Hindawi
Journal of Control Science and Engineering
Volume 2018, Article ID 1864703, 7 pages
https://doi.org/10.1155/2018/1864703

Research Article

A New Mining and Protection Method Based on Sensitive Data

Xiaoyao Zheng[1,2], **Yuqing Liu**[3], **Hao You**[1,2], **Liangmin Guo**[1,2] and **Chuanxin Zhao**[1,2]

[1]*School of Computer and Information, Anhui Normal University, Anhui, Wuhu 241002, China*
[2]*Anhui Provincial Key Laboratory of Network and Information Security, Anhui Normal University, Wuhu 241002, China*
[3]*School of Computer Science and Technology, Soochow University, Jiangsu, Suzhou 215000, China*

Correspondence should be addressed to Xiaoyao Zheng; zxiaoyao_2000@163.com

Received 29 June 2018; Accepted 6 November 2018; Published 25 November 2018

Academic Editor: Juan-Albino Méndez-Pérez

The traditional method of sensitive data identification for data stream has a large amount of calculation and does not reflect the impact of time on the data value, and the mining accuracy is not high. In view of the above problems we firstly adopt the sliding window mechanism to divide the data flow according to time and delay the dataset according to the characteristics of the data flow in the sliding window to achieve the purpose of saving time and space. At the same time, threshold sensitivity analysis is used to find out the optimal threshold. Finally, a K-anonymous algorithm based on dynamic rounding function is employed to achieve the protection of sensitive data. Theoretical analysis and experimental results show that the algorithm can effectively mine the sensitive data in the data stream and can effectively protect the sensitive data.

1. Introduction

With the rapid development of network technology, Internet platforms such as search engines, social networks, and e-commerce have generated a large amount of data when it is convenient for users. Now it is entering the era of big data where data is explosively growing. People are paying more and more attention to the protection of personal information, and data has become the most valuable thing at the moment. This has led to the mining and protection of sensitive information, that is, through the mining of super large amounts of data to obtain important information of users. However, mining sensitive data can also lead to privacy leakage. Therefore, many researchers began to focus on sensitive data mining and protection.

Baidu Encyclopedia defines sensitive information as follows: being used for improper behavior or being released or modified by others without the consent of the parties would be unfavorable to the implementation of the national interest or government plans or unfavorable to personal privacy rights enjoyed by individuals, including personal privacy information, business management information, financial information, personnel information, IT operation, and maintenance information. Among them, the data stream has strong time characteristics, and there is also the risk of sensitive information being tampered with and eavesdropped. However, expired stream data tends to be less valuable. The identification of sensitive data based on text content is a typical application of data mining. The method proposed in [1] is based on the threshold self-learning technology to improve computing efficiency. Massive text clustering and topic extraction based on sensitive data can obtain accurate and sensitive information, but it is not suitable for mining sensitive information in social networks [2]. There are various methods for mining sensitive data in social networks, such as ensuring close privacy while publishing sensitive data [3], analyzing sensitive data transmission in android, leak detection for privacy [4], and multipart support sensitive data mining algorithm [5]. Although the above method can efficiently mine sensitive data, it ignores the most important temporal characteristics of data flow. Based on this, Li Haifeng et al. proposed the *FIMoTS* algorithm in 2012 [6], Qi Xiangxia et al. [4] presented the *FIUT-Stream* algorithm in 2013, and Yin Shaohong et al. proposed the *SWM-MFI* algorithm in 2015 [7]. These algorithms are based on sensitive data mining algorithms on the time data stream and are more in line with the characteristics of data flow in today's social networks.

This paper summarizes the advantages of the above algorithm and proposes a threshold self-learning algorithm based on the sliding window, which can ensure the shortest mining time based on the mining of accurate information.

The protection of sensitive data is to prevent data from being leaked while ensuring the usefulness of the data [8, 9]. In order to protect the user's personal information, the traditional technique is to delete the user's sensitive data before the data is released. However, it has been found in practice that this method does not protect user information well, and it is difficult for data recovery, which destroys the usefulness of data. In view of the above problems, this paper mainly uses the following steps to mine and protect the data. Firstly, we segment the data stream and extract sensitive words. Then we divide the sliding window to mine sensitive data. Next, we find the optimal threshold. Finally we use the *DIFD* algorithm to implement the protection of sensitive data.

