Development of a Cyber Incident Information Crawler

Kazuki Ikegami[1(✉)], Michihiro Yamada[1], Hiroaki Kikuchi[2], and Koji Inui[2]

[1] Graduate School of Advanced Mathematical Sciences Meiji University, 4 -21 -1, Nakano Tokyo, Japan
cs192021@meiji.ac.jp, michi_datsuteni@yahoo.co.jp

[2] School of Interdisciplinary Mathematical Sciences Meiji University, 4 -21 -1, Nakano Tokyo, Japan
kikn@meiji.ac.jp, inui@meiji.ac.jp

Abstract. This paper studies a crawler system for public cyber incident information published from official press release websites. The incidents information dataset allows us to identify the primary source of incidents and countermeasures of the cyber incidents. However, organizations publish the incident press releases in arbitrary format, which prevents us from automating annotation and classification. To address this issue, we propose a machine learning technique in conjunction with a webpage crawler and reports the accuracy and the performance of the developed cyber incident information system.

1 Introduction

The Japan Network Security Association (JNSA) collects incident information from Internet news sites and press releases published from official websites. They report incident statistics analyzed manually every year including the list of companies and the number of personal data in [1]. However, a comprehensive incident survey is difficult to run manually. For instance, domestic newspaper, Asahi Shimbun "Kikuzo 2" [2], covered 134 incidents in 2015 that do not appear in the JNSA dataset. This suggests that the incidents reported by the media were distorted by the interests of the readers of the news media.

Our study aims to comprehensively collect and classify cyber incident data automatically without any distortions. To address the issue of overload of human analyst, we develop a website crawler system that automatically collects and classifies incident information from official press release websites. This system allows us to analyze collected incident data and identify the most significant determinants of incident causes. We describe the system architecture of the crawler system and propose an efficient algorithm for classification of incidents. Finally, we assess the classification accuracy and analyze the effect of the security managements in our dataset.

L. Barolli et al. (Eds.): IMIS 2019, AISC 994, pp. 447–455, 2020.
https://doi.org/10.1007/978-3-030-22263-5_43

2 Related Works

Frank conducted a survey of the insurance market in Sweden [3]. He interviewed 10 main insurance companies, along with two reinsurance companies and three insurance intermediaries. His results showed that the Swedish cyber insurance market is growing rapidly and that these companies are not willing to deal with customers who are immature and don't have a proper security.

In the United States, the total damage to corporations and consumers from identity theft is estimated to be 56 billion dollars for 2005 [4]. Of this, up to 35% was caused by corporate data breaches. Romanosky et al. estimated the extent to which identity theft decreased after the introduction of data breach disclosure laws using panel data from the US Federal Trade Commission from 2002 to 2009. They showed that data breach disclosure laws reduced the incident of identity theft caused by data breaches by 6.1% on average [4].

Edwards et al. used a popular public dataset and developed Bayesian generalized linear models to investigate trends in data breaches [5].

3 Crawler and Automatic Classification System

3.1 Overview

Our system collects official public statements published from official press release websites. We illustrate the system architecture of our system in Fig. 1. Our system collects online press release in the following way:

1. Provide a list of URLs of the target official press release website.
2. Retrieve the related materials including HTML sources and links to other websites.
3 Convert the collected HTML into clear text.
4 Output the text database that includes specific keywords.

According to the results of previous work [6], we identify some keywords "apology" , "unauthorized" , "leak" , "breach" and "defect".[1] In Step 2, the system conducts recursive crawling as many as arbitrary specified time. Steps 2 to 4 are repeated for each company URL.

The date and the number of leaked personal data are extracted simply by the regular expression with the specific pattern.

3.2 TF-IDF Values

Term Frequency (TF) is the frequency of index term t, $n(t, d)$, in document d, that is,

$$tf(t, d) = \frac{n(t, d)}{\sum_{s \in d} tf(s, d)}. \quad (1)$$

[1] Actually, we use Japanese words "owabi" , "fusei" , "rouei" , "ryushutsu" and "fuguai", correspondingly.

Inverse Document Frequency (IDF) measures how many documents include index term t in all documents. Letting N be the number of all documents and $df(t)$ be the number of documents including index term t, the IDF is defined as

$$idf(t) = \log \frac{N}{1 + df(t)}. \tag{2}$$

The details are in [7].

