Bulletin of Electrical Engineering and Informatics
Vol. 7, No. 3, September 2018, pp. 393~399
ISSN: 2302-9285, DOI: 10.11591/eei.v7i3.847

An Approach for Risk Estimation in Information Security Using Text Mining and Jaccard Method

Prajna Deshanta Ibnugraha[1], Lukito Edi Nugroho[2], Paulus Insap Santosa[3]
[1,2,3]Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia
[1]School of Applied Science, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Nov 26, 2017
Revised Aug 09, 2018
Accepted Aug 23, 2018

Keywords:

Information security
Information value
Jaccard method
Risk analysis
Text mining

ABSTRACT

Involvement of digital information in almost of enterprise sectors makes information having value that must be protected from information leakage. In order to obtain proper method for protecting sensitive information, enterprise must perform risk analysis of threat. However, enterprises often get limitation in measuring risk related information security threat. Therefore, this paper has goal to give approach for estimating risk by using information value. Techniques for measuring information value in this paper are text mining and Jaccard method. Text mining is used to recognize information pattern based on three classes namely high business impact, medium business impact and low business impact. Furthermore, information is given weight by Jaccard method. The weight represents risk levelof information leakage in enterprise quantitatively. Result of comparative analysis with existing method show that proposed method results more detailed output in estimating risk of information security threat.

Corresponding Author:

Prajna Deshanta Ibnugraha,
Department of Electrical Engineering and Information Technology,
Universitas Gadjah Mada, Yogyakarta, Indonesia.
Email: prajna.deshanta.i@mail.ugm.ac.id

1. INTRODUCTION

The digital information supports business of enterprises by giving knowledge to users such as staffs, investors, customers and business management. Most of digital information consist of high sensitivity value, so protection method is needed to be applied in system of information technology [1]. However, data leakage due to flaws of IT system isstill occurred and it causes serious impact to enterprises. Average cost from incident of data leakage is about US$3.8 million [2]. In order to reduce impact of incident, enterprise must identify threat and perform mitigation. Appropriate mitigation procedure can be formulated after enterprises know about level of risk in threat. However, estimation of risk level is not simple thing. Enterprises must use risk model as reference to calculate risk level.

In data leakage case, risk level can be estimated by qualitative or quantitative method. Importance level of information is used commonly in qualitative method and financial value is used in quantitative method [3]. The problem of financial approach in quantitative method is difficult to be implemented because user must know representation of information in financial metrics Users are also required to have direct access to financial report. It becomes new challenge to identify new approach to estimate risk of information value in quantitative method.

An approach of information value estimation is performed by giving weighting for information term. Pribadi et al. gave weight for information value in automated short answer scoring case [4]. Five classes were used to represent information value, i.e. highly important term, very important term, important term, fairly important term and not important term.Weighting of information term from Pribadi et al. study can be adopted to estimate risk of information security. Meanwhile, high business impact (HBI) term,

medium business impact (MBI) term and low business impact (LBI) term can be used as classes in forming of risk level [5]. Based on previous study related weighting and classification of information term above, this paper has objective to develop new approach for risk estimation in information security area.

In order to reach objective, we divide this paper to several sections. Section 2 reveals previous study related risk estimation and information measurement. Differences between previous study and proposed method are explained in this section. Section 3 is research method where it contains methodology to achieve goal. Experimental details are described in section 4. In section 4, we explain process to form risk level of information security threat from data source. Result from experimental details is revealed in section 5. Comparation between proposed method and previous method is also explained in section 5.

2. RELATED WORK

In previous study, information value in enterprises is related with meaning of information to business [3]. In risk analysis, information value has three categories namely High Business Impact (HBI), Medium Business Impact (MBI) andLow Business Impact (LBI) [5]. High Business Impact (HBI) is data that has a severe impact for information owner and organization in case of data leakage. Information that has impact in reputation damage, is included Medium Business Impact (MBI) category, whereas Low Business Impact (LBI) is information that has limited impact to owner of information or organization.

