Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 5303329, 10 pages
https://doi.org/10.1155/2023/5303329

Research Article

Using an Efficient Detection Method to Prevent Personal Data Leakage for Web-Based Smart City Platforms

Chih-Chieh Chiu [1] Pang-Wei Tsai [2] and Chu-Sing Yang [3]

[1] Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
[2] Department of Information Management, National Central University, Taoyuan, Taiwan
[3] Miin Wu School of Computing, National Cheng Kung University, Tainan, Taiwan

Correspondence should be addressed to Pang-Wei Tsai; pwtsai@ncu.edu.tw

Received 30 December 2022; Revised 6 April 2023; Accepted 19 April 2023; Published 10 July 2023

Academic Editor: Tien-Wen Sung

Many Internet of Things and information exchange technologies bring convenience, cost-efficiency, and sustainability to smart city solutions. These changes have improved our day-to-day quality of life, with impacts on: (a) lifestyle (e.g., automation and robotic reaction), (b) infrastructure (efficient energy consumption), and (c) data-driven management (data sensing, collection, and investigation). It is common to integrate Web-based interfaces and such solutions for developing platforms. When software and hardware components store, retrieve, and transfer such information, people may suffer from personal data leakage. This paper introduces a privacy information detection method, using a data weighting mechanism to save time and cost in finding personal information leaks over Web services. According to an initial evaluation, the proposed method can reduce time by 62.19% when processing 8,000 crawled files, and roll-back verification shows that it maintains 90.08% accuracy for finding marked content.

1. Introduction

Since technologies related to the Internet of Things (IoT) have been increasingly used in smart cities and homes, applications related to agriculture and fisheries, water quality monitoring, power monitoring, road monitoring, and other applications have expanded exponentially [1]. However, the IoT primarily uses HTTPS or MQTTS for information transmission, and the transmission process meets information security requirements [2], most of which import HTTPS data encryption for basic data protection. Hence it is an emerging requirement to enable protective measures on data storage and smart information solutions.

In forensic cases, it is common to find security breaches caused by inadequate settings [3]. In these security accidents [4], the data collectors utilized to gather sensing data form a key point. Because they only collect and aggregate data via an IoT gateway [5], they do not understand the content of the transmitted data, making it impossible to determine if it is sensitive information. Without strict checks during data processing, such data may be exposed, especially on Web-based smart city service platforms.

End users should know whether their personal information has been disclosed. Platform administrators may have to pay the cost, as they are responsible regarding accidental leaks. For example, in 2018, British airways [6] compromised information of 380,000 customers, for which the company was fined more than 180 million euros. Based on a survey [7] on data privacy, about 70% of consumers feel concerned and experience consequences if their private information is leaked. Hence data security has become a complex challenge for smart cities [8]. This research reduces data processing time while maintaining investigative accuracy on smart city service websites, and introduces a method for detecting privacy information leaks over public Web services.

Several situations may cause personal information leaks, such as uploading sensitive data or maintaining registration data in an inappropriate location. Storage with low-end security protection can be compromised, and stored plaintext data can be easily exposed. Many countries have enforced privacy laws [9] at various levels of government to force stakeholders to preserve sensitive data, and industry and business associations, and social groups have issued regulations and guidelines. Nevertheless, with data mining [10] and machine learning [11] mechanisms, it is helpful to accelerate the process of determining whether collected data involve personal information. Using a scanner or crawler [12] to secure data privacy has become a common practice.

The remainder of this paper is organized as follows. Section 2 introduces background notions and related work, and in brief discusses the measurement of personal data detection. Section 3 explains the system design, platform development and workflow of the investigation. Sections 4 and 5 present the data comparison process, propose methods that improve the time and cost, and demonstrate the verification and evaluation results of experiments. Section 6 provides our conclusions and discusses future work.

2. Background Knowledge and Related Work

2.1. Personal Information Data Leakage. As the global network continues to grow, the internet accelerates the exchange of information. However, personal data might be accidentally disclosed on networks [13]. Researchers have discussed internet leaks and countermeasures. Tbahriti et al. [14] illustrated the requirement of making a privacy policy to protect Web services, and proposed a dynamic model with privacy at the levels of data preservation and processing. Tiwary et al. [15] discussed the challenge of effective privacy protection in data service composition. Depending on the route for transmission and data access, authentication and encryption procedures must be activated to prevent leaks [16]. If there is no way to guarantee data security, allowing data queries over autonomous data services by masking sensitive content is an option [17, 18]. Due to a critical leakage event, many countries have launched data protection laws requiring government departments and enterprises to define strict policies for processing data with personal information.

2.2. Case Studies. Case 1 involves an actual government unit during a research period, which uses requires a nondisclosure agreement (NDA) to preclude leaks of information. The data collector compiles a daily report stored in an unencrypted CSV file, which allows system administrators to download statistical reports. When the data collector's information security was checked, the database and file contained names and ID information (e.g., country identification code used in voting, medical treatment, telecommunications, and household registration), with each file containing on average 500–3,000 pieces of personal data. Manual verification revealed that the data collection system primarily collected clock-in and clock-out information for each employ in a building. However, because the government department lacked an employ number, ID data and names were used to identify and store data, and the data collector transmitted encrypted data through nearly 20 IoT devices in the building, which is a case of IoT data leakage. In this case, the mechanism was able to stop the publication of personal data on the internet.