2. Related Work

2.1. Sensitive Data Mining. The sensitive data recognition method based on text content mentioned in [1] is to judge sensitive information by simple feature selection extraction of text content and threshold determination method based on learning. The advantage is that the threshold determined by self-learning can make the accuracy of data extraction the highest, but when comparing threshold effects, a large number of calculations are generated, which greatly reduces the mining efficiency. The *FIMoTS* algorithm mentioned in [6] is more in line with the characteristics of data flow in today's social networks, which emphasizes the influence of time on the value of data. Based on the sliding window processing method, the time period is the processing unit, which increases the computational efficiency. However, the selection of the threshold value during the process is too arbitrary, and it is difficult to ensure the reliability of the mining result only through user customization. The mining algorithm used in this paper is combined with the *FIMoTS* algorithm in [1, 6]. First, the *FIMoTS* algorithm is briefly introduced as follows.

The algorithm mainly uses enumeration tree as the data structure to save data. Firstly, the enumeration tree is initialized according to the initial sliding window dataset and absolute support degree; then the algorithm uses the time characteristics of data arrival and departure to mine sensitive data and prunes the enumerate trees. Finally, the algorithm sets the upper and lower boundaries of data changes to improve the mining efficiency.

For the sliding window $S(|S| = z)$, given the data item *Item*, the relative support is set to represent the a/b $(b \geq a)$ in a fractional form, where b represents the total number of data items contained in the sliding window S and a indicates the total number of itemsets appearing in the sliding window S. Minimum support threshold $\lambda r = c/d$ $(d \geq c)$, where minimum support is set by the user, and relative support is greater than minimum support, which is sensitive data. And we use $Tree$ to save the enumeration tree. In the enumeration tree, the parent node data item is a subset of the child nodes. When a child node of sensitive data is nonsensitive data, the node is set as a leaf node. The enumeration tree uses the form $< item, sup, time >$ of the triple to represent each node's information, where *item* represents the information of each data item, *sup* represents the relative support of the data item, and *time* represents the update time of the node.

This paper improves the algorithm *FIMoTS* in the process of mining sensitive data. In the literature [1], the dataset is first processed by word segmentation. Then the frequent itemset mining algorithm based on the sliding window is used to dynamically change the threshold and finally obtain the optimal threshold. So, we rename the algorithm as threshold self-learning-sensitive data mining algorithm, namely, *SL-SDMA*.

2.2. Sensitive Data Protection. With the continuous advancement of mining technology, it is becoming easier for people to obtain sensitive data, and personal privacy is seriously threatened. Therefore, how to effectively protect the sensitive data excavated becomes another important research area. Current methods for protecting sensitive data include the privacy protection method based on K-anonymity [10], anonymized privacy protection technology based on clustering [11], and differential privacy protection [12]. Although the algorithm of [11] reduces the risk of privacy leakage to a certain extent, the proposed K-anonymity model cannot solve the problems of homogeneous attacks and background attacks. The method in [12] can deal with attackers with arbitrary background knowledge and improve the usability of data clustering. However, this method cannot solve the privacy leak security problem in distributed environments. In view of the shortcomings of the above methods and the characteristics of social network datasets, this paper proposes the optimization of the K-anonymous algorithm based on the rounding partition function [10]. Dynamically changing the processing dataset can fully reflect the time characteristics of the data flow, and K-anonymity sensitive data protection method is the earliest proposed privacy protection mechanism. After years of research, the technology has matured, and the operation is simple and has strong practicality.

3. Sensitive Data Mining and Protection

3.1. Dataset Preprocessing. The dataset studied in this paper mainly comes from online commentary. The dataset *C_comment* consists of multiple online reviews. Let $C\_comment = \{C_1, C_2, ...C_i\}$, where C_i denotes the i-th network comment. We first segment the online reviews into words through the TULAC word segmentation system developed by Tsinghua University's Natural Language Processing and Social Humanities Computing Laboratory and also identify the parts of speech of a word when it is segmenting, such as noun (_*n*), person's name (_*np*), and verb (_*v*).