3.3 Classification of Incidents

We classify the set of incidents by the cause of the leakage, e.g., "Loss/ Misplacement", "Theft", "Unauthorized Access", and "Malware". Our algorithm consists of the following steps.

1. Count the appearance of keywords in the given document and compute TF-IDF value, which gives features of the document.
2. Determine the top 49 keywords in TF-IDF value as features of the cause of the leakage.
3. Compute TF-IDF values of 49 dimensions of an incident release whose cause is unknown.
4. Compute the cosine similarity of the vector of features between the given unknown incident and each of the main causes of leakage, and calculate each cosine similarity. Finally, identify the likely cause of each incident, which has the one with highest similarity for all causes.

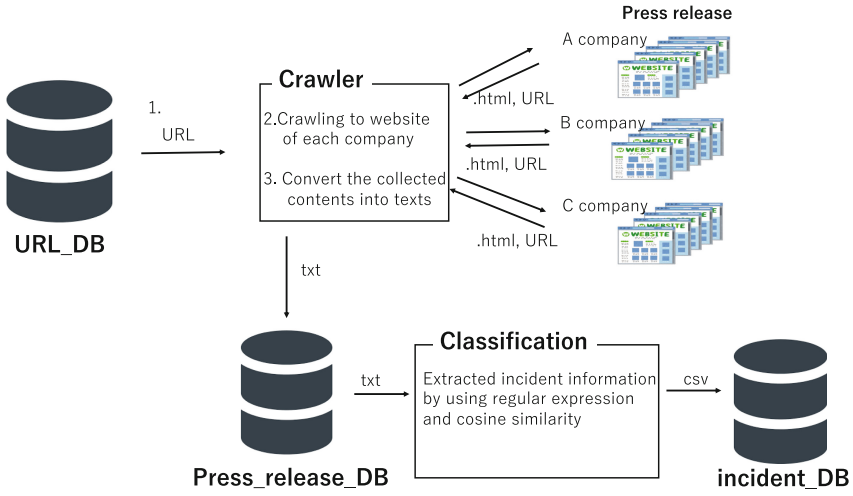

Fig. 1. System architecture of our developed system

3.4 Company Database

The Toyo Keizai Corporate Social Responsibility (CSR) Survey has been conducted every year since 2005 [8]. The investigation service Toyo Keizai sent questionnaires to 3580 listed and major unlisted companies. Our study targets 537 companies chosen from the CSR database to examine on the effects of security management. According to the survey, of the 537 companies, 169 (31%) introduced an Information Security Management System (ISMS) and 150 companies (28%) appointed a Chief Information Officer (CIO).

3.5 Crawling Result

We implemented the crawler system in the Ubuntu 18.04.2 LTS, 3,6GHz Intel Core i7 and ran in November 2018. A list of the companies covered in the CSR Dataset is presented in Sect. 3.4. We show the result of the crawling analysis in Table 1. From the collected press releases, only 191 articles (or 1%) were about cyber incidents. Besides, among 191 press releases, we found some incidents that were duplicated in several webpages, e.g., both detail and general pages. Hence, unique incidents among 191 press releases is 178.

Table 2 shows the date of the incidents, the number of victims, and the accuracy of the estimate of the causes of the incidents. The accuracy is defined as the fraction of the incidents with correctly estimated causes out of all the target incidents. The accuracy of the estimated date and number of victims exceeds 70% but falls to 50% when some attributes are combined.

Table 1. Statistics of crawling

Period	# Companies	# Collected articles	# Collected article related incident	Rate
2004/10/1–2018/11/2	537	17,957	191	(0.01)

Table 2. Accuracy of estimates

	Date	# Victims	Cause of leak	Date & victims & cause
Accuracy	0.882	0.792	0.719	0.505
	157/178	141/178	128/178	90/178

4 Evaluation

4.1 Comparison with the JNSA Dataset

We show the quality of the corrected incident information in this section. Table 3 shows the numbers of companies that suffered cyber incidents in conjunction with number of incidents. Our dataset contains 34 companies and 141 incidents from 2005 to 2016, that is, fewer incidents than the JNSA dataset.

Figure 2 illustrates the changes in the number of incidents from 2005 to 2016. The incidents reported by the JNSA dataset occurred in 537 companies.