Case 2: Assume there is a website managed by a law-enforcement agency that preserves criminal personal information such as criminal behavior, age, masked name, masked ID card, and masked address. Hackers can easily modify unencrypted and unprotected information such as time and area, and the data collector may be attacked from two aspects. The first is a direct attack, in which the hacker only modifies the records of accomplices to avoid the detection of substantial amounts of abnormal data. The second is to forge the transmission packet of the electronic shackles and regularly pass a fake packet, so as to prevent criminal accomplices from being restrained by electronic shackles. In this case, the electronic shackles are IoT devices, and the possibility of data collectors being attacked is high. Therefore, in addition to data encryption and protection, data security protection, and personal data de-dentification should be performed as much as possible to avoid destroying the reputation of crime-proof smart cities.

Table 1: Keywords and counting numbers of websites.

Item	Remarks
Website *A*	{*k*101(c101), *k*102(c102)–*k*115(c115)}
Website *B*	{*k*201(c201), *k*202(c202)–*k*228(c228)}
Website *C*	{*k*301(c301), *k*302(c302)–*k*352(c352)}
*k*115	Keyword of number 115 on Website *A*
*c*301	Counting number on *k*301 keyword

2.3. Discussion. The authors have observed that most websites are publicly accessible, making it difficult to protect sensitive personal data, especially for websites that frequently update content. Comparing the scanning task of Web content with a local file, the former increases much of the work of searching and determining valuable information for investigations. Many personal data breach incidents have occurred in recent years, drawing attention to governmental departments [19], enterprises, and sole proprietorships. Hence research and commercial benefits exist for the development of application services to help stakeholders practice information security and protect personal information.

2.4. Challenges. This research focuses on reducing the data required for analysis during preprocessing, because excluding useless datasets can reduce the time to finish pattern-comparison work. Support vector machine (SVM) and neural network (NN) algorithms are expected to initially classify collected data from the target website. However, a manual examination of training data collected from the internet found that datasets involving personal information on websites are diverse, and most trained samples cannot be used as a factor for subsequent prediction.

For instance, in Table 1, k indicates the keyword of the privacy data in the target website, and c is the total number of times the keyword appears on the website. After calculation, Website A has 115 keywords, $k101$–$k115$, with corresponding numbers of occurrences of $c101$–$c115$. A total of 95 keywords between $k101$–$k115$, $k201$–$k228$, and $k301$–$k352$ were found. The result reveals that these keywords are not duplicated. Therefore, no similarity of personal information

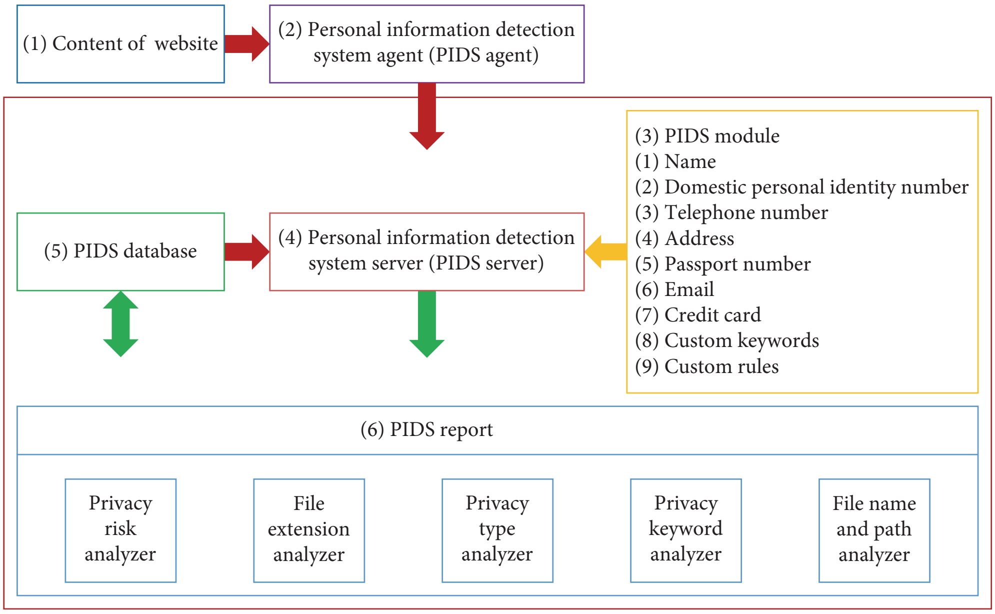

Figure 1: System architecture.

keywords exists between the websites. In terms of analysis, classification algorithms such as K-nearest neighbors [20] or SVM [21] cannot be used to identify the real aim of personal information for a website. The authors have tried other methods to overcome such problems. The proposed mechanisms are further discussed in Sections 3 and 4.

3. System Design

To guard against personal data leakage, the authors previously built a prototyping system, personal information detection system (PIDS) [22], which can use word analysis and pattern-based detection to determine whether scanned data violate detection rules. However, this system only supports the scanning of static data stored on local machines. As global networks continuously grow, the World Wide Web has become a vulnerable place that suffers from personal data leakage accidents. Hence we introduce a Web-scan system based on the original PIDS design, which aims to explore and examine imported Web content types, characteristics, and the chances that content contains sensitive data. This research seeks to reduce the time cost of data processing while maintaining good investigation accuracy on the websites of smart city services.