Let the phrase after lexical analysis be $C_i = \{< wg_{i1}, wf_{i1} >, < wg_{i2}, wf_{i2} >, \cdots, < wg_{ij}, wf_{ij} >\}$. Among them, C_i denotes the phrase after the word segmentation of the i-th dataset, wg_{ij} denotes the i-th phrase after the lexical analysis of the j-th group dataset, and wf_{ij} denotes the part of speech of the phrase wg_{ij}. The phrase C_i was used to analyze the word and word frequency to obtain a new phrase $C_i' = \{wg_{i1}', wg_{i2}', \cdots, wg_{ij}'\}$. Among them, C_i' denotes the phrase

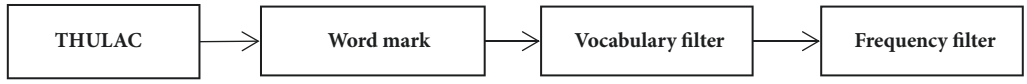


Figure 1: Data preprocessing flow chart.

after the segmentation of the i-th dataset. wg'_{ij} represents the word after the i group phrase was filtered, and the part of speech is noun. The specific steps are shown in Figure 1.

3.2. Sensitive Data Mining Algorithm. The threshold cannot be dynamically changed in the algorithm *FIMoTS*, so it is not suitable for the mining of sensitive data of various sizes of datasets, resulting in low mining efficiency. The *SL-SDMA* algorithm adopted in this paper changes the threshold value dynamically. After the enumeration tree is initialized, data items are inserted and deleted continuously as the sliding window moves. According to the algorithm *FIMoTS*, when add_n data items are inserted, the upper bound of the type change of $Item$ becomes $(u - add\_n)$. If $Item$ is sensitive data, the lower bound of the type change becomes $(b - [(c/(d - c))add\_n])$; if $Item$ is nonsensitive data, the lower bound of the type change becomes $(b - [((d - c)/c)add\_n])$. When there are sub_n itemsets removed, the lower bound of the type change of $Item$ becomes $(b - sub\_n)$. If $Item$ is sensitive data, the upper bound of the type change becomes $(u - [((d - c)/c)sub\_n])$; if $Item$ is nonsensitive data, the upper bound of the type change is $(u - [(c/(d - c))sub\_n])$. After sensitive data is mined using the type change upper and lower bounds, the mined time and sensitive data redundancy of sensitive data under different thresholds are compared, and the optimal threshold is finally obtained, which can maximize the mining efficiency.

Use Fi to save sensitive data, and a single sensitive dataset consists of $< fi\_u, fi\_b >$ and two datasets, which represent changes in the upper and lower bounds of sensitive data; use If to save nonsensitive data, and a single nonsensitive dataset consists of $< if\_u, if\_b >$ and two datasets, indicating nonsensitive data type changes. Upper and lower bounds: the *Initialize* function is used to initialize the enumeration tree. The *FIMoTS* algorithm implements mining of sensitive data. The specific algorithm steps are as shown in Algorithm 1.

3.3. Sensitive Data Protection: DIDF Algorithm. This paper optimizes the K-anonymity algorithm [10] known as *Flexible Partition* algorithm based on the rounding partition function, which regards time as an important attribute. By dynamically changing the dataset, it can guarantee the real-time performance of the data and make it independent at different time periods. The data is relatively independent, and then *Flexible Partition* algorithm processing is performed on the dataset of each time period to obtain the maximally anonymous group. The *Flexible Partition* algorithm is briefly described below.

Assuming that table T(d) contains $n = d \times k + r$ records (where k denotes K- anonymity and r is a positive integer smaller than k), theoretically table $T(d)$ can partition $k + 1$ anonymous groups. Then for any anonymous group $n' = d' \times k + r$ of $T(d)$ (where k denotes K-anonymity and r is a positive integer smaller than k), if d' is odd, the number of anonymous groups generated by the two-division method is $(d' - 1)/2 \times k + (r' + k)/2$ and $(d' - 1)/2 \times k + (r' + k)/2$, respectively. Anonymous groups can be divided into W $d' - 1$ anonymous groups.

