Scientific Research Publishing

Journal of Computer and Communications, 2023, 11, 131-141
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Method to Improve the Accuracy of Personal Information Detection

Chih-Chieh Chiu[1], Chu-Sing Yang[2], Ce-Kuen Shieh[1]

[1]Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan
[2]Miin Wu School of Computing, National Cheng Kung University, Tainan
Email: q38021026@mail.ncku.edu.tw, csyang@ee.ncku.edu.tw, shieh@ee.ncku.edu.tw

How to cite this paper: Chiu, C.-C., Yang, C.-S. and Shieh, C.-K. (2023) A Method to Improve the Accuracy of Personal Information Detection. *Journal of Computer and Communications*, **11**, 131-141.
https://doi.org/10.4236/jcc.2023.116010

Received: May 16, 2023
Accepted: June 27, 2023
Published: June 30, 2023

Abstract

It is necessary to confirm the personal data factors and the rules of verification before conducting personal data detection. So that the detection method can be written in the subsequent implementation of the automatic detection tool. This paper will conduct experiments on common personal data factor rules, including domestic personal identity numbers and credit card numbers with checksums. We use ChatGPT to test the accuracy of identifying personal information like ID card identification numbers or credit card numbers. And then use personal data correlation to reduce the time for personal data identification. Although the number of personal information factors found has decreased, it has had a better effect on the actual manual personal data identification. The result shows that it saves about 45% of the calculation time, and the execution efficiency of the accuracy is also improved with the original method by about 22%, which is about 2.2 times higher than the general method. Therefore, the method proposed in this paper can accurately and effectively find out the leftover personal information in the enterprise.

Keywords

Data Leakage, Privacy, Personal Information Factors

1. Introduction

1.1. Disclosures of Personal Data

The authors have observed that a lot of personal data breach incidents have occurred in recent years in governmental departments, enterprises, and sole proprietorships [1]. What will happen if the names, mobile phones, addresses, photos, credit card numbers, medical information, and other personal information

of ordinary people are accidentally leaked on the Internet [2]? If these data fall into the hands of malicious persons, they can be used to carry out criminal activities, such as tracking the victim's work habits, children's schooling status, friendship behavior, bank and stock market transactions, and so on, and then carrying out counterfeiting, tracking, and threatening victims with equal payments [3].

According to the latest survey data, in January 2022, the number of online frauds reached a peak of nearly 19,000 in the past six months. With the launch of Netflix, Disney+, and others, many phishing scams involving fake streaming audio-visual platforms have appeared on the Internet [4]. By falsely claiming that there is a problem with the account that needs urgent remediation via messages such as there is a problem with payment information, Account has been locked, someone has hacked into your account, your account will be deleted, your account has been deleted, new login, and so on. The public is induced to enter phishing websites that resemble the official streaming audio-visual platform, allowing thieves to steal the account number and credit card information associated with the account [5].

Congratulations on getting a new mobile phone! What will happen if you press like and add an ID card identification number? There are other messages that unsuspecting users may click on, causing their personal information to be leaked unknowingly. When a user follows the instructions on the fraudulent fan page, their communication software's account (which may include name, mobile phone number, email, gender, company, job title, location, and graduate school) [6] may be leaked together, and then this information may be used as a profit-making tool and sold to other businesses. Consequently, the communication software receives unsolicited commercial advertisements.

1.2. Impact of Personal Data Leakage

Another common fraudulent method involves telephone fraud [7]. Based on the information leaked from communication applications on mobile [8], the fraudsters will investigate an individual's net worth, age, graduate school, circle of friends, and other details to determine whether it is worth defrauding them. Then, the user might receive an unfamiliar call from someone claiming to have legitimate business with them. For a parent of a teenager, the fraudster may say, I am your child's school teacher. If the user is a salesperson, they might be deceived by being told, I am interested in purchasing products from you. As criminals become increasingly adept at tailoring their scams to individuals by using their personal data, the best way to protect oneself is to avoid disclosing too much private information [9].

