ICIC Express Letters
Part B: Applications
Volume **13**, Number **12**, December **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 1339–1344

PERSONAL DATA PRIVACY PROTECTION USING PROCESS MINING: FOCUSED ON A GENERAL HOSPITAL CASE

Bu-Yong Choi[1] and Nam-Wook Cho[2,*]

[1]Department of Information Technology
Korea Institute of Radiological and Medical Sciences
75 Nowon-ro, Nowon-gu, Seoul 01812, Korea
bychoi@kirams.re.kr

[2]Department of Industrial and Information Systems Engineering
Seoul National University of Science and Technology
232 Gongneung-ro, Nowon-gu, Seoul 01811, Korea
*Corresponding author: nwcho@seoultech.ac.kr

Received April 2022; accepted June 2022

Abstract. *As the risk of personal information leakage in medical institutions has increased, the protection of personal information requires careful attention. However, the research on personal information protection in medical institutions is still limited to qualitative approaches, such as surveys or regulations. This paper proposes a methodology for detecting and preventing personal information leakage by using a process mining technique based on the log data collected from a general hospital. A process mining technique has been utilized to construct and analyze process models. An outlier detection technique has been presented to detect outliers that might impose risks to privacy protection effectively. An experiment has been conducted to show the effectiveness of the proposed methodology. This paper is expected to provide an effective way to protect sensitive personal information in the healthcare industry.*
Keywords: Process mining, Privacy protection, Medical information, Outlier detection

1. **Introduction.** In today's information society, vast amounts of personal data are collected, stored and processed [1]. Although personal data are generally used for the benefit of the community, they can also be easily abused. Despite various efforts, privacy protection still faces challenges, causing damage to both institutions and individuals [2].

Since much of the data in the healthcare industry has a sensitive nature, the importance of privacy protection cannot be overestimated, requiring a higher level of management and security. In the case of medical information managed by hospitals, patient treatment and physical characteristics recognized for therapeutic purposes, past medical history, and even family medical history are collected and utilized. If patient information is leaked for abnormal purposes, the impact on individuals is fatal, which can seriously interfere with daily life. A critical issue in implementing security for streaming health information is to offer data privacy and validation of a patient's information over the networking environment in a resource-efficient manner [3].

Recent studies on process mining have paid much attention to privacy issues [4]. Batista et al. presented a privacy-preserving process mining method based on a micro aggregation technique [5]. A group-based privacy preservation technique was proposed in [6] To deal with location-oriented attacks such as restricted space identification and object identification attacks, a privacy-preserving process mining technique based on the uniformization of event distributions has been presented [7].

DOI: 10.24507/icicelb.13.12.1339

However, research on process mining in healthcare has focused on enhancing the efficiency of its processes. Amantea et al. utilized a process mining technique to discover and improve the operations of the Hospital-at-Home service [8]. Maruster et al. traced frequent users of emergency medical services by process mining [9]. Therefore, more attention should be paid to privacy protection in healthcare [10].

According to [11], the major types of personal information leakage include hacking and information leakage by an internal employee. Therefore, it is urgent to establish a management system that can safely handle medical institutions' personal information and analyze hospitals' business processes.

This paper aims to propose a methodology for detecting and preventing personal information leakage by using a process mining technique based on the log data collected from a general hospital. To that end, a process mining technique has been utilized. Based on the log data, a process model has been constructed. In addition, an outlier detection technique has been presented to detect outliers that might impose risks to privacy protection effectively. An experiment has been conducted and the results have been verified through qualitative analysis.

The rest of this paper is organized as follows. Section 2 provides a research framework and the methodologies used in the paper. Section 3 explains the results of the experiments conducted with the actual data of a general hospital. Finally, Section 4 discusses the benefits and limitations of our research.

2. **Methods.** The overall procedures of our paper are illustrated in Figure 1. As shown in Figure 1, the log data have been collected from hospital information systems. Then, the data have been processed so that a process mining technique has been applied to the preprocessed data to discover a process model. Given the process model, outlying process activities in terms of privacy protection have been identified and evaluated.