For a large dataset $X = \alpha \times k + \beta$, it can be divided into two anonymized groups $X_1 = \alpha_1 \times k + \beta_1$ and $X_2 = \alpha_2 \times k + \beta_2$, where $\alpha_1 + \alpha_2 \leq \alpha$, and when the $\beta_1 + \beta_2 = \beta$ equation is established. Based on the above analysis, it can be seen that the rounding partition function is

$$X_1 |X_1| = \frac{d'}{2} \times k + \frac{r'}{2}$$
$$X_2 |X_2| = \frac{d'}{2} \times k + \frac{r'}{2}, \quad (1)$$

and the opening in the function is rounded up, and the opening down is rounded down. This paper takes time as an important factor to divide the dataset according to the time period and dynamically to change the processing dataset, while considering the edge data processing method. The method proposed in this paper helps to maintain the datasets relative independence in different time periods while making the protection of sensitive data more accurate with strong operability. We renamed the algorithm based on the dynamic rounding function K- anonymous algorithm as *DIDF* algorithm, and the specific algorithm steps are as shown in Algorithm 2.

The *DIDF* algorithm always tries to split the dataset of a single time period into more anonymous groups and has stronger advantages in processing the data on the boundary line, which can fully reflect the time characteristics of the data flow in the social network. For example, when $k = 2$ and dataset $|X| = 5$, $|X| = 2 \times 2 + 1$ is known based on the *DIDF* algorithm, so after the algorithm operation, two anonymous groups $|X_1| = 1 \times 2 + 0$ and $|X_2| = 1 \times 2 + 1$ are generated.

4. Comparison of Experimental Results

The experiment was run on a PC with a 1.90GHz Core i3 processor, 44GB of memory, and a Windows 8.1 operating system. The lexical analysis processing was implemented using python programming language, and the *SL-SDMA* and *DIDF* algorithms were implemented in C# on Tongcheng datasets originated from online user reviews.

4.1. Sensitive Data Mining. The dataset selected in the data mining uses two kinds of tourist reviews of the Tongcheng tourism website as an experimental dataset. Table 1 lists the data characteristics of the two datasets. The data used to support the findings of this study are available from the corresponding author upon request or the url "https://pan.baidu.com/s/1-LEzNrk9YjG8o0hOhi0WWA."

