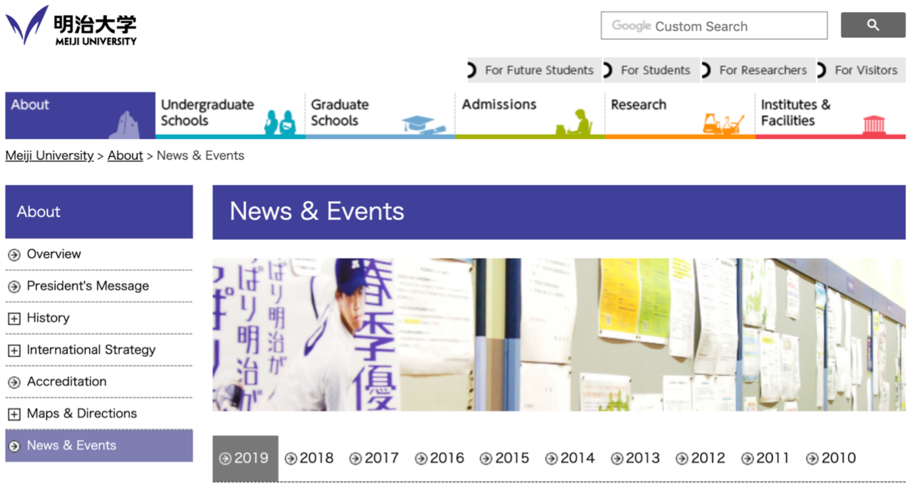
Development of Cyber Incident Information Crawler

Kzuki ikegami,* Michihiro Yamada,* Hiroaki Kikuchi and Koji Inui

*Graduate School of Advanced Mathematical Sciences Meiji University,

Background

- Increasing of unauthorized access
- Damage to companies
  - Lost and altered data
  - Business interference
  - reputation
- Necessity of security management
  - Certification of Information Security Management System ISMS decreases cyber incident risk by 20%.[1]



明治大学 MEIJI UNIVERSITY

Google Custom Search

For Future Students | For Students | For Researchers | For Visitors

About | Undergraduate Schools | Graduate Schools | Admissions | Research | Institutes & Facilities

Meiji University > About > News & Events

About

- Overview
- President's Message
- History
- International Strategy
- Accreditation
- Maps & Directions
- News & Events

News & Events

2019 2018 2017 2016 2015 2014 2013 2012 2011 2010

Personal data of 1147 students were disclosed

[1] M. Yamada, K. Ikegami, H. Kikuchi and K. Inui, Assessment of the effect of decreasing breach by the management situation (2), CSS2018, 2018.

Previous study

- Number of Cyber incidents / year [2]

name	JNSA: Japan security network association	Asahi Shimbun domestic newspaper	common
way	Some medias	a news paper	
# incident	788	279	145

[2] K.Ikegami,H.Kikuchi,Dataminingofreasonsofdatabreachbasedontheinformationleakage data set,
The 80th National Convention of IPSJ, 2W-06, vol.3, pp. 543–544, 2018.

