, and q11, and the rest of the questions are all Y items. The researcher is interested in the X items in the order of q7, q10, and q11. The answer dataset for each of the Y items first has a relationship with the three X items. However, according to the anonymity level examined, it drops from the layer A_3 toward the layer A_0 (the layer A_n represents the answer dataset that retains its relationship with n X items.) For example, A_0 comprises the answer data for q6, q8, and q9 but has no relationship with an X item. A_1 comprises the answer data of q2 and q5 but has one relationship with an X item (q7).

In this framework, the anonymity level is calculated with a combinatorial-approach measure which takes the form of log (${}_MC_N$). It measures the degree of difficulty in dividing members into particular categories, like the positive-answer group and the negative-answer group. (In the formula of log (${}_MC_N$) M typically denotes the total number of all people and N denotes the number of people who answered positively (or negatively). Please refer to section 4 of [2] for the mathematical details of this measure.)

It should be noted that the question items should belong to the same layer if the researcher conducts a multiple variable analysis. In this example, it is assumed that the researcher intends to conduct a multiple variable analysis with the answer results of q6, q8, and q9. These questions comprise a question block, for which the result of the anonymity level test equals the result of the lowest test result obtained for each question independently.

However, in this original framework, the order of elimination should be predetermined by researchers and the result of the calculation is dependent on the order. In an unfortunate case, it would result in elimination of a large number of X category items and the limitation of obtained information. For an example, if the predetermined order leads to the elimination of all X category items for each of Y category item, the researcher will have no information about the correlation between an X category item and a Y category item.

3. Optimization Design of the Attributes Elimination Order

This section presents a new design for automating the selection of the best elimination order. The basic approach of this paper is as follows: 1) the system attempts to calculate all the possible elimination orders, and 2) it determines the result that retains the greatest amount of X-item information and outputs it.

One of the most critical points to design is the measure for evaluating the remaining X-item information. This paper employs an evaluation method of counting the remaining X-item data (each number is incremented by 1 in preparation for further use).

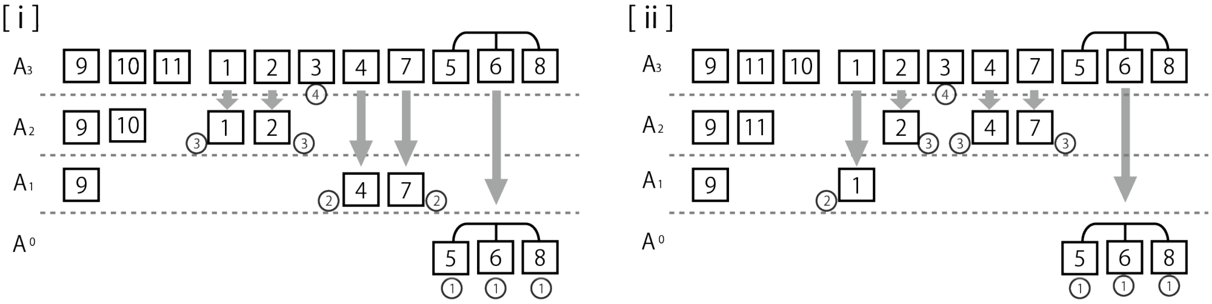

Figure 2 A Comparison of Two Modification Processes

Figure 2 shows two application examples of the counting method. In each case, it is assumed that the question sheet originally has three X items: q9, q10, and q11 but no predetermined order was set by the researcher. The rest of the questions are Y items. In case [i], the system attempted to use the elimination order of q11, q10, and q9. The summation of the X-item number for each Y-item equals 17 (each circle refers to the X-item numbers). In case [ii], the system attempted to use the elimination order of q10, q11, and q9. This result equals 18 and is better than that obtained in case [i]. The number of possible elimination orders equal |X|!, which is the same as the number of repetitions of this calculation.

Although this test does not reflect relational aspects such as the multiple analysis possibility, it can compare multiple possible output data patterns efficiently and is considered to be a reasonable comparison method.

4. Formal Description of Expanded Privacy-Enhanced Survey System

This section details the design of the expanded privacy-enhanced survey system. The essential incremental part for the purpose of this study is introducing ZMAP and ZSET functions. While the design follows the hospital evaluation version described in [4], the core functions are also applicable to other versions. Furthermore, although the formal description of some functions differs from previous related studies, without these core functions and their application in the routines, the process would be mathematically equivalent to the process described in [3] and [4] (both are accessible online).

