Indian Journal of Science and Technology, *Vol 12(17), DOI: 10.17485/ijst/2019/v12i17/143328, May 2019*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Apriori Algorithm Application on the Prevalence of Computer Malware

Devine Grace Doble Funcion*

IT Faculty, Leyte Normal University, B.Sc. Computer Science/MSc in Information Technology, Philippines; dgd_35@lnu.edu.ph

Abstract

Objective: The study aims to identify the characteristics, sources of computer malware. **Methods:** Data mining technique was explicitly utilized the Apriori algorithm to determine the characteristics, types,and sources of malware once it infiltrates the computer system. The processof gathering the data was through a survey questionnaire using Google form where there are two hundred five (205) IT students answered the survey form. **Findings:** The analysis shows that Apriori algorithm 98% accurately generates association rules on computer malware. Hence, computer malware can quickly spread through the use of flash drives and common malware that infects computer laboratory is a Virus. **Application/Improvements:** In the formulation of the computer laboratory policy, there must be clear policy on the use of flash drive inside the laboratory to avoid spreading computer malware that can damage the computer hardware/software. Additionally, a comprehensive study on the use of Apriori Algorithm is recommended.

Keywords: Apriori Algorithm, Computer Malware, Data Mining, Flash Drive, Knowledge Discovery in Database Process

1. Introduction

In this modern age where data grows in complexity and is rambling beyond the horizon. Tiny datasets are drifting around waiting to unravel essential information through data mining process. Data mining is a technique used to interpret large data and elucidate hidden patterns and information[1]. Further, data mining is in search of patterns and relationship in large databases[2]. Data miningin the medical field, banking firms, food, and the drug is widely used to ensure customers safety by using artificial intelligence analysis, usually applied to large-scale datasets[3]. Moreover, revealing hidden patterns and relationships uses functions in data mining such as (clustering, classification, prediction, and association). One essentialtask in data mining is that of the association rule. Association rule was first introduced in 1993, as a technique in data mining thatidentifies and extracts frequency patterns, association, correlations and relationships among sets of items in databases[2]. Example of association rule was used to analyzecustomer buying habits and behavior. The goal of the market –the basketwas to identify the customer frequently purchased products. If the customer bought ITEM A (apple, sandwich), then Item B(drinks) will most likely be purchased. This data can be used to organize and display the products, approximately close to each other and making it more accessible to the customer to purchase the product[4,5]. Also, association rule mining in microeconomics product selection based on the parameters that are frequently applied by retailers to endorse their product selection decision-making process[6]. Hence, results showed that the model was capable of identifying cross-selling effects implicitly by using frequent item sets, instead of having to calculate cross-selling parameters explicitly. An essential characteristic of association rule mining is that it separates the problem of mining into sub-problems to do efficient computing. One problem is finding frequent item sets from the database, and the other problem generates association rules from the database[7,8]. Hence, association rule using the Apriori

*Author for correspondence

algorithm was used to elicit significant information in the medical field specifically those patients with diabetic conditions. It is vital to note that the nature of medical information is categorized, quantitative or Boolean. Data mining with association rules as described is concerned merely getting meaning of Boolean data[9].

Moreover, association rule in mining can be used to identify computer malware that commonly destroys files and computer hardware. Computer malware isa computer program designed to damage computer system and application[10]. Computer malware comes in different forms, including spyware, ransomware, viruses, worms, Trojan horses, adware, or any malicious code that infiltrates a computer. According to Panda Security research, everyday there are about 230,00 new malware produced, and it is predicted to keep on increasing each year[11]. As technology advances, hackers use email to collect information to generate money using ransomware attack. However, according to the researcher from Erlangen-Nuremberg University, people are not cautious about the effect of opening unknown links send through email. About 78% of people still open unknown email or spam delivered by the unknown sender[12]. Furthermore, 81% of internet users become a victim of data breach. Internet users do not have a system that will self-detect data breach[13].