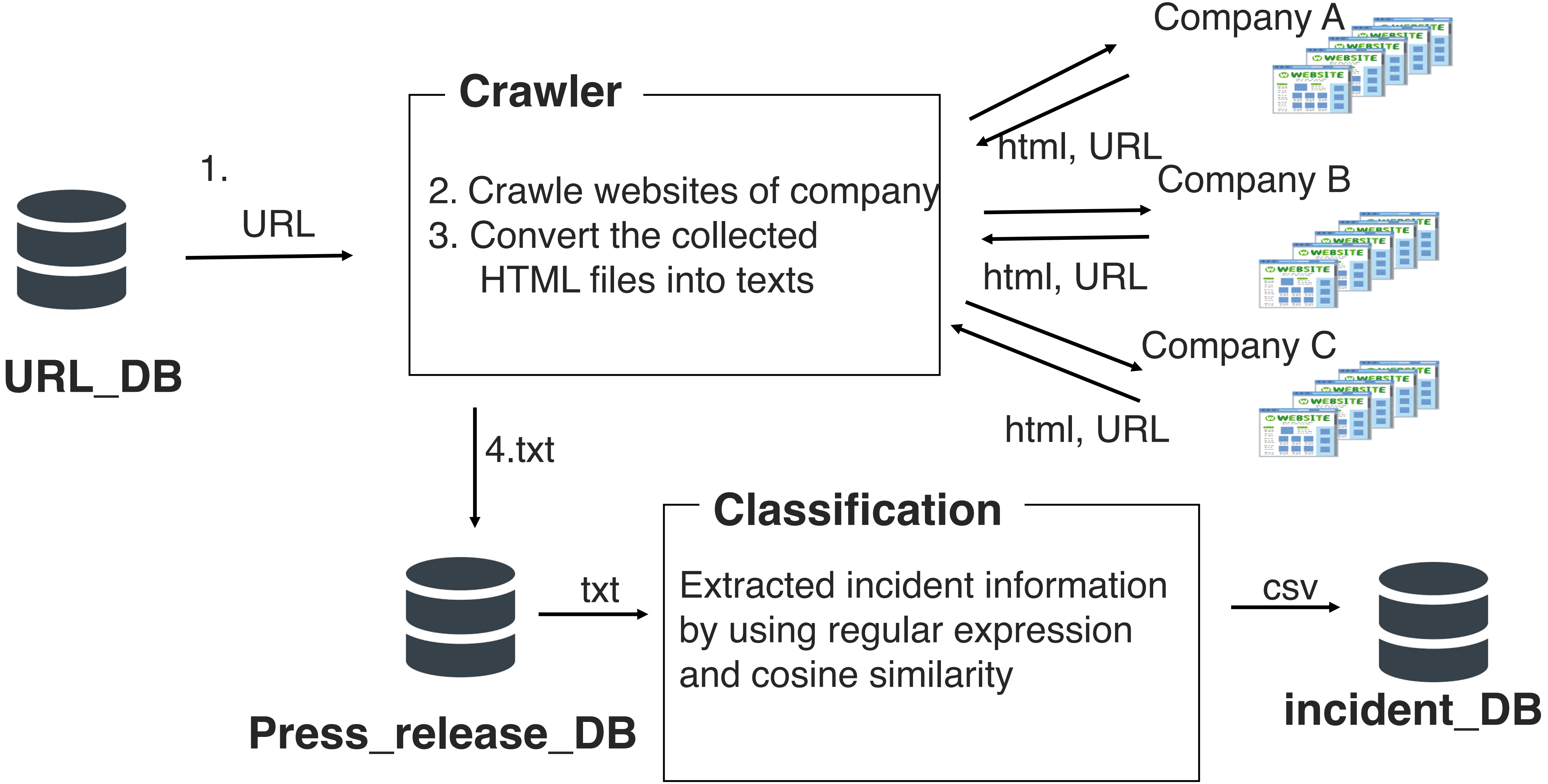
Problems

- Distortion
  - The incident reported by media were distorted by interests of the readers of news media.
- Cost
  - Investigation of all cyber incident for one year takes 3 days.

Our study

- Purpose
  - To comprehensively collect and classify cyber incident data automatically without any distortions.
- Approach
  - Develop a website crawler system for collecting cyber incident.
  - Develop a system which automatically classifies cyber incidents into some causes.

How our system works



Press release

Company A

Company B

Company C

Crawler

2. Crawle websites of company
3. Convert the collected
   HTML files into texts

1.

URL

URL_DB

html, URL

html, URL

html, URL

4.txt

Press_release_DB

txt

Classification

Extracted incident information by using regular expression and cosine similarity

csv

incident_DB

Features: TF-IDF values and 49 dimensions vector

- TF : frequency of index term *t* in document *d*.
- IDF : inverse frequency of documents that include index term *t*.

	Term	TF	IDF	TF-IDF
Accept	*password*	0.006	2.696	0.016
	website	0.003	3.572	0.011
	unauthorize	0.004	2.696	0.011
Reject	*server*	**0.001**	3.635	0.003
	account	0.004	**2.283**	0.009
	countermeasure	0.002	2.879	0.006

Ex) Classification

Extract by regular expression

Input(release)

Feb/2/2019
Some PCs were unauthorized accessed in A company.
The number of compromised Personal data was 3000.

features	input
lost	0
unauthorized	1
PC	1
entrustment	0
E-mail	0

	Human error	**Unauthorized access**	**Insider**
	3	0	0
	0	2	0
	1	1	4
	1	1	3
	2	0	0
Cosine similarity	0.33	0.87	0.45