In banking [10], property insurance, life insurance, investment credit, securities, and other industries in the financial industry, the loss of personal data leakage is often more serious than system abnormality. If many customers of a single bank suffer credit card losses, other bank customers may think that their bank information has been stolen from the bank itself [11]. Consequently, to

avoid their credit card information being stolen, they will suspend the bank's credit card [12] and transfer their money to other banks that pay more attention to information security. Thus, the bank can lose the trust and business of existing customers and enterprises when many of its clients become victims of online fraud.

2. Research of Personal Data

2.1. Personal Data Legacy Case

The first step in protecting personal data is to understand how many data files are kept in the organization. Basically, the source of personal data can be divided into two categories: the personal data of personnel within the organization and the system on the external service. User profile. These two types of personal information may appear on the organization's information system or computers. The author has studied hundreds of cases of personal information import and testing services in financial, life insurance, government, and medical institutions, and the complete list is in the information.

The supervisor entrusts an employee with the handling of employee travel insurance matters, which include the names, ID card identification number, dates of birth, and other information of all colleagues, but after the processing is completed, the files are still stored in employee travel insurance Information.xlsx on the desktop. If this information is accidentally leaked, all colleagues' names, ID card identification number, date of birth, and other information will be leaked. Along with this basic information, fraudsters can use the mobile phone numbers of employees to trade with the commonly used credit card bank [13]. It is easy to pass the first checkpoint of the bank phone review [14].

2.2. Challenges with Manually

It is not easy to find personal data owned by the organization [15]. Paper data may be manually read, but accessing digital data without violating the personal information law by mistake is a major challenge to be faced next. Auditors often use rules of thumb to search for keywords, such as customer information, statement, list, ID card identification number, and other keywords, but this method can only search for file names, and it is impossible to search for the rules of personal information. For example, we found a file with the words ID card identification number, which may be a bank's credit card signing template, but the content of the file does not contain real personal information and does not violate personal information. Therefore, simply searching for keywords is not enough to prevent personal data leakage.

Many organizations manually fill out the report of personal information. However, even if the paper inventory is manually filled out, it is not guaranteed that there are no personal information files. Employees may still maintain some of this information on their systems. However, most companies simply fill in the option of none (which means there is zero personal information) on the inven-

tory report to avoid explaining and tracking the follow-up process to identify individual information with auditors.

Personal data inventory can be manually self-reported or automatically counted using certain tools. If manual inventory is easy, each colleague writes his own report freely without personal data files, and through the auxiliary detection of tools, it is easier to find some personal data files that may have been accidentally leaked. However, when the number of files that need to be manually counted exceeds 50 or the file list or the number of folders exceeds 15 or more, employees may inadvertently fill in incorrect information.

3. Discussion on ChatGPT's Personal Data Detection

3.1. About ChatGPT

ChatGPT [16] is an advanced artificial intelligence technology launched by the American artificial intelligence research laboratory OpenAI at the end of November 2022. ChatGPT learns and internalizes through the training of a large amount of data, so it has a huge amount of knowledge and the same human-like language ability. However, occasionally the answer may be inaccurate, but it is indeed a great technological innovation to be able to get a complete text reply through such a simple input command. A month after launch, more than 1 million people used it, and the number of users exceeded 100 million in January 2023. ChatGPT is currently able to handle some tasks because ChatGPT has good language skills, such as replying to letters, handling overcomes, translating, writing lyrics, sorting out text points, correcting grammatical errors, even writing stories, and writing website code. However, it should be noted that ChatGPT was trained with data before 2021. In general, if you ask ChatGPT about things after 2022, it will not know anything.

The author also conducted related research on ChatGPT's personal data detection. I wish that ChatGPT could correctly detect the characteristics of personal data. Personal data may exist in physical paper or electronic files. This chapter focuses on the text detection of electronic documents and does not consider the use of ChatGPT in images [17], sounds [18], or videos.

3.2. Factor of Personal Information

There are many types of personal information. Therefore, personal information is easy to understand literally, but because of the complexity of society, some personal data is not directly named; it is disclosed to identify a specific person, which infringes on personal privacy. Therefore, when the law is revised, personal information such as a nickname or artist name that can directly or indirectly identify the personal data will be required. Personal information can also be used to indirectly identify the individual's information, such as the address of the floor of the resident, the job title of the CEO of a company, or medical information such as the ward of the patient. This information may allow interested parties to indirectly find relevant information.