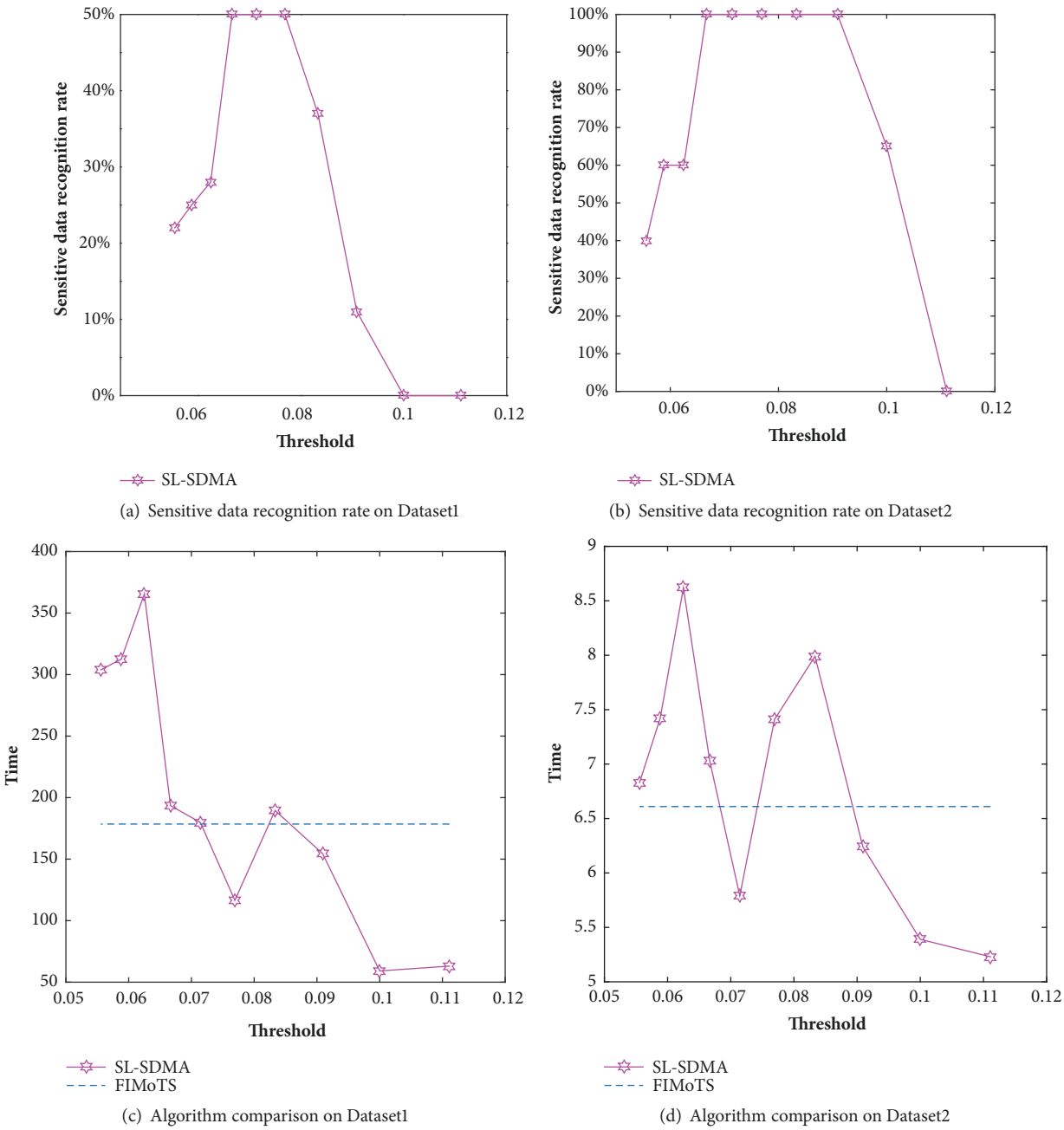
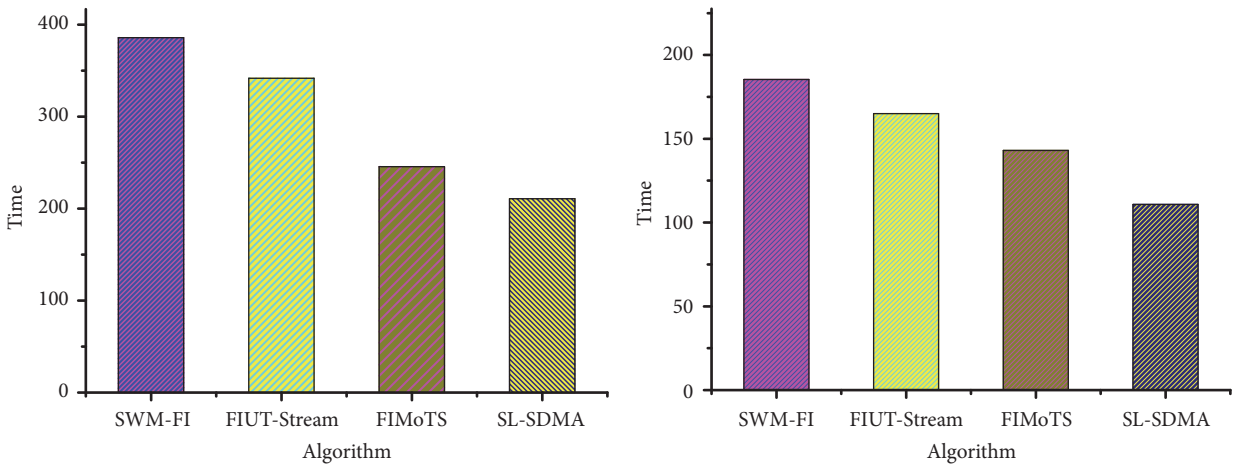


Figure 2: Threshold determination time comparison.

The experiment acquires the total data length, the longest data item length, the shortest data item length, and the running time. By modifying the threshold multiple times, the relationship between sensitive data mining time and threshold is finally determined, as shown in Figure 2.

Figures 2(a) and 2(b) show the relationship between the threshold and the sensitivity data recognition rate; there are maps showing that when the threshold range is [1/14, 1/12], sensitive data identification rate is highest, up to 100%. With the increase of the threshold, [1/11, 1/9], the standard for extracting sensitive data is increased, and complete sensitive data cannot be obtained, which leads to a decrease in the recognition rate of sensitive data; as the threshold decreases, [1/18, 1/15], the standard for extracting sensitive data is reduced, and more redundant data are obtained, which reduces the recognition rate of sensitive data.