Output : {*date* : Feb/2/2019, *scale* : 3000, *cause* : unauthorized access}

The extracted items

Input

DeNA Co., Ltd.
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)

Output

	Extracted	Correct(JNSA)
Company name	DeNA	DeNA
industry	IT companies	IT companies
date	2016/4/1	2016/4/1
Number of victims	104847	104847
Cause of leakage	Unauthorized access	Unauthorized access
Summary of incident	n/a	✓
URL	n/a	n/a
Social responsibility	n/a	normal
Kind of breach	n/a	Personal information
means of leakage	n/a	internet
Post response quality	n/a	normal

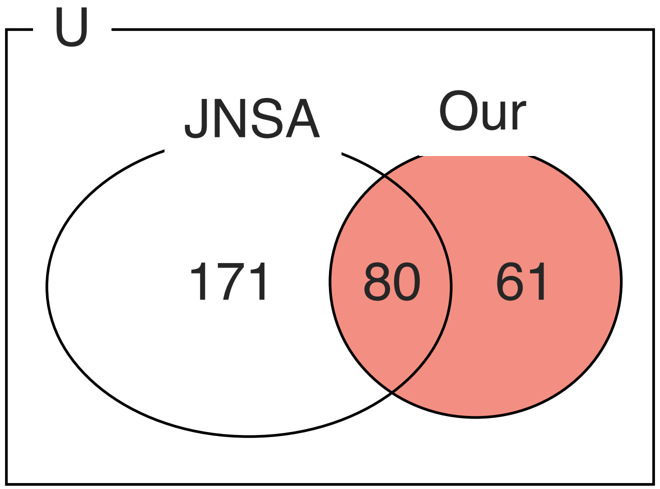
Statistics

- **statistics**

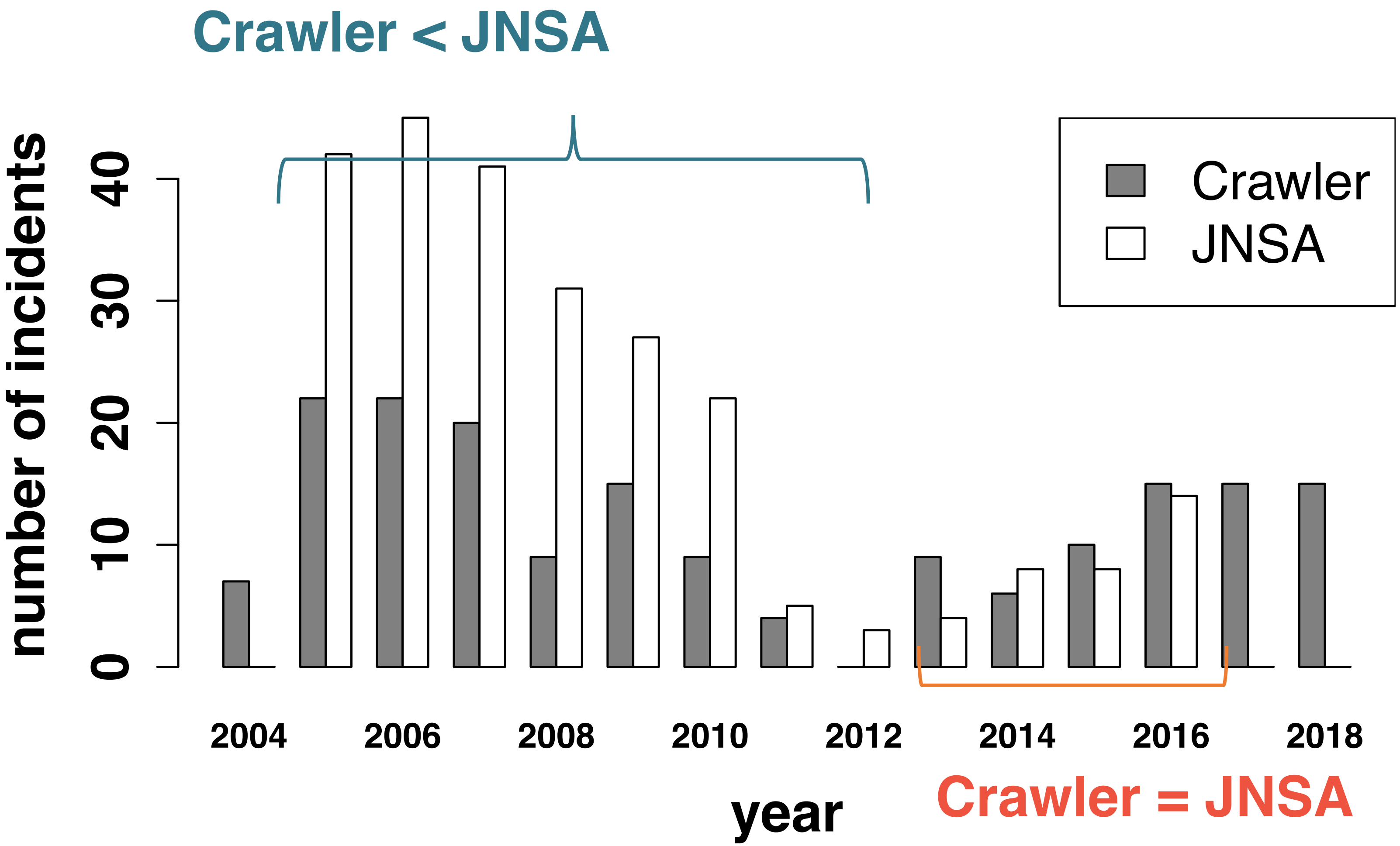
duration	# companies	# collected Press releases	# collected press releases related incident	rate
2004/10/1 - 2018/11/2	537	17,957	191	1%