After analysis and research, there are three types, including identity numbers, credit card numbers, and email addresses, that contain the verified number of rules. Regional rules include phone, mobile, address, and other personal data factors. The results of this experiment are to ask the ChatGPT whether the samples are accurate or not. Therefore, after reviewing the above personal information rules, this study only uses identity numbers and credit card numbers to provide ChatGPT for detection.

3.3. Detection Results of ChatGPT

We generated 20 real ID card identification numbers and credit card numbers for testing data. The identity number adopted the identification rules of the country where the author belongs, which included the first character codes of different regions. 20 credit card numbers were used, including the types of VISA, Mastercard, JCB, and American Express cards. These 20 samples are the testing data for the ChatGPT. The scoring board for these three questions is as follows: **Table 1**.

First of all, let's ask the ChatGPT about the description of identity number rules, which contains verification codes, but the presentation of the results is not precise enough. The verification of 20 real data points is also completely unrecognizable.

Looking at the above testing results, we found that ChatGPT can answer any question, but if the question includes the verification of the check code, it is easy to cause the misjudgment or non-answer of ChatGPT, as shown in **Table 2**. The author also asked ChatGPT about the tax number of enterprises, the number of vehicle registration plates, and the motherboard numbers of computers. The result shows that the description of each question is barely correct. It may be incomplete or incorrect information on the internet that leads to the mistakes in the description. So that the verified code of rules cannot be well known by the CHATGP and the real samples can't be analyzed correctly.

4. Experimentation and Discussion

4.1. Improved the Detection Accuracy

If there are still a lot of mistakes after using the methods described in the previous

Table 1. Scoring board for the three questions.

Question 1	Question 2	Question 3	Sum of Points
About the verified numbers	About the rules for verifying numbers	20 samples to verified	Sum of points by Question 1 - 3
0 - 3 points	0 - 3 points	0 - 4 points	0 - 10 points
0: Can't answer 1: Totally mistakes 2: Few mistakes 3: Exactly correct	0: Can't answer 1: Totally mistakes 2: Few mistakes 3: Exactly correct	Each sample is worth 0.5 points if it is correct	0 - 3: Poor 4 - 6: Good 7 - 10: Excellent

Table 2. Scoring of the five items.

Question Items	Question 1 (Description)	Question 2 (Rules of Verify Code)	Question 3 (20 Samples to be verified)	Sum of Points
Identity number	2	3	0	5
Credit card number	1	1	0	2
Tax number of enterprises	1	0	0	1
Number of Vehicles registration plate	1	0	0	1
Motherboard number of computers	1	0	0	1

section to apply the detecting, the accuracy can be improved by using the methods described in this section.

- Method 1: Individual factor intersection

File A, for instance, just has 12 names and 8 addresses, with no further identifying information. File A will be designated as the critical risk in rule 1 because there are more than ten names of personal information components in the file, as mentioned in the risk rule table of ideas in the preceding chapter. Even if there are less than ten addresses, the first risk condition is still satisfied. This type of situation typically occurs in lengthy PDF documents. The reason for this is because it is quite simple for mixed texts to lead to incorrect conclusions.

However, if we enhance the judging method of the personal information factor and switch to intersection (as shown in **Figure 1**), the 12 names and eight addresses will only meet the high risk of risk rule 2, meaning that there is a very high probability of name misjudgment in the file and the real data may only have eight items of personal information. Accordingly, the system's misjudgment may be greatly decreased by the confluence of data factors.

Passport and residency certificate numbers were among the criteria used in the follow-up. The passport number was especially problematic because there was no clear check code or set of rules to follow when evaluating it. The number on a person's residence certificate, on the other hand, functions similarly to that on an identity card in that it includes a check code to validate the most recent digital formulae, making it less prone to error in judgment. It's important to treat five high-risk files on a computer differently than it would treat five thousand files containing sensitive personal information. Mistakes can be eliminated manually one by one in a computer with five personal data files, but a computer with five thousand personal data files cannot. Thus, correct regulations will decrease file-related mistakes.

- Method 2: The personal information correlation factor

The writers helped out with various agencies and their testing of private information; they discovered that the way certain pieces of data are presented depends on the specific circumstances. One such linkage of personal information is

the appearance of 20 characters before or after the name in the ID card identifying number.