In the Philippines, the Department of Information and Communication Technology (DICT) formulated a National Cybersecurity Plan 20222 ensuring the safety of each Filipino people in using the Internet. The primary goals of this Plan are as follows: (1) tocontinuallyprovide the cooperation of public and military networks, (2) implement a quick response on cybersecurity threats during and after the attack, (3) efficiently coordination with law enforcement agencies and (4) to educated society about cybersecurity[14]. Access to the Internet is a fundamental element to improve the quality of education. Some State, College,and Universities (SUC's) have acquired internet services to provide better training. Instructors utilize the net to access online materials to supplement new learning, and students havea widerange ofaccess to learning not just at the library but also seeing a virtual library on the net.Interactive teaching methods, supported by the Internet, enable teachers to pay more attention to individual students' needs and support shared learning.

Moreover, the approach to the Internet helps educational administrators to scale down the monetary values and improve the caliber of schools and colleges[15]. Also, having internet access at school entails great responsibility in accessing information. Teachers and students should know the possible treat attached to the internet. Some email contains computer malware, once the virus penetrated the computer system it can affect the computer of other users[16]. However, educating users about computer malware is essentialto avoid becoming a victim of computer malware.

There are many association rule algorithms like Apriori Algorithm, Eclat Algorithm, and FP-growth Algorithm[17]. Moreover, the study utilizes the association rule specifically the apriori algorithm to determine and raise awareness to the students on how computer malware affects the computer and destroys the data and files stored by the student. Hence, this will likewise identify the characteristics, sources of computer malware.

2. Methodology

The study deploys the use of Knowledge Discovery in Databases (KDD). The KDD is a techniquethat will unveil hidden knowledge in large databases[18]. The study undergoes the process of KDD as shown in Figure 1.

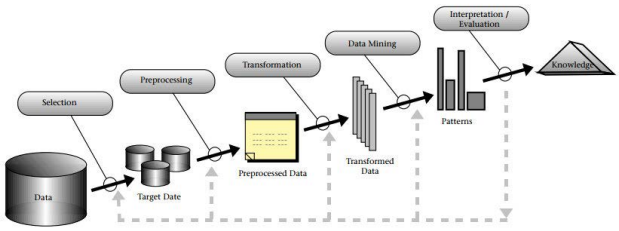

Figure 1. Steps in the KDD process.

2.1 Data Collection Stage

Table 1. Respondents of the study

Year Level	Total No. of Students	Actual Survey
First Year	200	132
Second Year	20	19
Third Year	46	32
Fourth Year	56	22
Total	322	205

The collection of data isthrough a survey questionnaire which was created using Google form. The research asked permission from the teacher in-change to conduct the study by giving the links generated in google form to

answer the survey questionnaire. A total of 205 out of322 respondents or 63.66% Information Technology students participated in the conduct of the study. Table 1 shows the total respondents per year level.

2.2 Preprocessing

In the preprocessing stage, it improves the data reliability by removing some of the attributes that are insignificant in the process of data mining presented in Table 2. Example, the *Age, Year Level,and Knowledge in computer virus*informationwas foundirrelevant in the process of identifying computer malware was discarded. Moreover, multiple answers in *Characteristics of Malware and Sources of Malware*was broken into individual responses and placed it in one column.

2.3 Transforming

In the transformation stage, the researcher carefully translated the answers of the respondents into codes. After, in Weka application, the ARFF file was loaded after the data cleaning of data found in Table 3 has been conducted.

2.4 Data Mining

The researcher utilized the Apriori Algorithm. Apriori is a seminal algorithm in finding patterns for frequent

Table 2. Survey Questionnaire questions

Parameter	Values	Type of Malware
Age	• 16-20 years old • 21-26 years old • 27-32 years old • 33 and above	It will identify the age group of IT students
Year Level	• First Year • Second Year • Third Year • Fourth Year	These are the year level of the respondents in the study
Knowledge in computer virus	• Highly knowledgeable • Moderately knowledgeable • Somewhat knowledgeable • Slightly knowledgeable • No idea at all	These will identify the level of expertise among the BSIT students in terms of computer malware
Types of malware	• Virus • Worm • Trojan Horse • Spyware • Phishing	These are the type of computer malware that can cause damage to the hardware and software component of the computer
Sources of malware	• Email • Computer Network • Flash Drives and other external media • Infected computer software	These are the possible sources of how computer malware will infect the computer system
Characteristics of malware	• Programs taking longer than usual to load and execute • Disk access taking longer than usual • Malfunctions of computer hardware • Disappearing files • Unusual file appearing • Pop-up ads keep on coming back • Increase in program file size	These are the characteristics of computerinfected malware

itemsets. The sorting of the itemset in Apriori transaction through lexicographic order[8].