3.1. System Architecture. We introduce the system architecture. Figure 1 illustrates its design and module functionality. During operation, the input data comprise the website content, and the output data are the detected personal information records. This design enables the system to support distributed parallel processing. For security, confidential parameters are encrypted by AES256, including the password of the database and the administrator's login account.

3.2. Personal Information Detection System Web Agent. Several customization settings are available on the PIDS Web Agent (PWA) to support various requirements and work progress. The PWA employs a Document Content Parsing Module (DCPM) on the file by setting the file name extension type to be scanned. The PIDS server side tells the PWA to investigate specific types of personal information, such as email addresses, phone numbers, addresses, mobile numbers, credit card numbers, identification numbers, passport numbers, and names. After the scan, the PWA dashboard displays the investigation results and risk levels.

3.3. Personal Information Detection System Server. Based on past work, the authors investigated particular categories of privacy data, including names, IDs, phone numbers, mobile numbers, addresses, passport numbers, email addresses, credit card numbers, keyword customizations, and personal data analyses, and built corresponding customized rules, using various principles and methods to identify information. For instance, the identification criteria of email addresses are simple. The "@" symbol must be preceded by at least one character and followed by at least one character, including a period. The following explanations detail the comparison policy and rules for determining personal information in collected data.

(1) Name: The family name plus first name is used for a string comparison. The detect rule set involves 2,251 common English family names and 2,368 given (first) names. Capital letters, lowercase letters, and abbreviations in English names, such as "Jr.," are also detected.

(2) Domestic personal identity number: The ID number has specific formula verification rules; hence it is easy to verify whether an ID is a Taiwan ID number, for example, the checksum is used to verify whether the number is a legal ID number.

(3) Telephone number: The telephone number uses a list of reasonable area codes (022–089) at the beginning

of detection, and identifies a total number of 9–10 digits.

(4) Address: Addresses includes 43 county and city names. The condition includes a road and address number (e.g., rule of county or town plus road and number) to identify the address feature.

(5) Passport number: The passport number is a combination of nine digits starting with "3." The rate of misjudgments of passport number rules is extremely high because it has no checksum. Consequently, it is challenging to identify whether personal data are real if the file contains various passport numbers.

(6) Email: The identification criteria of email addresses are that the text preceding the "@" symbol must include at least one character and be followed by one or more characters that include a period.

(7) Credit card: Credit card numbers include 16-digit numbers for Visa, Mastercard, and JCB, and 15-digit numbers for American Express, each with their own checksum rules. The check code is a combination of numbers. Hence the misjudgment rate of credit card numbers is extremely low. This is indispensable in the PIDS.

(8) Custom keywords: Clients have inspection category keywords for particular industries, which can be added individually (e.g., account information for the financial industry, consumer or staff number for the insurance industry, or license plate number for the property insurance industry).

(9) Custom rules: The system can add custom rules using plugins if rules require additional checking (e.g., rules for bank remittance account numbers, material number rules for manufacturing, or student ID rules for universities). These are additional criteria for personal data identification to increase the accuracy rate.

The DCPM of the Web agent gathers data from common file types (e.g., doc, docx, ppt, pptx, xls, xlsx, odp, ods, odt, pdf, rar, zip, 7z, tar, txt, csv, htm, and html), and converts the file format, such as Microsoft or Open Office, into a plain text file for analysis. General text files (e.g., txt, csv, htm, and html) are also supported. The DCPM must still put considerable effort into parsing files in various character encoding schemes, such as UTF-8, ANSI, ISO, and Big5.

Examining compressed files, such as with rar, zip, 7z, and tar extensions, typically requires decompression to obtain the original content. However, the compressed file has a specific file title, which is convenient for parsing and reading. If a compressed file can only be opened with a password, the PIDS detects the file-type behavior as the file is encrypted, and cannot directly decompress it to analyze whether the content has personal data.

Another case is the PDF file, such as policy content or customer statements of account. PDF files are sent through email for customers to view. There have been recent cases in which financial professionals have manually tampered with the content of PDF statements to deceive customers and steal account assets. If the content of a PDF file is encrypted or limits the function of printing or copying text, the DCPM presents an error message.

3.4. Personal Information Detection System Database. The PIDS database used MariaDB, and 46 data tables were initially created. In terms of rule configuration, the database preserves personal information criteria, allowed or disallowed keywords, allowed or disallowed file paths, file size limits, time limits, and character number limits. The investigation results are stored here, including scanning information, date/time, risk level, risk number of files, and risk percentage relevant information. The database supports the limiting of the amount of data and number of characters to avoid excessive data being returned to the database and to prevent transmission failure.

3.5. Personal Information Detection System Report. The PIDS report saves the system settings, including personal rules and report settings. The system setting functions are to set the initial screen, apply rules, auto-update settings, and support external command modes. Personal data settings include detection of personal data conditions, customizing rules, exclusion paths, and risk levels. The report function includes the configuration and automatic schedule. Other configurations include startup restrictions, appointment schedules, and acceleration and encryption modes. Once the system scan is complete, the detection results for all people and the real-time scanning process are available on the PIDS report. This system is used by more than 30 banks and life insurance companies.