Some methods are also used to estimate information value in previous study. Sajko et al. developed method to measure information value by calculating volume of information. Volume is calculated by three variables like meaning information in business, time and cost for producing information [3]. Dimension from volume of information can be presented in Figure 1.

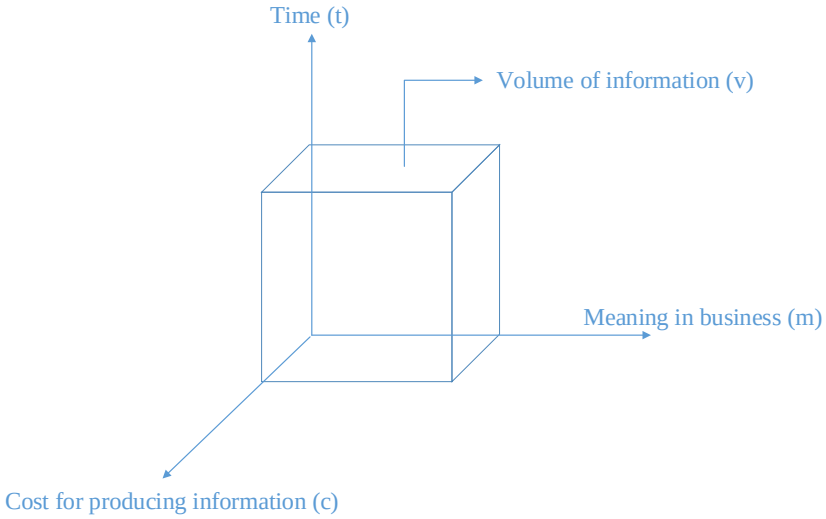

Figure 1. Dimension of information volume

Information value was represented by volume of information ($\boldsymbol{V_{inf}}$). Meanwhile, volume of information was measured by involving three variables namely meaning information in business (***m***), time (***t***) and cost for producing information (***c***). Relation between volume of information and its variables can be descrbed in Equation 1.

$$Information\ Value = V_{inf}\{m, t, c\} \quad (1)$$

Weight for each variable in Formula 1 was obtained from survey method. Assessment tool was built in questionnaire. Experts as respondents of assessment chose ordinal value or interval value as option to represent weight of variables.

However, use of expert opinion for filling weight of variables gives subjective grade in information value estimation. Therefore, Gao et al. developed new approach to estimate information value because use of expert opinion in assessment was old ways that increased complexity in operation [6]. Clustering method and Fuzzy algorithm were used by Gao et al. to estimate information value. Fuzzy algorithm was used to quantify

information risk factors. Furthermore, results of Fuzzy processing were divided in four clusters namely L1, L2, L3 and L4. L1 represents the minimum value and L4 represents the highest value. Clustering method used by Gao et al was K-Means. In comparison, method from Gao et al. is more objective than method from Sajko et al. However, method from Gao et al needs minimum number of data as data training in early process.

This paper uses different approach to view business risk of enterprises. Number of data leakage becomes point for estimating risk. Enterprise has high risk if it has big number of data leakage that involves sensitive information. Therefore, involvement of text mining and Jaccard method becomes important thing to develop new approach of risk analysis in this paper. Text mining is used to classify sensitivity of information and Jaccard method is used to calculate weight of information. Use of text mining to classify information was ever used by Data Leakage/Loss Prevention (DLP) [7], [8] whereas Jaccard method was used for weighting in similarity function of information retrieval system [9], [10] and plagiarism detection [11]. Jaccard method is possible to be implemented for estimating value of information in document by calculating number of specific term that represents a category in this paper [4].