Rules such as "mobile must appear in the first 20 characters before or after the name" and "address must appear in the first 20 characters before or after the mobile" are examples of those that apply. All of these guidelines for detecting and bettering the accuracy of personal in-formation detection are based on the correlation between the preceding and subsequent characters. **Figure 2** is an example of a diagram for thinking about the context of your personal data before and after a correlation keyword.

Some auditors and government organizations agree with this concept because some pieces of personal data are only significant when placed in the context of a larger whole. This is the reason when we fill out an application form, we often provide our personal factors in the following order: name, ID card identification number, phone, mobile, e-mail address, address, credit card number, etc. Therefore, the frequency with which certain pieces of private data occur depends on the context in which they are used. That's why it's important to take into account the contextual connection of personal information variables when making risk-level assessments; it'll boost the precision of your PI detection.

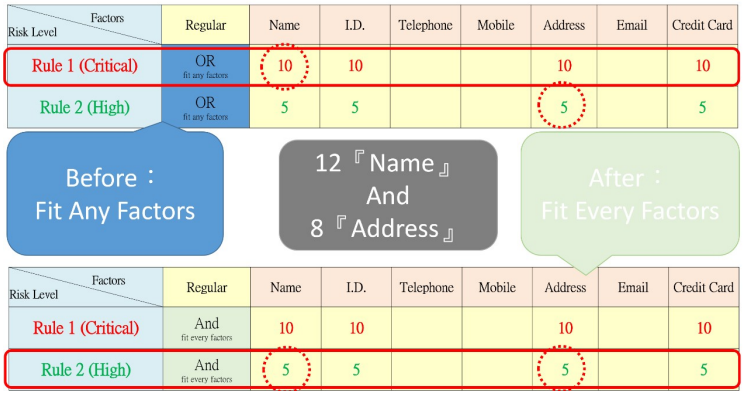

Figure 1. Table of risk rules enhanced by intersection.

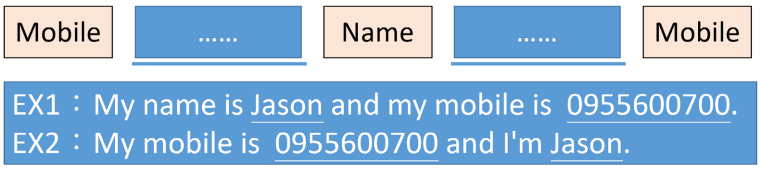

Figure 2. The personal information correlation factor.

4.2. Experimental Results

The experiment used 500 sensitive documents by using the ID card identification number as the main analysis index and the name and address as the auxiliary personal information factors of the ID card identification number. Methods 1 and 2 in the previous section for comparative analysis a complete, workable scan is complete when at least one ID card identification number is

detected. Method 1: When the identity card, name, and address must appear at the same time within one document, the result can only be considered personal information. Method 2: When an ID card identification number appears, at the same time, the first and last 100 characters must also include the name or address, and the result is considered personal information.

4.3. Preliminary Judgment of Accuracy

The statistical results are shown in **Table 3** below. It took about 66 minutes for the full scan to find 84 ID card identification numbers in 500 sensitive documents; method 1 took 59 minutes to find 49 ID card identification numbers; and method 2 took 24 minutes to find 56 ID card identification numbers.

The accuracy is 1 based on the number 84. Method 1 and Method 2 were 0.58 (49/84) and 0.67 (56/84) respectively. The weight value calculated by the previous research method is C squared/D, and the weight value of method 2 is 0.22 higher than the original method.

Since there are a total of 500 files in the samples, the ID card identification number rule has a verify code and has passed the check code formula; therefore, it is first determined that all ID card identification numbers have been found. The result shows only 84 entries of all ID card identification numbers appearing in 500 files. The results of this experiment are helpful for the rapid prediction of the future. For example, how long does it take for ChatGPT to detect 10 ID card identification numbers? According to the above experimental results, the original method takes 7.85 (66/84 * 10) minutes, method 1 takes 12.04 (59/49 * 10) minutes, and method 2 only takes 4.28 (24/56 * 10) minutes, which can save about 45% (1 − (4.28/7.85)) of the time efficiency compared with the original method. In other words, with the same accuracy, Method 2 can provide correct personal information in only 55% (4.28/7.85) of the detection time. This undoubtedly makes a significant contribution to the large amount of data gathered for research.