3.6. Web Content Crawling. In the first step, the system downloads all website files to search the personal information on a website [23]. It contains a list of websites to store multiple candidates, which is convenient for subsequent automatic downloads. The download time depends on the number of pages. The system arranges 5–10 execution sequences simultaneously before personal information analysis.

Several predefined parameters are used to avoid downloading failure. For example, "robots = off" means to turn off the detection of robots, "wait = 2" means to wait for 2 s, "random-wait" means the download seconds are random, "user-agent" means to provide user-side browser program parameters, and "no-check-certificate" means that the download can be done regardless of whether there is an SSL certificate.

In the real world, when avoiding the personal data leakage of a smart city website, we usually download the entire website for a data check. A website containing 2,000–5,000 pages takes 10 hr to download, and a more extensive website, with 8,000–20,000 pages, takes longer. The collected Web content becomes the training and testing data in research.

4. Model Development

4.1. Training Data. For verification and evaluation, the authors collected 50 public Web-based platforms from education institutions as the major for crawling tasks. It took about 10 days to download the entire content. Thus, there

Table 2: Statistical information for 50 websites.

Total files for one website	No. of websites (A)	Total files of (A) (B)	No. of files with personal data (C)	No. of repeated personal data (D)	Average files per website (B/A)	No. of repeated personal data per website (D/A)	No. of repeated personal data per file (D/C)
0–500	29	3,639	2,784	42,279	125	1,458	15
501–2,000	12	12,439	7,652	209,948	1,037	17,496	27
2,001–5,000	4	13,285	8,453	405,145	3,321	101,286	48
5,001–10,000	4	25,074	20,880	343,699	6,269	85,925	16
10,001–50,000	1	23,457	22,148	269,831	23,457	269,831	12
Total	50	77,894	61,917	1,270,902			

Table 3: Statistical information for 50 websites.

File name extension	Word	Excel	PPT	PDF	HTML	Total
Number of files	3,381	513	87	9,649	64,264	77,894
Percentage of total	4.34%	0.66%	0.11%	12.39%	82.50%	100.00%

Table 4: Classification of file name extensions for 50 websites.

File name extension	No. of total files (A)	No. of files with personal data (B)	Percentage of files with personal data ($C = B/A$)	No. of repeated personal data (D)	No. of repeated personal data per file ($E = D/B$)	No. of nonrepeated personal data (F)	Average nonrepeated personal data per file ($G = F/B$)
Word	3,381	2,546	75.30%	25,146	9.88	8,952	3.52
Excel	513	362	70.57%	12,852	35.50	3,772	10.42
PPT	87	66	75.86%	318	4.82	258	3.91
PDF	9,649	7,679	79.58%	201,083	26.19	37,488	4.88
HTML	64,264	51,264	79.77%	1,031,503	20.12	31,692	0.62
Total	77,894	61,917		1,270,902		82,162	

will be fewer realistic and duplicate personal data when performing relevant research. The testing results are expected to increase system performance and reduce computation time.

In the first step, the collected data from these 50 websites must be categorized. The Table 2 displays the statistical findings. These websites took 54.7 GB of storage space, and had 77,894 recognized file formats (an average of 1,558 pages per website). The processing equipment was a workstation with an Intel i7-12700 CPU, 32 GB DDR4 memory, and 2 TB of SSD space.

The number of personal information files is provided in Table 2. A total of 77,894 Web pages were scanned, and 1,270,902 personal records were present on 61,917 of these pages. The number of nonrepeated personal data from 1,270,902 is around 51,692 count by computer calculation, and the top three repeated data occurrences are approximately 21,254, 12,002, and 6,436, thus indicating that the number of repeated personal data is very high. The nonrepetitive data were used in the follow-up research. A file contained an average of 20.5 (1,270,902/61,917) repeated personal data items during preprocessing. It took nearly 20 days, due to the number of files. This study aims to determine how to decrease scanning times while maintaining a specified level of system accuracy.

Table 2 displays statistical information for 50 websites, preliminarily categorized according to a site's number of webpages, with categories defined as 0–500, 501–2,000, 2,001–5,000, 5,001–10,000, and more than 10,000 webpages, and an average website has 1,558 (77,894/50) webpages. Subsequent experimental samples gathered 500–2,000 webpages for analysis, to ensure their quality and validity.

4.2. Method 1: File Type Classification. We review the mentioned 50 websites, and present classification results for several file name extensions in Table 3.

According to the classification summary, the percentage of HTML files with personal information is the highest (Table 4). HTML was the most common file name extension, at 82.50% of all files, and 79.77% (51,264/64,264 records) of all HTML files (64,264) that included privacy data.

About 4.88 unique personal information instances were found in a PDF file on average, and the top three files had 4,417, 3,937, and 3,695 repeated records. Moreover, 7,679 of the 9,649 files (79.58%) had personal information in PDF format.

For the PPT format, 66 of 87 files (75.86%) had personal information. The top three files had 846, 777 and 660 records with repeated data.

Table 5: Percentage of HTML files.