The formal description of the system implementation is as follows:

[Sets of Respondents and Questions]

P denotes the set of all the respondents or patients. Using the respondent number $i(i \leq i \leq N)$, we can define P as follows:

$$P = \{1,2,\ldots,N\}$$

Q represents the set of all the questions in a question form. Each element of Q is classified as either X and Y items. Each element of X is classified as BPI and MD. The X items $x_1,x_2,\ldots,x_n$ represent individual attributes that are observable by medical staff. The Y items $y_1,y_2,\ldots,y_m$ represent individual attitudes that are not observable even by medical staff. BPI items $bpi_1,bpi_2,\ldots,bpi_\mathrm{i}$ represent basic personal information and MD items $md_1,md_2,\ldots,md_\mathrm{d}$ medical data.

$$BPI = \{bpi_1,bpi_2,\ldots, bpi_i\}$$
$$MD = \{md_1,md_2,\ldots, md_d\}$$

$$X = \{x_1,x_2,\ldots, x_n\}$$
$$Y = \{y_1,y_2,\ldots, y_m\}$$

$$X = BPI \cup MD$$
$$Q = X \cup Y$$
$$Z = \{z_1,z_2,\ldots, z_n\} = ZSET(zmap),\ \ zmap \in ZMAP(X)$$

The function $ZMAP(X)$ returns the set of all possible one-to-one mappings from X to Z(=X)=$\{z_1,z_2,\ldots, z_n\}$. The function $ZSET(zmap)$ returns the set Z according to the mapping information $zmap \in ZMAP(X)$. The index numbers of Z items $z_1,z_2,\ldots, z_n$ denote the supposed priority (an item with a relatively large index number is eliminated faster in the database). Since X is a finite set, one can add a total order to X. That is to say, X can be regarded to be a totally ordered set. The function $ZMAP(X)$ transforms the total order of X. The function $ZSET(zmap)$ returns one of the possible totally ordered sets whose base set is the same as that of X and whose total order is obtained by $ZMAP(X)$.

[Respondents' Answers]

For $\forall q \in Q$, $D(q)$ represents the domain of the answer to the question q. The 3-tuple (i,q,a) indicates that a respondent $i \in P$ selected the answer $a \in D(q)$ for the question $q \in Q$. T_0 denotes the set of all such 3-tuples and contains the information of all the answers provided by all the respondents. For $\forall q \in Y$, $D_C(q)$ is defined as the set of all $D(q)$ elements that are sensitive alternatives that require their selectors to be concealed. That is, $Dc(q)$ represents the set of the alternatives of a negative evaluation.

[Question Block]

A question block is defined as a non-empty subset of Y. Different questions that belong to a question block are expected to be analyzed together using multivariate analysis methods. Each question belongs one of the question blocks. When the total number of question blocks is M ($Block_j(1 \le j \le M)$), for all $j(1 \le j \le M)$, the following holds true.

$$Block_j \neq \phi$$

$$i \neq j \rightarrow Block_i \cap Block_j = \phi$$
$$Y = \bigcup_{1 \le j \le M} Block_j$$

For each $Block_j(1 \le j \le M)$, (the initial value of) the set of answers $AB_j(1 \le j \le M)$ can be defined as follows:

$$AB_j = Select\ (T_0,\ 2,\ Z \cup Block_j)$$

Here, $Select(S_1, j, S_2)$ is a function that selects the 3-tuple of S_1 when the j-th element of the 3-tuple belongs to set S_2, and returns the subset of S_1 . Similarly, $Del(S_1, j, S_2)$ is a function that eliminates the 3-tuple of S_1 when the j-th element of the 3-tuple belongs to set S_2, and returns the subset of S_1. $Prj(S, j)$ is a function that returns the set of all j-th elements of the 3-tuples that belong to S . Similarly, $Rand(S, j)$ is a function that swaps the j-th elements of all the 3-tuples of S and returns the set that can be obtained as a result of the calculation.

[Grouping of Respondents by Individual Attributes]

For $Prj\ (AB_j,\ 2) \cap Z = \{z_1,\ z_2,\ ...,\ z_k\}$,

$$Dz\ (AB_j) = D(z_1) \times D(z_2) \times \cdots \times D(z_k)$$

is defined. As each element of $\{z_1, z_2, \ldots, z_k\}$ represents a question regarding individual attributes, the answer of $i \in Prj(AB_j,1)$ is related to one point of $Dz\ (AB_j)$ (in the case of $Dz\ (AB_j) = \phi$, we consider this point to be ϕ). The respondent group that relates to point z of $Dz\ (AB_j)$ is denoted as $Grp\ (AB_j, z)(\subset P)$(in the case of $Dz\ (AB_j) = \phi$, $Grp\ (AB_j, z) = P$).