Figures 2(c) and 2(d) show the relationship between the threshold and the extraction time of sensitive data. As can be seen from the figure, the blue dashed line is the dividing line, and the line graph of the red threshold and time is roughly divided into three parts. These correspond to the three ranges of change in the recognition rate of sensitive data in Figures 2(a) and 2(b), respectively: Firstly, when the threshold range is between 1/11 and 1/9, the running time decreases with the



(a) Mining time on Dataset1

(b) Mining time on Dataset2

Figure 3: Mining time of experiment on two datasets.

Table 1: Dataset characteristics.

Data set	Minimum data item length	Maximum data item length	Total data item length
The underwater world (Dataset1)	1	6	14900
Nanjing Presidential Office (Dataset2)	1	4	4857

Input: *items*, $z(|S| = z)$, N;
(1) *Set a threshold value*
(2) *Initialize* (*items*,λr,z);
(3) **for each** $Fi[i]$ **do**
(4) $Fi[i].fi\_b = Fi[i].fi_b - N - \left[\frac{c}{d-c}N\right]$;
(5) $Fi[i].fi\_u = Fi[i].fi_u - N - \left[\frac{d-c}{c}N\right]$;
(6) **end for**
(7) **for each** $Fi[i]$ do
(8) **If** $Fi[i].fi\_b <= 0 \parallel Fi[i].fi\_u <= 0$ **then**
(9) Update the *itemsets* and *enumerated tree*
(10) **end if**
(11) **end for**
(12) *If* [i] *the method is the same as* $Fi[i]$
(13) **for each** New node join
(14) Initialize(*items*, λr, z);
(15) **end for**
(16) *Statistical run time*
(17) *More sensitive data redundancy*.
Output: *Sensitive data*, *Optimal threshold*

Algorithm 1: Threshold self-learning-sensitive data mining algorithm: *SL-SDMA* algorithm.

increase of threshold value. However, the current sensitive data recognition rate is not the highest, and the complete sensitive data cannot be obtained. Secondly, the point at the upper left threshold is in the range of [1/18, 1/15]. The sensitive data recognition rate is 100%, and it is not difficult to see that when the threshold value is 1/13, the running time is the least, that is, the optimal threshold point, 3. The intermediate threshold value is in the range of [1/14, 1/12]. At this point, the running time of the range decreases with the decrease of the threshold, but the recognition rate of the sensitive data obtained in this range is not the highest, and there is more redundant data. Therefore, when the dataset size is 28900, the threshold value is 1/13, which can not only guarantee the highest recognition rate of sensitive data, but also guarantee the shortest running time, which is the optimal threshold point in this paper.

The experiment in Figure 3 shows the mining time of *FIMoTS*, *SWM-FI*, *FIUT-Stream*, and *SL-SDMA* algorithms. The time complexity of these four algorithms is $O((1/2)n^3)$, $O(n^2 \log n)$, $O(n^2 + An)$, $O(n^2 + Bn)$, respectively, where A and B are constants. Through the sensitive data mining experiments on the above two datasets, the running time is shown in Figure 3. However, *SWM-FI* and *FIUI-Strean* are based on sliding window to mine sensitive data. They divide sliding windows according to transaction size and do not fully consider the impact of time on the datasets in social networks. Furthermore, they use matrix to store data for storing data, which wastes a lot of space. The *SL-SDMA* algorithm proposed in this paper uses the storage structure of enumeration tree to save data. Because our algorithm stops judging when the parent node is insensitive data, it not only saves time, but also avoids the waste of space. In addition, the sliding window in our algorithm is divided according to the time, which can fully reflect the time characteristics

Input: table $T(d)$, Ω

(1) Get a set of data sets

(2) Redefine $X = \Omega$, $|X| = \alpha \times k + \beta$, where β is a positive integer smaller than k

(3) Select any one-dimensional i of Ω and find a suitable positive integer j such that $|\bigcup_{T(d)\cap X \le X(i,j)}| \ge \alpha/2 \times k + \beta/2$ and $|\bigcup_{T(d)\cap X \le X(i,j)}| \ge \alpha/2 \times k + \beta/2$