Table 3. Table of experimental result.

	Number of I. D. (A)	Executing Time (min) (B)	Precision (C) = (A)/84	Percentage of Time (D) = (B)/66	Efficiency of Weight (C) * (C)/(D)
Full Scan	84	66	1.00	1.00	1.00
Method 1	49	59	0.58	0.89	0.38
Method 2	56	24	0.67	0.36	1.22

5. Conclusions

Recently, in the technology industry, worker job-hopping occurred where workers carried company knowledge and know-how to a new company; it happens all the time. For important internal business secrets, it is necessary to discuss them in detail from the perspectives of personnel, time, land, and objects. It is impor-

tant to distinguish who can know the company's business secrets and confidential documents, how detailed the information can be, when it is a suitable time to get the data, and where the information will be stored and monitored. Further, what information cannot be disclosed to the external the lifeblood of an enterprise is these important personal data files and confidential documents. Even if 100% prevention of file leakage cannot be achieved, at least reduce the probability of data leakage as much as possible. Relevant control procedures must be done within the enterprise, including the retention, place on files, backup, file transfer in and out of personal data files and confidential files, etc., to protect important assets of the company and avoid huge losses after data leakage. Nowadays, personal data leakage incidents are emerging one after another, which not only causes panic among the public but also proves that both individuals and companies should better understand and implement the mechanisms of personal data protection.

In view of this, in order to prevent personal data leaks, it is necessary to strengthen the computer's personal data protection mechanism. In order to prevent these issues, the company can regularly conduct personal data protection education and training for employees [19]. Thus, personal data scanning on computers, websites, and databases also helps. Smart personal data encryption will greatly reduce the risk of personal data leakage due to employee negligence or system program loopholes. Therefore, prone to negligence and incorrectness in paper filling, more and more enterprises and organizations use automated personal data inventory tools to carry out the final computer personal data inventory and supplement manual inventory.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chiu, C.-C., Tsai, P.-W. and Yang, C.-S. (2021) PIDS: An Essential Personal Information Detection System for Small Business Enterprise. 2021 *International Conference on Electrical, Computer, Communications and Mechatronics Engineering* (*ICECCME*), Mauritius, 7-8 October 2021, 1-6. https://doi.org/10.1109/ICECCME52200.2021.9590950

[2] Lee, H.-J. Lee, K. and Won, D. (2011) Protection Profile of Personal Information Security System: Designing a Secure Personal Information Security System. 2011 *IEEE* 10*th International Conference on Trust*, Changsha, China.

[3] Naomi, J.F., Vasanthageethan, A., Roshini, G. and Kumar, J.S. (2021) Data Privacy Preserving Recommendations for Social Media. 2021 7*th International Conference on Advanced Computing and Communication Systems* (*ICACCS*), Coimbatore, 2021, 1229-1232. https://doi.org/10.1109/ICACCS51430.2021.9441870

[4] Vishnevskaya, J.A., Kovalenko, M.V. and Skvortsova, M. (2021) Analysis of Methods of Illegal Personal Data Distribution of Messengers, Social Networks and Search Engines Users. *Russian Young Researchers in Electrical and Electronic En-*

gineering (*ElConRus*), Moscow, 26-29 January 2021, 2404-2409.

[5] Wang, S., Shi, L., Hu, Q., Zhang, J., Cheng, X. and Yu, J. (2021) Privacy-Aware Data Trading. *IEEE Transactions on Information Forensics and Security*, **16**, 3916-3927. https://doi.org/10.1109/TIFS.2021.3099699

[6] Chia, S.Y., Xu, X., Ding, M., Smith, D., Paik, H.-Y. and Zhu, L. (2023) A Selection Model of Privacy Patterns. 2023 *IEEE* 20*th International Conference on Software Architecture* (*ICSA*), L'Aquila, 13-17 March 2023, 1-11. https://doi.org/10.1109/ICSA56044.2023.00009

[7] Zhang, D., Yao, L., Chen, K., Yang, Z., Gao, X. and Liu, Y. (2021) Preventing Sensitive Information Leakage from Mobile Sensor Signals via Integrative Transformation. *IEEE Transactions on Mobile Computing*, **21**, 4517-4528. https://doi.org/10.1109/TMC.2021.3078086

[8] Liu, J., *et al.* (2023) CPAHP: Conditional Privacy-Preserving Authentication Scheme with Hierarchical Pseudonym for 5G-Enabled IoV. *IEEE Transactions on Vehicular Technology*.