Association rule mining:

Let **I**= { a_1, a_2, a_3, a_n….} be the attributes called items. Let **TD**= { Dt_1, Dt_2, Dt_3, Dt_n….} be set of transaction database. Each transaction in **DT** has a unique transaction ID and contains a subset of the items in **I**. A rule is defined as in every DT of records, X ⇒ Y means, a record of I contains X then I also contains Y[19]. The item set X and Y is called support and consequent of the rule respectively.

The support supp(X) has the rule of:

$$\text{supp}(\mathbf{X}) = \frac{\text{number of transactions which X appears}}{\text{total number of transactions}}$$

Consequent rule: (**X→Y**) = supp (**XUY**)/supp(**X**)

Steps in Apriori Algorithm

1. Start the item sets containing just a single item, such as {virus} and {flash drive}
2. From the transaction database, get the support S for each itemset, where S>=min_sup.
3. Using the item set generated in Step 1, get all possible patterns of item set
4. Repeat Step 1 and 2 until no new patterns or item sets generated[20].

Apriori Algorithm: Pseudocode

- Join Step: joined C_k and Lk-1 generatedwith itself
- Prune Step: Any (k-1)-item set that is not frequent cannot be a subset of a frequent
- k-itemset.

C_k: Candidate item set of size k
L_k: frequent item set of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) do begin
C_{k+1} = candidates generated from L_k;
for each transaction t in database do
increment the count of all candidates in C_{k+1}
that are contained in t
Lk + 1 = candidate in Ck + 1 with min_support
end
return $\cup_k L_k$;

3. Result and Discussion

Apriori
=======
Minimum support: 0.35 (71 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large item sets:

Size of a set of large item sets L(1): 8

Size of set of large itemsets L(2): 21

Size of set of large itemsets L(3): 10

Size of set of large item setsL(4): 1

Best rules found:

1. Virus=YES Unusual file appearing=YES 87 ==>Flash Drives=YES 85 <conf:(0.98)> lift:(1.06) lev:(0.02) [4] conv:(2.27)
2. Virus=YES Disappearing files=YES Unusual file appearing=YES 74 ==>Flash Drives=YES 72 <conf:(0.97)> lift:(1.06) lev:(0.02) [3] conv:(1.93)
3. Virus=YES Disappearing files=YES 106 ==>Flash Drives=YES 102 <conf:(0.96)> lift:(1.04) lev:(0.02) [4] conv:(1.66)
4. Virus=YES Malfunctions of computer hardware=YES 77 ==>Flash Drives=YES 73 <conf:(0.95)> lift:(1.03) lev:(0.01) [2] conv:(1.21)
5. Programs taks longer load and execute=YES Unusual file appearing=YES 76 ==>Flash Drives=YES 72 <conf:(0.95)> lift:(1.03) lev:(0.01) [1] conv:(1.19)
6. Disappearing files=YES 141 ==>Flash Drives=YES 133 <conf:(0.94)> lift:(1.02) lev:(0.01) [3] conv:(1.23)
7. Virus=YES Programs taks longer load and execute=YES 88 ==>Flash Drives=YES 83 <conf:(0.94)> lift:(1.02) lev:(0.01) [1] conv:(1.15)
8. Infected Software=YES Disappearing files=YES 83 ==>Flash Drives=YES 78 <conf:(0.94)> lift:(1.02) lev:(0.01) [1] conv:(1.08)
9. Programstaks longer load and execute=YES Disappearing files=YES 83 ==>Flash Drives=YES 78 <conf:(0.94)> lift:(1.02) lev:(0.01) [1] conv:(1.08)
10. Disappearing files=YES Unusual file appearing=YES 99 ==>Flash Drives=YES 93 <conf:(0.94)> lift:(1.02) lev:(0.01) [1] conv:(1.11)

Table 3. Data Coding

IVirus	Worm	Phishing	TrojanHorse	Spyware	Email	Computer Network	Flash Drives	infected software	Programs taks longer load and execute	Disk access taking longer than usual	Malfunctions of computer hardware	Disappearing files	Unusual file appearing	Pop-up ads keeponcomingback	Increase in program file size
YES					YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES					YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
YES							YES	YES	YES	YES	YES	YES	YES	YES	
		YES				YES	YES		YES		YES		YES		
YES							YES	YES			YES	YES	YES		

In the conduct of the study, the following assumptions were made. It was assumed the

1. Computer malware can be spread quicklythrough the use of flash drive with infected software.
2. That most infected computer with Virus show characteristic of Unusualfileappearing or Disappearing files.
3. That flash drive is the common carrier of computer malware.