- **comparison(2004-2016)**

	JNSA	Our data	common
# companies	65	34	23
# incidents	251	141	80



Change in number of incidents



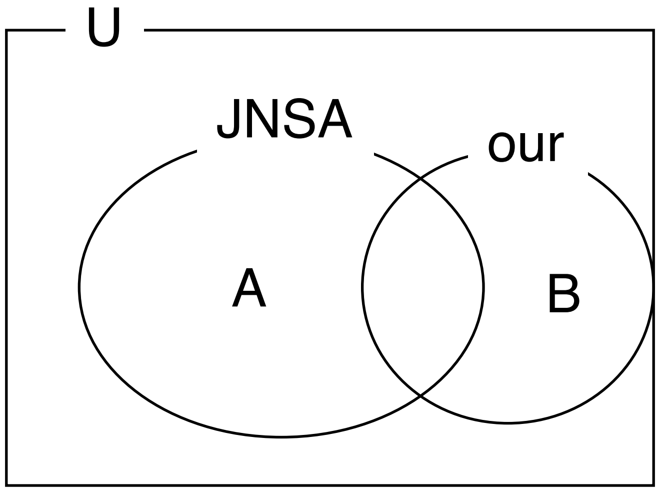
Example of incidents

A High interests

CyberAgent, Inc.
Jan 1, 2010
Probably 450 user IDs and passwords were compromised.

B Low interests

Tokyo Gas Co., Ltd.
December 8, 2016
An employee lost 3,463 receipts including customer information.



Accuracy of estimates

$$\text{Accuracy} = \frac{\text{the incidents with correctly estimated causes}}{\text{all the target incidents}}$$

	date	# victims	cause	date & victims & cause
accuracy	0.882 157/178	0.792 141/178	0.719 128/178	0.505 90/128

The accuracies of each of items exceed 70%
but fall to 50% when some attributes are combined.

Effect of Security Management

- The probability of incident occurring : $p = \frac{1}{1+e^{-z}}$

$$z = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_m X_m$$

  - β_i : coefficients of each variable
  - X_i : vector of explanatory variables. E.g., # employee, management and industry

- The adjusted odds ratio in β_1 is

$$\mathrm{OR} = e^{\beta_1} = \frac{p}{1-p} / \frac{q}{1-q}$$

	With *M*	Without *M*
Incident	p	q
No incident	$1-p$	$1-q$

M : management

Result of logistic regression

	(Intercept)	# employee	ISMS	CIO	External inspection
Estimate(β)	-23.26	0.399	1.222	0.0002	-0.959
Odds ratio	0.000	1.49	3.32	1.00	0.383

The probability of the incident occurring in ISMS certified companies is three times higher than that of company without certification.

Conclusions

- We have developed automatic crawler system that collected more than 190 articles related with incident.
- The accuracy of single item exceeds 70%.
- The coverage of incidents we collected are as same as JNSA Dataset after 2013.
- As future research, we will consider how to improve the identification accuracy of causes, increase the coverage of companies, and will provide open databases for incidents.