Our crawler collected more incident press releases than contained in the JNSA databases for the period 2013 to 2016; however, it collected fewer incidents from 2005 to 2012. This was because the older press releases about cyber incidents had expired and were unavailable on official press release websites.

Table 3. Comparison of our investigation with the JNSA dataset

	JNSA	Our investigation	Common
# Companies	65	34	23
# Incidents	251	141	80

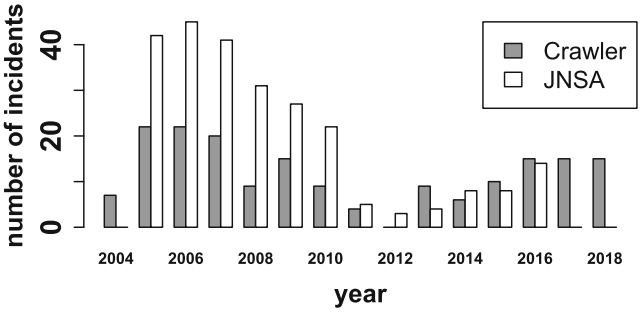

Fig. 2. Change in number of incidents

How accurately does our system retrieve information from a press release statement? Table 4 shows examples of press releases reporting cyber incidents (input) and Table 5 shows the estimated incident attributes (output). Most incident attribute values, e.g., name, date, firm type, and number of personal-information records, were extracted accurately from the incident reports. In our analysis, we regard the attribute values of the JNSA dataset as being correct. In Table 5, all extracted items are correct, but the extracted values were not always correct.

Table 4. Sample press release (input)

DeNA company 2016/04/01 In Mobage, a portal and social network for games serviced by DeNA, a malicious third party impersonating a victim user illegally gained access to the system. The total number of compromised IDs was 104,847. *(The original Japanese statement was translated into English.)*

Table 5. The extracted items (output)

	Extracted	Correct (JNSA)
Company name	DeNA	DeNA
Industry	IT companies	IT companies
Date	2016/4/1	2016/4/1
Number of victims	10,4847	10,4847
Cause of leak	Unauthorized access	Unauthorized access
Summary of incident	n/a	✓
URL	n/a	n/a
Social responsibility	n/a	Normal
Classification	n/a	Personal information
Route of leak	n/a	Internet
Post response	n/a	Normal

4.2 Estimated Classification Accuracy

Our system classifies incident information into sets of causes. The overall classification accuracy is less than 50% on average. We break down the sets of incidents by cause in Table 6, where most causes are classified correctly as indicated in the orthogonal cells. Unauthorized access includes a general security incident that exploits a known vulnerability of software and a malicious attack on a computer system. Major false classification was made in difference between the "Loss/ Misplacement" and the "Administration Error", which was made up of the 75%. As the difference between the "Loss/ Misplacement" and the "Administration Error" is subtle, even a human analyst could have difficulty in making the distinction. Therefore, we claim that the classification quality of our system is good in practical use.

5 Effect of Security Management

5.1 Method

We use a logistic regression to estimate how much security management reduces risk of cyber incident without suffering confounding factor. The probability p_{iy}

Table 6. Number of incidents classified by cause

Estimated \True	Loss Misplacement	Administration Error	Theft	Operational Error	Unauthorized Access	Malware	Others
Loss/ Misplacement	53	9	1	0	0	0	2
Administration Error	5	11	1	0	0	0	0
Theft	3	2	29	0	0	0	4
Operational Error	0	0	0	12	2	0	1
Unauthorized Access	0	0	0	1	10	1	2
Malware	0	0	0	0	1	5	1
Others	0	1	0	0	4	2	8

of an incident occurring in company i in year y is

$$p_{iy} = \frac{1}{1 + e^{-z_i}}. \tag{3}$$

Then, we assume it as

$$z_i = \alpha + \beta_i b_i + \beta_y c_y + \beta_d d_d + \beta_{x_1} x_1 + \cdots + \beta_{x_m} x_m, \tag{4}$$

where b_i, c_y and d_i are dummy variables to counteract effect of incident, e.g., industry, survey years and scale of company, respectively. And, x_m is a vector of explanatory variables. The use of security management or not is represented by Boolean values. α is a constant and βs are coefficients of each variable. The adjusted odds ratio in x_1 is

$$OR = e^{\beta_1}. \tag{5}$$

5.2 Results

ISMS is designed to reduce the risk of cyber incidents. However, it has been unclear to what degree security management prevents an organization from being compromised. To address this issue in this section, we use the crawling incident data in conjunction with the CSR dataset and conduct a multivariable logistic regression, where the occurrence of an incident is the target variable of the regression and the explanatory variables include various types of security management such as ISMS, auditing, and CIO.