3. RESEARCH METHOD

Proper mechanism for processing unstructured data is text mining [12]. It can be used in classification function that categorizes sets of string and inputs appropriate word into a category [13]. Regular expression is technique that can be used to recognize word for a category by pattern matching or keyword matching. Steps of research in this paper are shown in Figure 2:

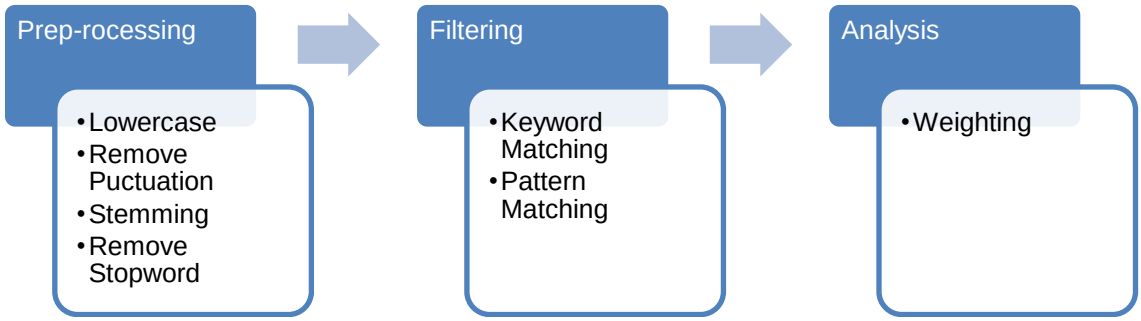

Figure 2. Steps of research method

In pre-processing step, sets of string in document are processed through lowercase conversion, punctuation removing, stemming, tokenization and stopwords removing. Pre-processing step prepares document of data sourceso it can be processed in filtering step.

Filtering step is to define categories and criteria. This paper refers to Ruivo et al. categories and criteria [5]:

- High Business Impact
  It consists of words: passwords, bank account, credit card number
- Medium Business Impact
  It consists of words: information of customer specification
- Low Business Impact
  It consists of words: gender, address

Categories and criteria are implemented in wordlist. Regular expression technique refers to that wordlist for recognizing term in document ofdata source [14].

Categorized words are processed in analysis step by estimating weight for every word in a category. Weight total of each category is sum of weight of words in that category. Jaccard method is used to estimate weight for each word in category. It calculates ratio between number of occurrences word that defined in wordlist (***W***) and total number of unique words in document (***D***) [15]. Result of Jaccard method is coefficient. Jaccard coefficient (***Jaccard (W, D)***) is determined from division operation of intersection size (***W∩D***) and union size (***W∪D***). Calculation of Jaccard coefficient uses Equation 2.

$$Jaccard(W,D) = \frac{W \cap D}{W \cup D} \quad (2)$$

4. EXPERIMENTAL DETAILS

This paper defines disclosed information from information security assessment as data source. In order to obtain disclosed information, SQL injection attack is used as assessment method of information security. Table 1 is result of information security assessment.

Table 1. Experimental Data Source

username	password	lastlogin	status	email
User1	af59b75d998d4e6869caea0b22bc8f5c	15th December 2015 03:17PM	active	email1
User2	b2805c093f83761e5aba2a145067ddc7	2nd November 2015 03:34AM	active	email2
User3	b2805c093f83761e5aba2a145067ddc7	5th November 2015 01:07AM	active	email3
User4	8e721d1c51f5109c989c77d9275fcf61	1st November 2015 12:24PM	active	email4
User5	b2805c093f83761e5aba2a145067ddc7	6th November 2015 09:23AM	active	email5
User6	b2805c093f83761e5aba2a145067ddc7	8th September 2015 09:40AM	active	email6
User7	b2805c093f83761e5aba2a145067ddc7	16th October 2015 09:42PM	active	email7
User8	b2805c093f83761e5aba2a145067ddc7	28th August 2015 11:15PM	active	email8
User9	b2805c093f83761e5aba2a145067ddc7	3rd October 2015 06:22AM	active	email9
User10	b2805c093f83761e5aba2a145067ddc7	27th December 2014 03:42AM	active	email10
User11	b2805c093f83761e5aba2a145067ddc7	29th December 2014 07:00AM	active	email11
User12	b2805c093f83761e5aba2a145067ddc7	19th April 2005 04:15PM	active	email12
User13	b2805c093f83761e5aba2a145067ddc7	19th April 2005 06:26PM	active	email13
User14	b2805c093f83761e5aba2a145067ddc7	13th October 2015 03:56AM	active	email14
User15	b2805c093f83761e5aba2a145067ddc7	20th April 2005 11:11AM	active	email15
User16	b2805c093f83761e5aba2a145067ddc7	24th November 2014 03:32AM	active	email16
User17	b2805c093f83761e5aba2a145067ddc7	22nd April 2005 11:14AM	active	email17
User18	b2805c093f83761e5aba2a145067ddc7	27th December 2014 03:44AM	active	email18
User19	b2805c093f83761e5aba2a145067ddc7	26th April 2005 02:24PM	active	email19
User20	b2805c093f83761e5aba2a145067ddc7	22nd April 2005 01:46PM	active	email20
User21	b2805c093f83761e5aba2a145067ddc7	24th April 2005 12:50PM	active	email21
User22	b2805c093f83761e5aba2a145067ddc7	25th April 2005 10:15AM	active	email22
User23	b2805c093f83761e5aba2a145067ddc7	05th February 2006 02:04PM	active	email23
User24	b2805c093f83761e5aba2a145067ddc7	12th November 2014 03:49AM	active	email24
User25	b2805c093f83761e5aba2a145067ddc7	6th November 2014 05:47AM	active	email25