(4) X is divided from j into two multi-dimensional rectangular regions X_1, X_2

(5) The data set is divided into two data sets $|X_1|$, $|X_2|$, where $|X_1| = \alpha/2 \times k + \beta/2$, $|X_2| = \alpha/2 \times k + \beta/2$

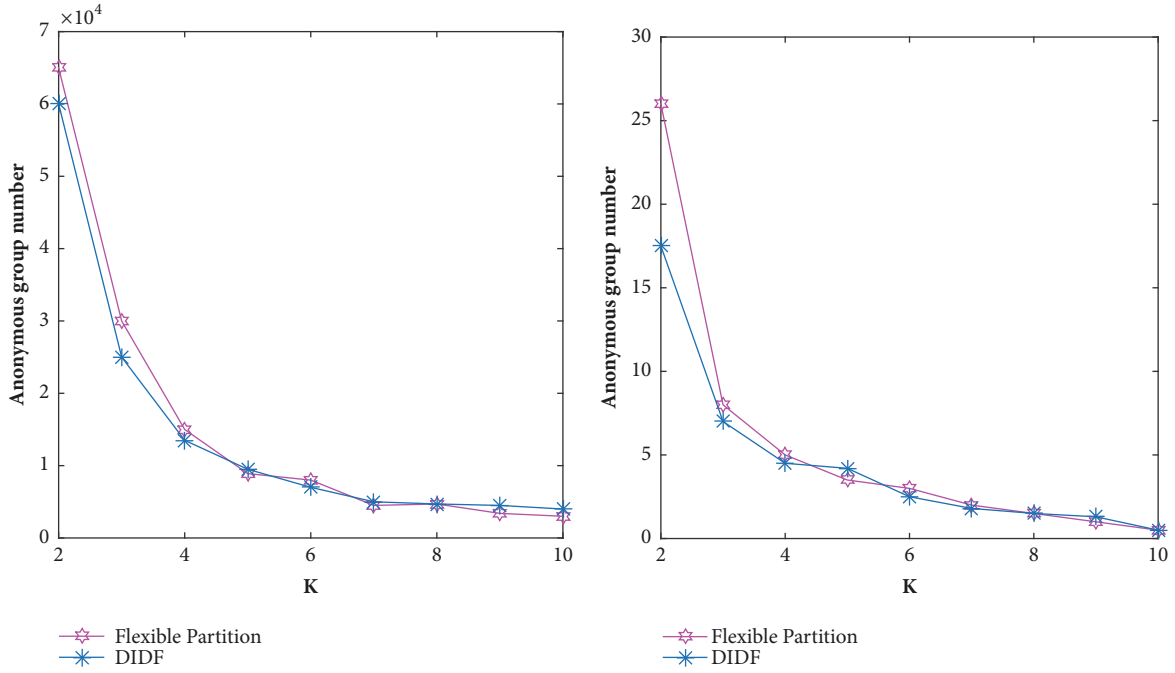
(6) If $|X_1| \ge 2k$, $|X_1|$ is recursively processed.

(7) If $|X_2| \ge 2k$, $|X_2|$ is recursively processed.

(8) Repeat (1)-(7)

Output: Anonymous group

Algorithm 2



(a) Anonymous group number comparison

(b) Time cost comparison

Figure 4: *DIFD* algorithm experimental results.

of the data stream, so the time efficiency of *SL-SDMA* is higher.

4.2. Sensitive Data Protection. The size of the dataset selected for the protection of sensitive data is 28,900. The content of the dataset includes the name of the visitor, the content of the comment, and the time of the comment. The change of the value, the number of anonymous groups obtained, and the time consumed are used to prove the feasibility of the algorithm.

The experimental comparison shows that the *DIDF* algorithm achieves the protection of sensitive data by dynamically acquiring the processed dataset. It not only can ensure the independence of the data in each time period, but also is more conducive to the centralized protection of data. The experimental results after the operation are similar to the experimental results of the *Flexible Partition* algorithm which is shown in Figure 4. That is, it is possible to obtain as many anonymous groups as possible, and the algorithm is also acceptable over time.

