Proceedings of International Conference on
Technology and Social Science 2021 (ICTSS 2021)

New Formalization Approach for a Privacy-Enhanced Survey System

Atsushi Iwai

Faculty of 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 demonstrates a new formalization approach for a privacy-enhanced survey system. The target survey system was proposed with a unique function-based formalization and 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 new formalization approach makes it easy to compare the design of the system with that of other conventional privacy-related systems.

1. Introduction

This study demonstrates a novel formalization approach for a privacy-enhanced survey system presented in previous studies, such as [1] and [2]. Unlike k-anonymity by Sweeny [4] and l-diversity by Machanavajjhala et al. [3], the system focuses on respondents' privacy from researchers or organization staff. This type of privacy is valuable in some surveys that collect evaluation data from users, such as class evaluation or hospital evaluation, because it prevents the quality of the obtained data from deteriorating. The target system is expected to be beneficial for such evaluation surveys.

However, the design of the system was proposed with a unique formalization based on database-related functions; hence, it is difficult to compare it with other conventional privacy-related systems. This study presents a new formalization approach for the attribute elimination method and paves a way for easier comparison of the system's design with that of other conventional privacy-related systems.

2. New Formalization of Attribute-Elimination Based Privacy-Enhanced Survey System

The design of the target system 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 and age. Y is defined as a category comprising individual attitudes, such as course evaluation. For each Y category question, a cross tabulation of several X-category questions is likely to yield special cells where only a small number of respondents exist; these cells are likely to cause 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., attribute elimination is used to protect privacy. The operation is described by the following function, which is used in Sub-step 1 of Step 1 in [Main Routine to Enhance Privacy] of [2] (see Section 4 of [2] for more details):

$$AB_j = Del\ (AB_j,\ 2,\ \{z_{|Prj(AB_j,2)\cap Z|}\})$$

By contrast, the formalization of l-diversity by Machanavajjhala et al. [3] employs an algebra-based notation. In the framework, $T = \{t_1, t_2, \cdots, t_n\}$ (a set of individuals) is a table with attributes

$A_1, \cdots, A_m$. $\mathcal{A}$ denotes the set of all attributes $\{A_1, A_2, \ldots, A_m\}$ and $t[A_i]$ denotes the value of attribute A_i for tuple t. If $C = \{C_1, C_2, \ldots, C_p\} \subseteq \mathcal{A}$, then the notation $t[C]$ represents the tuple $(t[C_1], \ldots, t[C_p])$, which is the projection of t onto the attributes in C (see [3] for more details).

As the elimination of a data table attribute is equivalent to a projection onto the attributes in C, the aforementioned function-based notation is expected to naturally rewritten using an algebra-based notation. This is the basic approach briefly presented in this study. In other words, when the remaining attributes are $C_x = \{C_{x1}, C_{x2}, \ldots, C_{xp}\}$, where C is a subset of attributes of the X category, the elimination procedure can be described as the projection of t onto the attributes in C_x.

3. Discussion

As the elimination of a data table attribute is also equivalent to a type of generalization discussed in [3], the proposed approach can also be described naturally using the terms for generalization procedures, such as $t_i \overset{\star}{\rightarrow} t_i^{\star}$ and $T \overset{\star}{\rightarrow} T^{\star}$ (see Section 2 of [3] for more details). However, the basic approach described in the previous section does not cover the precise procedure for preventing privacy data leakage. A more detailed formalization scheme is required for this purpose.

4. Concluding Remarks

This article demonstrates a new formalization approach for the attribute-elimination-based privacy-enhanced survey system. As it employs a conventional algebra-based notation, the new formalization approach is expected to make it easier to compare the design of the target system with that of other conventional privacy-related systems. However, the details of the description method remain undefined. This is a task for the next stage of this study.

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, " Optimization of Attributes Elimination Order in a Privacy-Enhanced Survey System", ***Proceedings of ICTSS2018(Proceedings of International Conference on Technology and Social Science 2018)***, 2018.

[3] 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.

[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", ***International Journal of Uncertainty Fuzziness and Knowledge Based Systems***, Vol. 10, No. 5, pp. 557-570, 2002.