[Finding Risky Elements]

$Level\ (AB_j, z, q)$, which represents the anonymity level observed at $z \in Dz\ (AB_j)$ and $q \in B_j$ with AB_j, is defined as follows:

$$Level\ (AB_j, z, q) = \log\ (\frac{|DATA|!}{|DATA - DATAc|! \times |DATAc|!})$$

Here, $DATA$ and $DATAc$ are the abbreviations of $DATA\ (AB_j, z, q)$ and $DATAc\ (AB_j, z, q)$, which are defined as follows:

$$DATA(AB_j, z, q) = Select(AB_j, 1, Grp(AB_j, z)) \cap Select(AB_j, 2, \{q\})$$
$$DATAc\ (AB_j, z, q) = DATA\ (AB_j, z, q) \cap Select\ (AB_j, 3, D_C(q))$$

$DATA(AB_j, z, q)$ represents the set of all data at $z \in Dz\ (AB_j)$ and $q \in B_j$ with AB_j. $DATAc(AB_j, z, q)$ represents the set of $DATA\ (AB_j,\ z,\ q)$ elements, the answer of which belongs to $D_C(q)$.

$Flag\ (AB_j)$ denotes the function used for determining whether the operation of modifying the database is required for protecting privacy with AB_j, i.e., it returns a value of 1 if the following condition holds and 0 if the condition does not hold.

$$\min_{z \in Dz(AB_j), q \in B_j, |DATAc|>0} (Level(AB_j, z, q)) < Level_{Th}$$

Here, $Level_{Th}$ represents the threshold value for determining the existence of a risk.

```
[Main Routine to Enhance Privacy]
  Step 1:
    For all $zmap \in ZMAP(X)$ do the following:
      Sub-step 1:
        For each $AB_j (1 \le j \le M)$, perform the following procedure:
          i) If $Prj\ (AB_j, 2) \cap Z = \phi$ or $Flag\ (AB_j) = 0$, then proceed to Step 2.
          ii) If $Prj\ (AB_j,\ 2) \cap Z \ne \phi$ and $Flag\ (AB_j) = 1$, $AB_j = Del\ (AB_j,\ 2,\ \{z_{|Prj(AB_j, 2) \cap Z|}\})$ and repeat Step 1.
      Sub-step 2:
          Perform the following procedure:
                i) $A_{Attributes} = Rand(Del(T_0, 2, Y),\ 1)$
                ii) For $k (0 \le k \le n = |Z|)$,
                        $A_k = Rand(\bigcup_{|Prj(AB_j, 2) \cap Z| = k} (AB_j),\ 1)$
                iii) Delete the data except for that of $A_{Attributes}, A_0, A_1, A_2, \ldots, A_n$
                iv) Record $A_{Attributes}, A_0, A_1, A_2, \ldots, A_n$
                      and the score of $\mathrm{EvalOrder} = \sum_{Layer=0}^{n}(|Select(A_{Layer}, 2, Z) + 1| \times |Select(A_{Layer}, 2, Y)|)$
  Step 2:
    i) For the maximum value of EvalOrder, output $A_{Attributes}, A_0, A_1, A_2, \ldots, A_n$.
    ii) If there are multiple $zmap$ s that give the same maximum value, one is selected at random.
```

5. Concluding Remarks

This study presents an optimization design for the attributes elimination order in a privacy-enhanced survey system. By expanding on a survey system designed in previously presented studies, a formal description of the new system was demonstrated. Although the implementation of the system described in the last section is still an ongoing task, it is expected to be feasible, as a prototype system described in [3] has already been implemented and evaluated.

The optimization function, however, can deteriorate the level of the respondents' privacy as it can enable people to guess the original input answers. There is a trade-off between the privacy of the respondents and the usability of the output data that the researchers obtain. However, from the viewpoint of the respondent's privacy, the original predetermining method is not the most secure either. For example, if the order of elimination is determined for each question at random by the system, it will be more difficult to guess the original input answer data. This is a new area of investigation and a task for the next stage of this study.

Acknowledgements

This study was supported by a grant from the Telecommunications Advancement Foundation.

References

[1] A. Iwai, "A Framework of Social Survey System that Prevents Personal Information Leakage by Automatic Modification of Questionnaire Design", in ***Proceedings of 18th symposium on socio-information systems***, pp. 127-132, 2012.

[2] A. Iwai, "Evaluation of an Anonymity Measure as an Index of Voting Privacy", ***Journal of Socio-Informatics***, Vol.5, No.1, pp11-25, 2012.

[3] A. Iwai, "Reviewing Privacy-Enhanced Social Survey System that Employs Combinatorial Anonymity Measure", ***IMECS2016 (International Multi-Conference of Engineering and Computer Scientist 2016) proceedings***, pp.311-316, 2016.

[4] A. Iwai, "A Design of Privacy-Enhanced Survey System that can be Used for Hospital Evaluation by Patients", ***Proceedings of ICMEMIS2017(1st International Conference on Mechanical, Electrical and Medical Intelligent System)***, 2017.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity", ***Proceedings of IEEE International Conference on Data Engineering (ICDE)***, pp. 24–35, 2006.

[6] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", ***International Journal of Uncertainty Fuzziness and Knowledge Based Systems***, Vol. 10, No. 5, pp. 55, 2002.