Percentage of HTML files	No. of non-HTML files (X)	No. of HTML files (Y)	No. of total files ($X + Y$)	Execution time (min)	Percentage of execution time (A)	No. of nonrepeated personal data	Percentage of nonrepeated personal data (B)	Value of weight $W1$ ($B \times B/A$)
10%	1,163	758	1,921	94	27.09%	9,976	68.51%	1.73
20%	1,163	1,516	2,679	114	32.85%	10,629	72.99%	1.62
30%	1,163	2,274	3,437	156	44.96%	13,222	90.80%	1.83
40%	1,163	3,032	4,195	174	50.14%	13,437	92.27%	1.70
50%	1,163	3,790	4,953	201	57.93%	13,479	92.56%	1.48
100%	1,163	7,580	8,743	347	100.00%	14,562	100.00%	1.00

There were 513 files in Excel format, and the system found 362 files with 12,852 types of privacy data. There were 3,772 types of nonrepeated records, and each file could contain more than 10.42 types of personal information on average. The top three Excel files had 786, 649, and 481 records with personal data. The HTML format had the highest proportion of personal data content (79.77%), but after deleting duplicate personal data, the personal data content of each HTML file was reduced from 20.12 to 0.62. There were 31,692 nonrepeated personal data items, and 1,031,503 repeated personal data items from files with the HTML extension. This research focuses on how to reduce the number of HTML calculations while maintaining a certain level of accuracy.

The analysis found that the HTML file name extension contained the lowest percentage of privacy data. However, HTML extensions accounted for the highest number of files, at 82.50% of the total. Therefore, the HTML file name extension is the first target for the implementation of a new method to reduce the time to search for privacy data.

4.3. Method 2: Statistical Analysis. The experimental findings were determined using 10 out of 50 websites as training data. The expected value estimation system divides the HTML format into 10%, 20%, 30%, 40%, 50%, and other proportional principles to conduct the tests and save around 50% of the time cost. The weight W was the square of the percentage of nonrepeated personal data/percentage of execution time as:

$$W1 = B^2/A. \tag{1}$$

Table 5 indicates that about 30% of HTML files provided the best results. Although the percentage of scan time decreased to 44.96%, the detection rate was maintained at 90.80%.

Furthermore, 30% of HTML file name extensions resulted in the maximum efficiency and detection rate, allowing 84.75 (13,222 records/156 min) personal data/min to be scanned, with a detection rate of up to 90.80%. The next-best results were for 10% and 40% HTML files, where 106.12 (9,976 records/94 min) and 77.22 (13,437 records/174 min) data items could, respectively, be scanned per minute. However, only 68.51% and 92.27% of data items, respectively, could be detected. Thus, the detection rate can be maintained at 90.80% even if the utilization is 30% of HTML files. Only 10% of all HTML files can be scanned to save 73% of the processing time, but the detection rate drops to 68.51%.

4.4. Method 3: Duplicate Analysis. This method leverages a detection software toolkit [24] to search for duplicate files, with a robust algorithm to identify the file content, size, and file name. As presented in Table 6, for 90% similarity, it takes about 13 min to reduce 8,449 HTML files to 3,248. This software also compares files accurately using their sizes, byte-to-byte. If two or more files have the same size, their data are compared. It is erroneous to assume that files with the same name are duplicates. The application never considers the file name because users can rename identical files or give different files with the same name; hence the results can be inaccurate. This method can find similar files regardless of their file types. To determine duplicates, it analyzes the file data and not just file attributes, like name and size, as standard clone removers do.

As previously explained, it takes too long to calculate website similarity for 50 samples; therefore, 10 websites with 500–2,000 pages were used for examination. The similarity of each file of these 10 websites was calculated. Website content to be searched for personal information can be decreased by filtering repeated and useless records, realizing a 50% reduction in computation time while maintaining a 70% detection rate. In the experiment, a website with 95% file similarity had only 3,927 files to be scanned from a total of 8,449 files, with a 98.26% detection rate. Even 95% file similarity takes 7 min of comparison time with 157 min of scanning time, and still saves 53.41% of the 352 min of scanning.

The formula involves multiplying the appearance ratio for each personal information type and the ratio of the personal information file, using the weighted statistical approach as detection rate times the detection rate divided by the comparison time and scan time as:

$$W2 = D^2/E. \tag{2}$$

In the best case (with a 90% threshold), as presented in Table 6, a 94.82% detection rate is achieved, and the investigation procedure only takes 40.91% of the time.

5. Evaluation

5.1. Testing Data in the Method. To evaluate the system performance, 10 websites (different from the training set) were randomly selected as testing data for the experiment. There were 39,492 files in the dataset, taking an average

Table 6: Percentage of file similarity.