*) users and emails are censored for security reason

Based on data format of Table 1, high business impact (HBI) category is represented in column name “username” and “password”, whereas low business impact (LBI) category is represented in column name “email”. Other content is defined as public information where it does not have impact to company business. It is caused public information having goal for public reader. Algorithm for classification can be described in Algorithm 1.

Algorithm 1: To identify unique word for as part of categories

Notation:

document	:	List of data
wordlist	:	list of string for defining token categories
token	:	word from output of text mining
tokens	:	list of tokens
num_tokens	:	number of tokens

Input: file
Var:
document, wordlist, token, tokens:string
num_tokens:integer
Begin
If(token Є wordlist)==true**then**
fori=1 **to** N **then**
tokens ←r**ead**(token Є document)
end for
num_tokens←**count**(tokens)
End

Output:

$$tokens = \sum_{i=1}^{n} token(i)$$

$$num\_tokens = \sum tokens$$

5. RESULTS AND ANALYSIS

Data source in Table 1 has twocategories namelyhigh business impact and low business impact. High business impact (HBI) consists of usernameinformation and password information. Meanwhile, email is categorized low business impact (LBI) information. Figure 3 represents data distribution from categories of result.

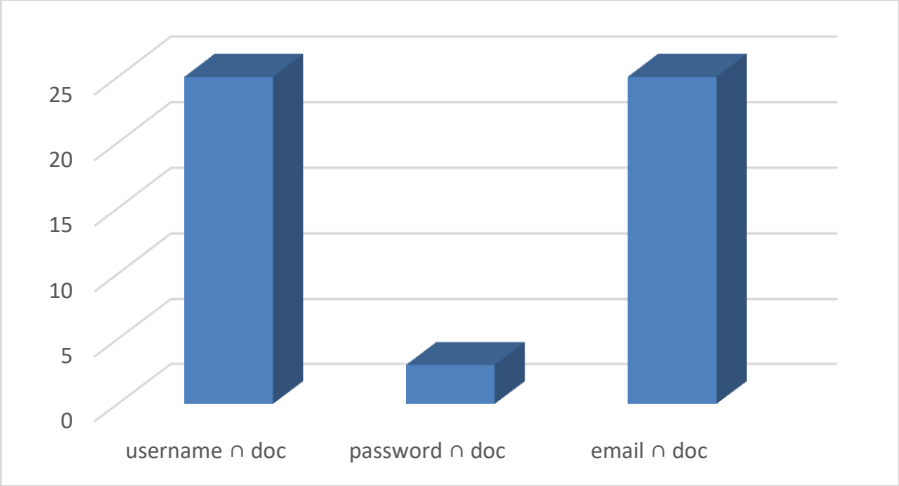

Figure 3. Data distribution from classification

Calculation of intersection and union size is conducted by involving data classification from previous process. It results three intersection data and one union data. Table 2 describes result of intersection and union calculation.