4. Conclusion

Apriori algorithm is an association rule in mining with a 98% confidence level that the algorithm will be able to create association rules. It was found out in the result that using USB or Flash drives can cause spread the *Virus* in the laboratory that could affect the student files and even the hardware/software of the computer.

5. Recommendation

1. there must be a strict policy in the computer laboratory in term of connecting USB or flash drives in the computer.
2. There must be proper computer maintenance done to check the computer for possible malware in the computer laboratory.
3. A comprehensive study should be done to validate the accuracy of the Apriori algorithm effectively.
4. Another researcher may use the technique in Apriori to generate association rule in identifying the frequency of the item sets.

6. References

1. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016.
2. Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering; 2013. p. 194–9. https://doi.org/10.1109/ICPRIME.2013.6496471
3. Hurria A, Togawa K, Mohile SG, Owusu C, Klepin HD, Gross CP, Klapper S. Predicting chemotherapy toxicity in older adults with cancer: a prospective multicenter study. Journal of Clinical Oncology. 2011; 29(25):3457. https://doi.org/10.1200/JCO.2011.34.7625. PMid:21810685 PMCid:PMC3624700
4. Ingle MG, Suryavanshi NY. Association rule mining using improved Apriori algorithm. International Journal of Computer Applications. 2015; 112(4):1–6.
5. Olson DL, Delen D. Advanced data mining techniques. Springer Science and Business Media; 2008.
6. Brijs T, Swinnen G, Vanhoof K, Wets G. Using association rules for product assortment decisions: A case study. KDD. 1999; 99:254–60. https://doi.org/10.1145/312129.312241
7. Khan MM, Rajavat A. An efficient algorithm for extracting frequent item sets from a data set. Proceedings of the International Journal of Advanced Research in Computer Science and Software Engineering; 2013. p. 1373–5.

8. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, Zhou ZH. Top 10 algorithms in data mining. Knowledge and Information Systems. 2008; 14(1):1–37. https://doi.org/10.1007/s10115-007-0114-2
9. Stilou S, Bamidis PD, Maglaveras N, Pappas C. Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. Studies in health technology and informatics. 2001; (2):1399–403.
10. Malware [Internet]. [cited 2019]. Available from: https://us.norton.com/internetsecurity-malware.html.
11. 27% of all recorded malware appeared in 2015 [Internet]. [cited 2018]. Available from: https://www.pandasecurity.com/mediacenter/press-releases/all-recorded-malware-appeared-in-2015/.
12. Even this expert on hackers got tricked into clicking a scam email [Internet]. [cited 2016 Aug 09]. Available from: https://www.businessinsider.com/expert-phishing-emails-2016-8?IR=T.
13. Ten hard-hitting cyber security statistics [Internet]. [cited 2016 Feb 07]. Available from: https://swimlane.com/blog/10-hard-hitting-cyber-security-statistics/.
14. National Cyber security Plan 2022 [Internet]. [cited 2017 May 02]. Available from: http://dict.gov.ph/national-cyber-security-plan-2022/.
15. Internet Access and Education: Key considerations for policy makers [Internet]. [cited 2017 Nov 20]. Available from: https://www.internetsociety.org/resources/doc/2017/internet-access-and-education/.
16. Harrington SJ. Why people copy software and create computer viruses. Information Resources Management Journal. 1989; 2(3):28–38. https://doi.org/10.4018/irmj.1989070103
17. Pujari AK. Data mining techniques. Universities press; 2001. p. 288.
18. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996; 39(11):27–34. https://doi.org/10.1145/240455.240464
19. Zhan J, Matwin S, Chang L. Privacy-preserving collaborative association rule mining. IFIP Annual Conference on Data and Applications Security and Privacy; 2005. p. 153–65. https://doi.org/10.1007/11535706_12
20. Mining Frequent itemsets - Apriori Algorithm [Internet]. [cited 2018 Mar 15]. Available from: http://dwgeek.com/mining-frequent-itemsets-apriori-algorithm.html/.