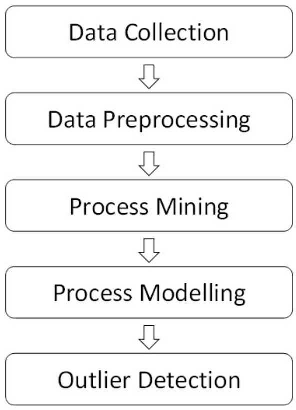

Figure 1. Research framework

2.1. **Process mining.** In this paper, process mining techniques were utilized to analyze the event logs of medical institutions. Process mining is a set of data-driven methods for diagnosing and enhancing business processes [12]. Process mining aims to create a consistent and explicit process model given an event log [13-15]. It includes the identification and diagnosis of issues between activities [16]. In this paper, DISCO has been utilized to discover a process model from log data.

2.2. **Process modeling.** As a result of process mining, a process model can be generated. In this paper, a process model is converted into a relevance matrix, representing the weighted relations between activities. Let us suppose an activity (A) is defined as a node, and the connection between activities is defined as a link. Then a process model can be converted into a graph with vertex set $U = \{A_1, \ldots, A_N\}$ and a weighted adjacency matrix, called a relevance matrix (RM), is constructed. The RM is a square $N \times N$ matrix and its element $RM_{i,j}$ represents the weight of a directed edge from vertex A_i to vertex A_j.

2.3. **Outlier detection.** An outlier can be defined as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [17,18]. It aims at finding abnormal observations that can be considered inconsistent with respect to the remainder of a dataset [19,20]. In this paper, an outlier is defined as an outlying activity that significantly deviates from normal activities which can cause risks in terms of privacy protection. In this paper, a relevance-based approach is proposed to detect such outliers. Three perspectives for relationship-based outlier detection were presented as follows [21].

C1: Frequency of activity occurrence
C2: Frequency of activity relationship occurrence
C3: Activities executed by appropriate resources

Then the relevance-based approach is based on the followings.

a) The activities with a low frequency of occurrence are likely to be outliers.

b) The department's activities with low relevance are likely to be outliers.

c) An activity is likely to be an outlier if its weight of link with a preceding activity is small.

d) An activity is likely to be an outlier if its weight of link with a following activity is small.

As seen in Figure 2, let us suppose an activity A_i is preceded by an activity A_{i-1} and followed by an activity A_{i+1}.

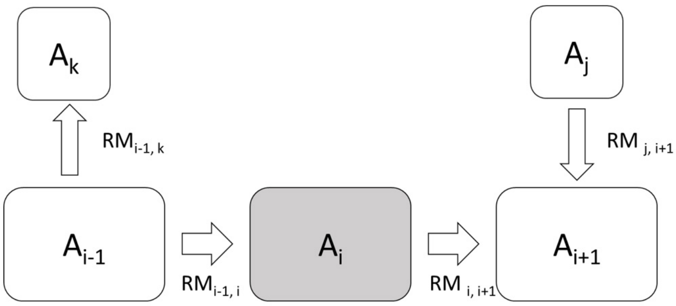

FIGURE 2. Relevance of activities

Then, the relevance score for an activity i is defined as follows:

$$Relevance\ Score[i] = F_i \times R_i \sqrt{\frac{RM_{i-1,i}}{\sum_{k=1}^{n} RM_{i-1,k}} \times \frac{RM_{i,i+1}}{\sum_{j=1}^{m} RM_{j,i+1}}}, \quad (1)$$

where F_i is a frequency of an activity A_i, and R_i is relevance of an activity A_i.

The pseudo-code of the proposed outlier detection algorithm is presented below.

```
♦ Variable
  var RM[i][j];   /* weighted adjacency matrix */
  var w, fa;      /* relevance of a department, frequency of an activity */
  var pl, nl;     /* weighted links of previous and next activities */
  var spl, snl;   /* sum of weighted links of previous and next activities */
  var rst;        /* result */
  var CV;         /* cut-off value */
```

♦ Relation-based outlier detection:

```
for i ← 1 : x                      /* Ego Activity */
    f_a ← f_a[i];                  /* Frequency of ith activity */
    w ← w[i];                      /* Weight of ith activity */
    pl ← RM[i-1][i];               /* the weighted link of previous activity */
    nl ← RM[i][i+1];               /* the weighted link of next activity */
    for k ← 1 : n
        spl ← spl + RM[i-1][k];    /* sum of out links of previous(i - 1) activity */
    for j ← 1 : m
        snl ← snl + RM[j][i+1];    /* sum of i-links of next(i + 1) activity */
    rst[i] ← f_a * w * sqrt(pl / spl * nl / snl);
    if rst[i] < BV then the activity is an outlier;
    else the activity is not an outlier.
```

A lower relevance score of an activity represents low relevance to personal data management processes, which means the activity is likely to be an outlying activity. In this paper, a cut-off value is used to determine outliers. The relevance scores are listed in descending order to determine the cut-off value. Then, a point with a significant difference between scores was selected as a cut-off value.