Table 2. Jaccard Variables from Data Source

Variable	Value
Intersection size between usernameand document from data source	25
Intersection size between password and document from data source	3
Intersection size between email and document from data source	25
Union size between (username,password,email) anddocument from data source	79

Jaccard coefficient is obtained by calculating intersection from each category over union of document. In high business impact (HBI) category, Jaccard coefficient is resulted from combination intersection of username and password over union of document. Meanwhile, Jaccard coefficient from low business impact (LBI) category is obtained from intersection of email over union of document. Result of Jaccard coefficient calculation is shown in Table 3.

Table 3. Jaccard Coefficient

Jaccard Variables	Jaccard Coefficient	Category
Jaccard((username,password), doc)	0,354	HBI
Jaccard(email,doc)	0,316	LBI

Jaccard coefficient represents weight of risk for every category. In caseabove, disclosed data from enterprise contains 67% of sensitive information where it consists of 35,4% in high business impact and 31,6% in low business impact. Measurement will result different output when it is implemented in different data from different enterprises. It is caused by differences ofvalue and characteristic from information in each enterprise.

In order to obtain description related result of proposed method, this paper makes comparative analysis with existing method for estimating risk. In general method, risk of information security is measured by two variables namely probability and impact [16]. Relation between risk, probability and impact can be described in Equation 3.

$$Risk = Probability\ x\ Impact \quad (3)$$

Open Web Application Security Project (OWASP) risk rating is one of methods where it implements relation between risk, probability and impact to measure risk [17]. In OWASP risk rating, probability is represented by likelihood variable and impact consists of technical impact and business impact. OWASP risk rating also considers business perspective to estimate risk so it has similar approach with proposed method. Therefore, OWASP risk rating can be chosen as comparison method in process of comparative analysis.Based data in Table 1, measurement of OWASP risk rating results medium risk level for technical impact and business impact. Description of result from OWASP risk rating is shown in Table 4.

Table 4. Measurement Result of OWASP Risk Rating

Aspects	Likelihood	Impact	Risk
Technical	Medium	Medium	Medium
Business	Medium	Medium	Medium

In order to show result of comparative analysis, this paper uses three categories namely method, aspect and experimental result. Comparison result between proposed method and OWASP risk rating can be shown in Table 5.

Table 5. Comparative RESULT

	Proposed Method	OWASP
Method	Text Mining + Jaccard Method	Probability x Impact
Aspect	Business	Business Technical
Experimental Result	High Business Impact Low Business Impact	Medium

In comparative result, proposed method and OWASP risk rating have different approach to estimate risk from threat of information security. Both methods also result different risk level in business aspect. However, proposed method has advantages in resulting more detailed risk because it examines each information from disclosed data. It is different with OWASP approach where OWASP view threat of information security generally with subjective measurement. In previous research, Jaccard method was faster than Cosine Distance algorithm in filtering data [18] so it becomes another advantage from proposed method.

6. CONCLUSION

This paper proposes different perspective to measure risk of information disclosure. It uses information value to determine risk of information disclosure. Information value in this paper is estimated by classification and weighting process. Sensitive information from data leakage is classified by High Business Impact (HBI), Medium Business Impact (MBI) and Low Business Impact (LBI). Text mining based on keyword and pattern matching is used as method to classify sensitive information from data leakage. Weight is given to classes by Jaccard method. Weight is used to give description related risk level quantitatively. Calculation of weight involves intersection and union size. In experimental details, data source from an organization results two categories of impact i.e. high business impact and low business impact. High business impact has weight about 0,354 and low business impact has weight about 0,316. The experimental result states that leakage data has 35,4% high sensitive information and 31,6% low sensitive information. In

order to obtain advantages of proposed method, comparative analysis is performed by comparing proposed method with OWASP risk rating. Comparison of both methods results conclusion that proposed method has more detailed output.