[9] Zhao, Y., Si, N., Sun, Y., Gao, X., Tong, H. and Yuan, G. (2022) An Automatically Privacy Protection Solution for Implementing the Right to Be Forgotten in Embedded System. *IEEE Access*, **10**, 35146-35161. https://doi.org/10.1109/ACCESS.2022.3162238

[10] Hema, P.N., Rathika, P.D. and Pushparaj, A. (2023) Privacy Preservation Using Federated Learning for Credit Card Transactions. 2023 *International Conference on Intelligent Systems for Communication, IoT and Security* (*ICISCoIS*), Coimbatore, India.

[11] Vithana, S. and Ulukus, S. (2023) Model Segmentation for Storage Efficient Private Federated Learning with Top *r* Sparsification. 2023 57*th Annual Conference on Information Sciences and Systems* (*CISS*), Baltimore, 22-24 March 2023, 1-6. https://doi.org/10.1109/CISS56502.2023.10089698

[12] Siva Alagesh, S., Bhavna, B., Swati, A., Aruna, J.R., Samrat, R. and Amit, K. (2022) Privacy Preservation Using Block chain for Credit Card Data. 2022 2*nd International Conference on Innovative Practices in Technology and Management* (*ICIPTM*), Gautam Buddha Nagar, 23-25 February 2022, 725-730. https://doi.org/10.1109/ICIPTM54933.2022.9754015

[13] Zhang, P., Zhang, N., Moini, A., Lou, W. and Hou, Y.T. (2020) Privacy Scope: Automatic Analysis of Private Data Leakage in TEE-Protected Applications. 2020 *IEEE* 40*th International Conference on Distributed Computing Systems* (*ICDCS*), Singapore, 2020, 34-44.

[14] Wang, X., Guo, Y., Zhao Y. and Yu, H. (2022) The New Progress and Methods of Privacy Protection on Medical and Health Big Data. 2022 14*th International Conference on Software, Knowledge, Information Management and Applications* (*SKIMA*), Phnom Penh, 2-4 December 2022, 73-78. https://doi.org/10.1109/SKIMA57145.2022.10029494

[15] Huang, Y., Li, Y.J. and Cai, Z. (2023) Security and Privacy in Metaverse: A Comprehensive Survey. *Big Data Mining and Analytics*, **6**, 234-247. https://doi.org/10.26599/BDMA.2022.9020047

[16] Zhang, J., *et al.* (2023) HiVeGPT: Human-Machine-Augmented Intelligent Vehicles with Generative Pre-Trained Transformer. *IEEE Transactions on Intelligent Vehicles*, **8**, 2027-2033. https://doi.org/10.1109/TIV.2023.3256982

[17] Chen, H. and Yamamoto, Y. (2021) Task-based Assessment to Evaluate Instagram Users' Capabilities for Personal Information Leakage Prevention. 2021 10*th Inter-*

national Congress on Advanced Applied Informatics, Niigata, 11-16 July 2021, 29-34. https://doi.org/10.1109/IIAI-AAI53430.2021.00005

[18] Zhang, R., Yan, Z., Wang, X. and Deng, R.H. (2023) VOLERE: Leakage Resilient User Authentication Based on Personal Voice Challenges. I*EEE Transactions on Dependable and Secure Computing*, **20**, 1002-1016. https://doi.org/10.1109/TDSC.2022.3147504

[19] Yu, H., Sun, H. and Xu, D., (2021) Research on the Dilemma and Countermeasures of Employees' Right to Privacy Based on Big Data. 2021 2*nd International Conference on Big Data and Informatization Education* (*ICBDIE*), Hangzhou, 2-4 April 2021, 21-28. https://doi.org/10.1109/ICBDIE52740.2021.00014