File similarity percentage	Comparison time (min) (A)	Scanning time (min) (B)	No. of files	No. of nonrepeated personal data (C)	Percentage of nonrepeated personal data (A) ($D = C/14{,}562$)	Percentage of time (B) ($E = (A + B)/352$)	Value of weight $W2$ ($D \times D/E$)
99%	4	228	5,770	14,439	99.16%	65.91%	1.49
97%	5.5	185	4,551	14,363	98.63%	54.12%	1.80
95%	7	157	3,927	14,309	98.26%	46.59%	2.07
93%	9	144	3,603	14,053	96.50%	43.47%	2.14
91%	11.5	136	3,370	13,856	95.15%	41.90%	2.16
90%	13	131	3,248	13,807	94.82%	40.91%	2.20
85%	19	129	2,705	13,410	92.09%	42.05%	2.02
80%	27	125	2,332	13,067	89.73%	43.18%	1.86
75%	35	122	2,031	12,506	85.88%	44.60%	1.65
70%	44	117	1,795	11,735	80.59%	45.74%	1.42
65%	55	113	1,664	11,518	79.10%	47.73%	1.31
60%	72	104	1,575	11,006	75.58%	50.00%	1.14
55%	88	97	1,503	10,458	71.82%	52.56%	0.98
50%	104	92	1,443	10,291	70.67%	55.68%	0.90
Full-scan mode	0	352	8,449	14,562	100.00%	100.00%	1.00

12.61 GB of storage space. Using the system described in Section 4, it took about 54 hr (3,238 min, precisely) to find 58,009 unrepeated personal information records involved in these 39,492 Web files in full-scan mode. The system marked 718,324 of these records (located in 29,946 files, about 76% of the dataset), which included 58,009 nonrepeated personal data items, and 660,315 repeated.

The experimental results indicate that each file may contain approximately 24 repeated and two unique personal information records on average. Manual investigation showed that these repeated 660,315 personal information records were usually home addresses or phone numbers frequently appearing on a webpage.

Table 7 compares the experimental results of the three methods. None of the methods computes the file download times, and all calculate the comparison and computation times for finding personal information records. Method 1 scans all file types and file source data, including 100% non-HTML and 100% HTML file formats, for comparison. In Method 2, only 30% of HTML files are grabbed without considering the file download time, and all file types are examined, including 100% non-HTML file formats, and a random 30% of HTML file formats. Method 3 removes website data from files that are 90% similar to one another. The process involves deleting all file types with a similarity score of more than 90%, including 100% non-HTML and HTML file formats. Comparing the time and accuracy of the investigation can determine which method has the best efficiency and accuracy.

5.2. Experimental Results. The required times for the three methods are listed in Figure 2. According to the results, Method 1 took 3,828 min to scan, Method 2 took 1,838 min, and Method 3 took the least time 1417.3 min, saving about 63% of the time compared to Method 1.

Table 7: Evaluated methods in experiment.

No. of method	Description
Method 1	Scan all file types and source data, including 100% non-HTML and 100% HTML file formats
Method 2	Grab 30% of HTML files without considering file download time, and examine all file types, including 100% non-HTML file formats and random 30% HTML file formats
Method 3	Remove files that are 90% similar to each other (including comparison and scanning time)

The numbers of nonrepeated personal information data items are 60,005, 44,522, and 53,347 from Methods 1, 2, and 3, respectively, as shown in Figure 3.

Another performance index is the weight value from calculating the detection count per time slot. The value of W (Figure 4) for Method 1 is 10 (as a standard), and it is 11.84 for Method 2 and 21.99 for Method 3, which clearly performs best.

In summary, Method 1 took about 63 hr to investigate 10 websites. Thus, if only 30% of the HTML files are used for analysis, using Method 2 can reduce the time to 30.63 hr, saving 51.99% of the time cost while keeping the detection rate at 74.20%. Method 3 can reduce the time to 23.62 hr, saving 62.19% of the time cost while increasing the detection rate to 90.08%. Thus, Method 3 has the best value and contributes the most, as indicated in Table 8 and Figure 5.

6. Conclusions and Future Work

We proposed a precise data processing mechanism to reduce the time for processing privacy data, which is expected to

	Method 1	Method 2	Method 3
🟦 WEB_1	620	205	9
🟧 WEB_2	255	121	58
⬜ WEB_3	429	147	91
🟨 WEB_4	694	328	319
🟦 WEB_5	278	93	121.5
🟩 WEB_6	57	18	18.1
🟦 WEB_7	527	186	103
🟫 WEB_8	59	19	6.2
⬛ WEB_9	550	375	350.5
🟫 WEB_10	359	346	341
🟦 Total	3,828	1,838	1,417.3

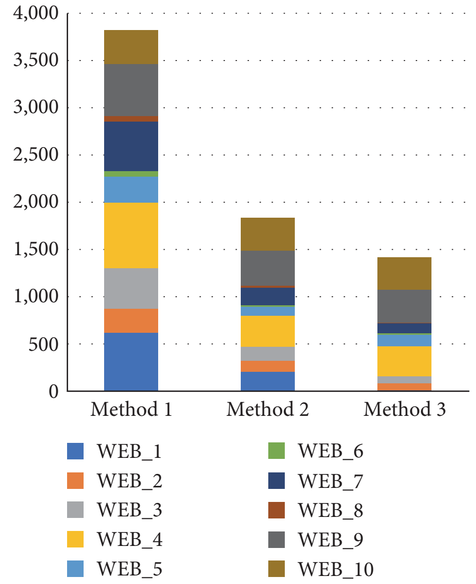

Figure 2: Execution time for each method.