Suppose an example process as shown in Figure 3. The relevance score of activity i (A_i) is determined as follows.

$$Relevance\ Score[i] = 10 \times 3\sqrt{\frac{2}{40} \times \frac{5}{30}} = 2.73. \tag{2}$$

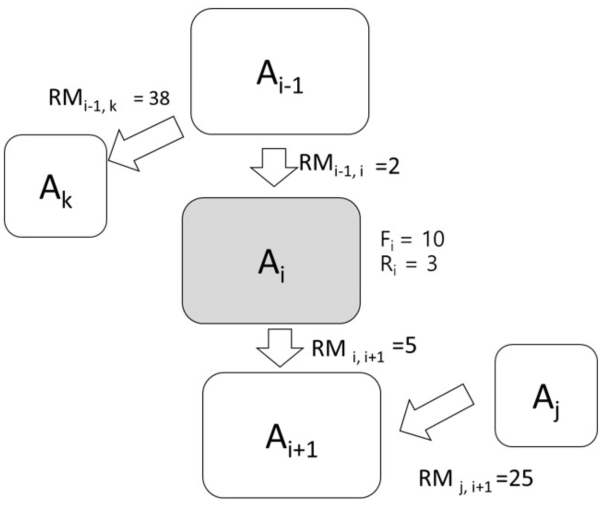

FIGURE 3. Relevance of activities: Example

3. **Results and Discussion.** An experiment has been conducted to show the effectiveness of the proposed methodology. The data used in this research were obtained from a general hospital in Seoul, South Korea. The hospital operates twenty-eight medical departments as public medical institutions, with 1,400 staff members. In the experiment, event log data of orthopedic outpatients that occurred for a month were analyzed. 2,789 events were used in the experiment.

First, a process modeling technique has been utilized to discover a process model from the given event log data. Then, the relevance matrix was generated. Before calculating the relevance score, the frequency and relevance of each activity are determined. Finally, the relevance scores of all activities are calculated and outliers are determined based on a cut-off value.

Table 1 shows the outliers detected in the experiment. Since the relevance-based outlier detection algorithm returns only theoretical outliers, verification is critical: results need to be carefully analyzed by healthcare experts to determine whether they impose actual risk in privacy protection. Activities such as 568 and 578 were identified as outliers because patient personal information should not be handled by non-medical departments, which requires careful attention from a privacy protection perspective. Activities such as 623, 1,800, and 1,238 were classified as outliers because of their low frequency.

Table 1. Experiments results

No.	Activity ID	Department	Relevance score	Note
1	623	electrocardiogram	0.02084	low frequency
2	568	R&D	0.05116	non-medical department
3	1,800	hematology	0.08034	low frequency
4	578	quality improvement	0.08075	non-medical department
5	1,238	gastroenterology	0.0825	low frequency

4. Conclusion.

This paper proposes a methodology for detecting and preventing personal information leakage by using a process mining technique based on the log data of a general hospital. Experiments were conducted with actual data from a general hospital and healthcare experts validated the results. The results show that the proposed method effectively detects outliers that might impose privacy protection risks, which is our paper's main contribution.

Despite the contribution of our study, it has some limitations. We cannot help admit that subjective judgment can determine the cut-off value. Therefore, a more sophisticated approach to determining the cut-off value can be a suitable future research topic. Another limitation is the scope of the study. Since only a portion of the entire hospital data was used, the method's validity can be limited; thus, the study might be insufficient for generalization. Consequently, more extensive experiments using real datasets should be conducted to reinforce our findings.

Acknowledgment. This paper was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean Government (MOTIE) (P0017123, The Competency Development Program for Industry Specialist).