REFERENCES

[1] S.K. Pandey and K. Mustafa, "A Comparative Study of Risk Assessment Methodologies for Information Systems", *Bull. Electr. Eng. Informatics*, vol. 1, no. 2, pp. 111–122, 2012.

[2] D. Gugelmann, P. Studerus, V. Lenders, and B. Ager, "Can Content-Based Data Loss Prevention Solutions Prevent Data Leakage in Web Traffic?", *IEEE Secur. Priv.*, vol. 13, no. 4, pp. 52–59, 2015.

[3] M. Sajko, K. Rabuzin, and M. Bača, "How to calculate information value for effective security risk assessment", *J. Inf. Organ. Sci.*, vol. 30, no. 2, pp. 263–278, 2006.

[4] F.S. Pribadi, T.B. Adji, and A.E. Permanasari, "Automated Short Answer Scoring using Weighted Cosine Coefficient", *2016 IEEE Conf. e-Learning, e-Management e-Services, IC3e 2016*, pp. 70–74, 2017.

[5] P. Ruivo, V. Santos, and T. Oliveira, "Data Protection in Services and Support Roles – a Qualitative Research amongst ICT Professionals", *Procedia Technol.*, vol. 16, pp. 710–717, 2014.

[6] G. Gao, X. Li, B. Zhang, and W. Xiao, "Information Security Risk Assessment Based on Information Measure and Fuzzy Clustering", *J. Softw.*, vol. 6, no. 11, pp. 2159–2166, 2011.

[7] B. Hauer, "Data and information leakage prevention within the scope of information security", *IEEE Access*, vol. 3, pp. 2554–2565, 2015.

[8] A. Shabtai, Y. Elovici, and L. Rokach, "Data Leakage Detection/Prevention Solutions", in *A Survey of Data Leakage Detection and Prevention Solutions*, 2012, pp. 17–37.

[9] K. Rinartha and W. Suryasa, "Comparative Study for Better Result on Query Suggestion of Article Searching with MySQL Pattern Matching and Jaccard Similarity", in *5th International Conference on Cyber and IT Service Management (CITSM)*, 2017, pp. 1–4.

[10] M. Erritali, A. Beni-Hssane, M. Birjali, and Y. Madani, "An Approach of Semantic Similarity Measure between Documents Based on Big Data", *Int. J. Electr. Comput. Eng.*, vol. 6, no. 5, p. 2454, 2016.

[11] S. Wang, H. Qi, L. Kong, and C. Nu, "Combination of VSM and Jaccard coefficient for external plagiarism detection", *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 4, pp. 1880–1885, 2013.

[12] N. Naw, "Relevant Words Extraction Method in Text Mining", *Bull. Electr. Eng. Informatics*, vol. 2, no. 3, pp. 177–181, 2013.

[13] R.S.A and S. Ramasamy, "Context Based Classification of Reviews Using Association Rule Mining , Fuzzy Logics and Ontology", *Bull. Electr. Eng. Informatics*, vol. 6, no. 3, pp. 250–255, 2017.

[14] S. Sun, Q. Li, P. Yan, and D.D. Zeng, "Mapping Users across Social Media Platforms by Integrating Text and Structure Information", pp. 113–118, 2017.

[15] J. Santisteban and J. Tejada-Carcamo, "Unilateral weighted Jaccard coefficient for NLP", *Proc. - 14th Mex. Int. Conf. Artif. Intell. Adv. Artif. Intell. MICAI 2015*, pp. 14–20, 2016.

[16] E. Gelbstein, "Quantifying Information Risk and Security", *ISACA J.*, vol. 4, pp. 1–6, 2013.

[17] OWASP, "OWASP Risk Rating Methodology", 2015. [Online]. Available: https://www.owasp.org/index.php/OWASP_Risk_Rating_Methodology.

[18] K. Song, J. Min, G. Lee, S. Chul Shin, and Y.S. Kim, "An Improvement of Plagiarized Area Detection System Using Jaccard Correlation Coefficient Distance Algorithm", *Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 76–80, 2015.