	Method 1	Method 2	Method 3
🟦 WEB_1	2,230	1,002	179
🟧 WEB_2	3,188	3,132	3,070
⬜ WEB_3	1,906	1,521	1,679
🟨 WEB_4	8,095	5,625	7,549
🟦 WEB_5	20,027	13,520	18,324
🟩 WEB_6	1,273	730	941
🟦 WEB_7	3,217	2,521	2,356
🟫 WEB_8	634	272	202
⬛ WEB_9	8,572	5,368	8,146
🟫 WEB_10	10,863	10,831	10,901
🟦 Total	60,005	44,522	53,347

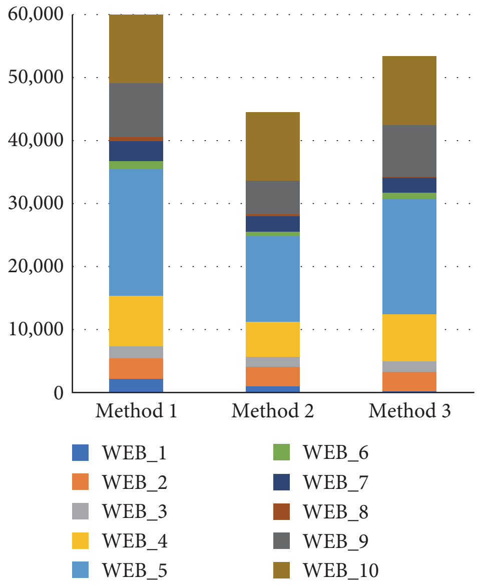

Figure 3: Number of nonrepeated personal information data items.

help find personal data leaks on the websites of smart city platforms, which often store a large amount of personal data, so our focus was to avoid the risk of data leakage. We developed a preprocessing method to prune useless data, so as to accelerate data processing. In previous experience, it was found that over 90% of Web content is unrelated to personal information. Hence, a weighting mechanism was proposed to predict and select useful raw data with a greater chance of involving personal information. According to the evaluation results, the implementation reduced the time by 62.19% while maintaining 90.08% accuracy in finding suspicious data that contained personal information.

In the next step, the authors aim to scale-up the computing capability of the system to gain throughput. Because data privacy protection is usually a continuous task, if the pattern-comparison process can be leveraged to multiple IoT nodes or integrated with current smart city implementations, it is expected to be more adaptable and flexible. The authors will continue to improve the proposed method, for lower processing times and better performance.

	Method 1	Method 2	Method 3
🟦 WEB_1	1	0.61	2.47
🟧 WEB_2	1	2.03	4.08
⬜ WEB_3	1	1.86	3.66
🟨 WEB_4	1	1.02	1.89
🟦 WEB_5	1	1.36	1.92
🟩 WEB_6	1	1.04	1.72
🟦 WEB_7	1	1.74	2.74
🟫 WEB_8	1	0.57	1.03
⬛ WEB_9	1	0.58	1.42
🟫 WEB_10	1	1.03	1.06
🟦 Total	10	11.84	21.99

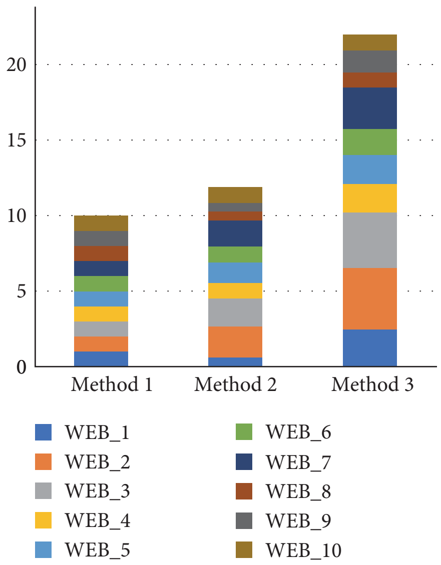

Figure 4: Performance of each method.

Table 8: Detection rate for each method.

	Method 1	Method 2	Method 3
Execution time (X)	100%	48.01%	37.81%
Keyword detection (Y)	100%	74.20%	90.08%
Detection rate ($Y \times Y/X$)	100%	114.66%	214.63%

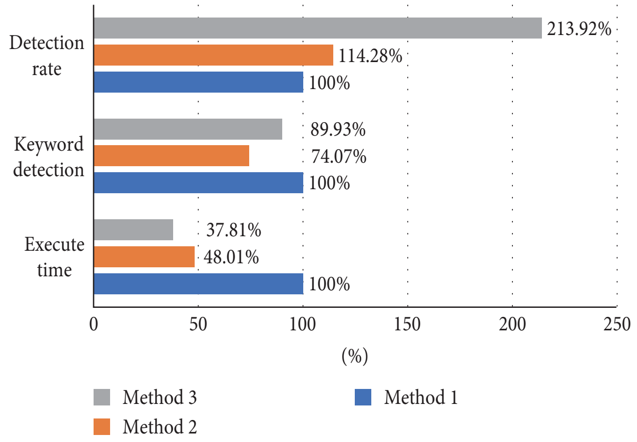

Figure 5: Comparison of detection rates.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable comments and suggestions on this paper. Authors are also grateful to TWAREN SDN and CDX teams for their great help. This research was in part supported by National Science and Technology Council of Taiwan under contract no. 109-2222-E-008-005-MY3 and 110-2218-E-006-013-MBK.