We show the results in Table 7. A positive coefficient means that the probability of the incident occurring increases when security management is adopted in the organization. In our analysis, external system auditing operates most significantly to reduce the chance of an incident. Some security managements such as ISMS, CIO, and internal auditing have positive coefficients whereas there are negative coefficients in [9]. We plan to study the reasons for this difference in future research.

Table 7. Logistic regression in our dataset

		Our dataset			[9]
		Estimate	Std.ERR	Pr(>|z|)	Estimate
a	(Intercept)	−23.260	2084.000	0.991	−8.300
b	Construction & Materials	16.280	2084.000	0.994	0.223
	Raw materials & chemicals	16.950	2084.000	0.994	−0.046
	Automobiles & transportation equipment	−0.279	4188.000	1.000	−0.334
	Steel & nonferrous metals	−0.012	3548.000	1.000	−0.838
	Electric Appliances $ Precision Instruments	16.260	2084.000	0.994	0.091
	IT/Services, others	18.120	2084.000	0.993	0.561
	Electric power & gas	20.330	2084.000	0.992	2.436
	Transportation & logistics	0.367	14320.000	1.000	0.829
	Commercial & wholesale trade	0.467	3777.000	1.000	0.066
	Retail trade	17.540	2084.000	0.993	0.904
	Financials (ex banks)	1.145	14510.000	1.000	0.209
	Machinery	−0.219	0.065	1.000	−0.219
c	2014	−0.350	0.724	0.629	0.221
	2015	−0.763	0.784	0.330	0.185
	2016	0.752	0.595	0.206	0.185
	2017	1.186	0.706	0.093**	−0.193
d	LOG(# employee)	0.399	0.366	0.275	0.948
x	ISMS	1.200	0.674	0.075**	−0.217
	CIO	0.000	0.719	1.000	−1.097
	Internal inspection	2.429	1.816	0.181	−0.207
	External inspectin	−0.959	0.428	0.025*	0.117
	Internal report window	−1.717	1.823	0.346	−0.050
	External report window	−0.969	0.789	0.220	−0.685

6 Conclusions

In this study, we crawled the websites of 537 companies and collected data on 178 incidents, which is more incidents than contained in the JNSA dataset for the period 2013 to 2016. The crawling analysis found that 1% of all public press releases we collected were classified in incident. We found 61 new incidents that were not covered in the JNSA dataset.

Our system automatically classifies incident data into classes with causes, and retrieves information about the incidents including date, number of victims, and cause of leak with accuracy of more than 70%. However, the overall accuracy of our system for all items is less than 50%.

Our future research will consider how to improve the identification accuracy of causes, increase the coverage of companies, and provide open databases for incidents.

References

1. Japan Network Security Association: Information security incident survey report (JNSA dataset), 2003–2016
2. Kikuzo 2: Asahi Shimbun Company, online article database
3. Franke, U.: The cyber insurance market in Sweden. Comput. Secur. **68**, 130–144 (2017)
4. Romanosky, S., Telang, R., Acquisti, A.: Do data breach disclosure laws reduce identify theft? J. Policy Anal. Manage. **30**(2), 256–286 (2011)
5. Edwards, B., Hofmeyr, S., Forrest, S.: Hype and heavy tails: a closer look at data breaches. J. Cybersecur. **2**(1), 3–14, (2016)
6. Ikegami, K., Kikuchi, H.: Data mining of reasons of data breach based on the information leakage data set. In: The 80th National Convention of IPSJ, 2W-06, vol. 3, pp. 543–544 (2018)
7. Tokunaga, T.: Information Search and Language Processing. University of Tokyo Press, Tokyo (2009)
8. Toyo Keizai Data Services, CSR DATA, 2014–2016
9. Yamada, M., Ikegami, K., Kikuchi, H., Inui, K.: Assessment of the effect of decreasing data breach by the management situation (2). In: Computer Security Symposium (CSS2018), pp. 376–384 (2018)