REFERENCES

[1] B. Claerhout and G. J. DeMoor, Privacy protection for clinical and genomic data: The use of privacy-enhancing techniques in medicine, *International Journal of Medical Informatics*, vol.74, nos.2-4, pp.257-265, 2005.

[2] B. Krishnamurthy, K. Naryshkin and C. Wills, Privacy leakage vs. protection measures: The growing disconnect, *Proceedings of the Web*, vol.2, pp.1-10, 2011.

[3] S. Pirbhulal, O. W. Samuel, W. Wua, A. K. Sangaiah and G. Li, A joint resource-aware and medical data security framework for wearable healthcare systems, *Future Generation Computer Systems*, vol.95, pp.382-391, 2019.

[4] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. F. Sani, A. Koschmider, F. Mannhardt, S. N. von Voigt, M. Rafiei and L. V. Waldthausen, Privacy and confidentiality in process mining: Threats and research challenges, *ACM Trans. Management Information System (TMIS)*, vol.13, no.1, pp.1-17, 2021.

[5] E. Batista, A. Martínez-Ballesté and A. Solanas, Privacy-preserving process mining: A microaggregation-based approach, *Journal of Information Security and Applications*, vol.68, 103235, 2022.

[6] M. Rafiei and W. M. van der Aalst, Group-based privacy preservation techniques for process mining, *Data & Knowledge Engineering*, vol.134, 101908, 2021.

[7] E. Batista and A. Solanas, A uniformization-based approach to preserve individuals' privacy during process mining analyses, *Peer-to-Peer Networking and Applications*, vol.14, no.3, pp.1500-1519, 2021.

[8] I. A. Amantea, E. Sulis, G. Boella, R. Marinello, D. Bianca, E. Brunetti, M. Bo and C. Fernandez-Llatas, A process mining application for the analysis of Hospital-at-Home admissions, *Studies in Health Technology and Informatics*, vol.270, pp.522-526, 2020.

[9] L. Maruster, D. J. van der Zee and E. Buskens, Identifying frequent health care users and care consumption patterns: Process mining of emergency medical services data, *Journal of Medical Internet Research*, vol.23, no.10, e27499, 2021.

[10] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, O. A. Johnson, M. Sepúlveda, E. Helm, V. Galvez-Yanjari, E. Rojas, A. Martinez-Millana, D. Aloini and I. A. Amantea, Process mining for healthcare: Characteristics and challenges, *Journal of Biomedical Informatics*, vol.127, 103994, DOI: 10.1016/j.jbi.2022.103994, 2022.

[11] Identity Theft Resource Center (ITRC), *2018 End of Year Data Breach Report*, 2019.

[12] P. Zerbino, A. Stefanini and D. Aloini, Process science in action: A literature review on process mining in business management, *Technological Forecasting and Social Change*, vol.172, DOI: 10.1016/j.techfore.2021.121021, 2021.

[13] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd Edition, Springer, 2016.

[14] W. van der Aalst and A. Weijters, Process mining: A research agenda, *Computers in Industry*, vol.53, no.3, pp.231-244, 2004.

[15] K. Okoye, S. Islam, U. Naeem and M. S. Sharif, Semantic-based process mining technique for annotation and modelling of domain processes, *International Journal of Innovative Computing, Information and Control*, vol.16, no.3, pp.899-921, 2020.

[16] C. dos Santos Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos and E. E. Scalabrin, Process mining techniques and applications – A systematic mapping study, *Expert Systems with Applications*, vol.133, pp.260-295, 2019.

[17] D. M. Hawkins, *Identification of Outliers*, Springer, 1980.

[18] Z. He, S. Deng and X. Xu, An optimization model for outlier detection in categorical data, in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang and G.-B. Huang (eds.), Springer Berlin Heidelberg, 2005.

[19] S. La and N.-W. Cho, A study on evaluation measures for unsupervised outlier detection, *ICIC Express Letters*, vol.14, no.5, pp.515-520, 2020.

[20] S. Kim, N. W. Cho, B. Kang and S.-H. Kang, Fast outlier detection for very large log data, *Expert Systems with Applications*, vol.38, no.8, pp.9587-9596, 2011.

[21] D. Kim, *Multi-Perspective Anomaly Detection in Process Mining*, Master Thesis, Postech, 2018.