5. Conclusion

This paper first uses NLP's lexical data package THULAC to preprocess the dataset. Then according to the temporal characteristics of the data stream, a sliding window-based sensitive data mining algorithm is proposed, which takes the most important attributes of time and adopts the data structure of the enumeration tree. The storage of the calculation results is realized. By defining the upper and lower bounds of the data item type, the enumeration tree and the data collection information are updated only when the relative

support degree reaches the upper and lower bounds of the type change, thereby saving the calculation time. Finally, the threshold self-learning function is used to determine the threshold for finding the minimum time spent in ensuring the accuracy of mining data. This method can determine the optimal threshold in the same dataset, thereby improving the experimental efficiency. In the protection of sensitive data, using *DIDF* algorithm to dynamically change the processing dataset not only can guarantee the independence of the dataset in each time period, but also can always obtain the maximum number of anonymous groups. Experiments show that the above methods can significantly improve the computational efficiency while ensuring the accuracy of the experimental results in the mining and protection of sensitive data and consider the time characteristics of the dataset, which has strong operability and feasibility. We intend to explore several directions in future work, including extending the algorithm to deal with the frequent pattern mining on data stream in distributed environment. Furthermore, because sensitive data mining may lead to personal privacy leakage, we intend to add a differential privacy method into our *SL-SDMA* method. We intend to explore several directions in future work, including extending the algorithm to deal with the frequent pattern mining on data stream in distributed environment. Furthermore, because sensitive data mining may lead to personal privacy leakage, we intend to add a differential privacy method to our *SL-SDMA* method.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request or from the following link: https://pan.baidu.com/s/1-LEzNrk9-YjG8o0hOhi0WWA.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Natural Science Foundation of China (No. 61772034, No. 61871412) and Natural Science Foundation of Anhui Province (No. 1808085MF172).

References

[1] L. Weiwei, T. Zhang, L. Weimin et al., "Research and implementation of sensitive data identification method based on text content," *Computer Engineering and Design*, vol. 34, no. 4, pp. 1202–1206, 2013.

[2] M. Peng, J. Huang, J. Zhu, J. Huang, and J. Liu, "Mass of short texts clustering and topic extraction based on frequent itemsets," *Computer Research and Development*, vol. 52, no. 9, pp. 1941–1953, 2015.

[3] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*, pp. 473–486, Vancouver, Canada, June 2008.

[4] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang, "AppIntent: analyzing sensitive data transmission in Android for privacy leakage detection," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '13)*, pp. 1043–1054, Berlin, Germany, November 2013.

[5] L. Xiongfei, Y. Senmiao, D. Liyan et al., "A data mining algorithm based on calculating multi-segment support," *Chinese Journal of Computers*, vol. 24, no. 6, pp. 661–665, 2001.

[6] H.-F. Li, N. Zhang, J.-M. Zhu, and H.-H. Cao, "Frequent itemset mining over time-sensitive streams," *Chinese Journal of Computers*, vol. 35, no. 11, pp. 2283–2293, 2012.

[7] Y. Shaohong, S. Kunyu, and F. Guidan, "Mining algorithm research of data stream maximum frequent itemsets in sliding window," *Computer Engineering and Applications*, vol. 51, no. 22, pp. 145–149, 2015.

[8] C. Zhao, C. Wu, X. Wang et al., "Maximizing lifetime of a wireless sensor network via joint optimizing sink placement and sensor-to-sink routing," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 49, pp. 319–337, 2017.

[9] C. Zhao, C. Wu, J. Chai et al., "Decomposition-based multi-objective firefly algorithm for RFID network planning with uncertainty," *Applied Soft Computing*, vol. 55, no. 6, pp. 549–564, 2017.

[10] Y.-J. Wu, Q.-M. Tang, W.-W. Ni, and Z.-H. Sun, "Algorithm for k-anonymity based on rounded partition function," *Journal of Software*, vol. 23, no. 8, pp. 2138–2148, 2012.

[11] A. Masoumzadeh and J. Joshi, "An alternative approach to k-anonymity for location-based services," *Procedia Computer Science*, vol. 5, pp. 522–530, 2011.

[12] L. Yang, H. Zhifeng, and W. Wen, "Research on differential privacy preserving k-means clustering," *Computer Science*, vol. 40, no. 3, pp. 287–290, 2013.