References

[1] M. Z. Alam, F. Reegu, A. A. Dar, and W. A. Bhat, “Recent privacy and security issues in internet of things network layer: a systematic review,” in *Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 1025–1031, IEEE, Erode, India, 2022.

[2] A. K. Ray and A. Bagwari, “IoT based smart home: security aspects and security architecture,” in *Proceedings of the IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 218–222, IEEE, Gwalior, India, 2020.

[3] S. Ghayyad and S. Du, “Overview on intrusion detection schemes for internet of things (IoT),” in *Proceedings of the International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pp. 1–6, IEEE, Mon Tresor, Mauritius, 2018.

[4] L. Wang, L. Pepin, Y. Li, F. Miao, A. Herzberg, and P. Zhang, “Securing power distribution grid against power botnet attacks,” in *Proceedings of the IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, Atlanta, GA, USA, 2019.

[5] M. S. Byri, M. Rajeshkumar, R. Santhakumar, and S. Balaji, “Privacy and security: internet of things,” in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–9, IEEE, Kuala Lumpur, Malaysia, 2021.

[6] P. Sandle, “BA apologizes after 380,000 customers hit in cyber attack,” September 2018, https://www.reuters.com/article/us-iag-cybercrime-british-airways/ba-apologizes-after-380000-customers-hit-in-cyber-attack-idUSKCN1LM2P6.

[7] "Consumers feel data leakage is inevitable so many have stopped caring," 2022, Available https://www.helpnetsecurity.com/2022/04/14/data-privacy-consumer-perceptions/.

[8] A. A. Abi Sen, F. A. Eassa, K. Jambi, N. M. Bahbouh, S. S. Albouq, and A. Alshanqiti, "Enhanced-blind approach for privacy protection of IoT," in *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 240–243, IEEE, New Delhi, India, 2020.

[9] A. Cavoukian, "International council on global privacy and security, by design," *IEEE Potentials*, vol. 35, no. 5, pp. 43–46, 2016.

[10] A. Alazab, S. Bevinakoppa, and A. Khraisat, "Maximising competitive advantage on e-business websites: a data mining approach," in *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 111–116, IEEE, Langkawi, Malaysia, 2018.

[11] L. Basheer and P. Ranjana, "A comparative study of various intrusion detections in smart cities using machine learning," in *2022 International Conference on IoT and Blockchain Technology (ICIBT)*, pp. 1–6, IEEE, Ranchi, India, 2022.

[12] P. Koloveas, T. Chantzios, C. Tryfonopoulos, and S. Skiadopoulos, "A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence," in *2019 IEEE World Congress on Services (SERVICES)*, pp. 3–8, IEEE, Milan, Italy, 2019.

[13] L. Li, K. Thakur, and M. L. Ali, "Potential development on cyberattack and prospect analysis for cybersecurity," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–6, IEEE, Vancouver, BC, Canada, 2020.

[14] S.-E. Tbahriti, C. Ghedira, B. Medjahed, and M. Mrissa, "Privacy-enhanced web service composition," *IEEE Transactions on Services Computing*, vol. 7, no. 2, pp. 210–222, 2014.

[15] G. P. Tiwary, E. Stroulia, and A. Srivastava, "Improving privacy in data service composition," *IEEE Access*, vol. 9, pp. 95716–95729, 2021.

[16] B. Lupton, M. Zappe, J. Thom, S. Sengupta, and D. Feil-Seifer, "Analysis and prevention of security vulnerabilities in a smart city," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 702–708, IEEE, Las Vegas, NV, USA, 2022.

[17] M. Barhamgi, C. Perera, C.-M. Yu, D. Benslimane, D. Camacho, and C. Bonnet, "Privacy in data service composition," *IEEE Transactions on Services Computing*, vol. 13, no. 4, pp. 639–652, 2020.

[18] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. S. Yu, "More than privacy: applying differential privacy in key areas of artificial intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 6, pp. 2824–2843, 2022.

[19] Y. Jiang, Q. Shen, J. Fan, and X. Zhang, "The classification for E-government document based on SVM," in *2010 International Conference on Web Information Systems and Mining*, pp. 257–260, IEEE, Sanya, China, 2010.

[20] J. Zhang, Y. Niu, and H. Nie, "Web document classification based on fuzzy k-NN algorithm," in *2009 International Conference on Computational Intelligence and Security*, pp. 193–196, IEEE, Beijing, China, 2009.

[21] G. Sahi, "Performance evaluation of artificial neural network for usability assessment of E-commerce websites," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1–6, IEEE, Pune, India, 2018.

[22] C.-C. Chiu, P.-W. Tsai, and C.-S. Yang, "PIDS: an essential personal information detection system for small business enterprise," in *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6, IEEE, Mauritius, Mauritius, 2021.

[23] M. R. Prathap, K. M. Nandhini, K. S. Vairavel, and M. V. Suraj, "Detection of data breaching websites using machine learning," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp. 1–6, IEEE, Coimbatore, India, 2021.

[24] "Fast-duplicate-file-finder," September 2022, Available https://www.mindgems.com/products/Fast-Duplicate-File-Finder/Fast-Duplicate-File-Finder-About.htm.