Journal of Engineering and Applied Sciences 12 (21): 5394-5398, 2017
ISSN: 1816-949X

# A Survey on Association Rule Mining Approaches for Malicious Detection

Nawfal Turki Obeis and Wesam Bhaya
College of Information Technology, University of Babylon, Babil, Iraq

**Abstract:** The quality of malicious detector is determined by the technique it uses. The extracting of interest and useful knowledge from huge data called data mining. It uses with many aspects of clustering, classification, association rule mining, frequent pattern mining, etc. Association rule mining is a significant technique to finds interesting relationships among items in various datasets. Recently, association rule discovery has turned to important topics in data mining with malicious detection. It attracts extracares because of its varied usability. The association rule mining is normally worked by generating of frequent itemsets and rules in which many researchers provided many effective algorithms. To discover these rules, it needs to find frequent itemsets. Based on these frequent itemsets, it can build blocks of association rules with a given support and confidence factors. Here in this study, a survey on association rule algorithms will be present. At the beginning, we present the concepts of association rules and some of the related research works which done on it. Then, a discussion of the limitations and advantages of association rule algorithms will provide.

**Key words:** Association rule, malicious detection, frequent pattern mining, data mining, survey, itemsets

## INTRODUCTION

Security technology has become out to be significant in ensuring government-computing infrastructure. Modern malicious detection applications confronting complex issues. Malicious detection is an area developing in importance as an ever-increasing number of delicate information are stored and processed in computer or networked systems. An extensive Malicious Detection System (MDS) requires a lot of human mastery and time for advancement. Using MDS in data mining doesn't need a lot of expert experience and provide a good implementation. Malicious program are activities meant to bargain the privacy, integrity or potentially accessibility of a PC or PC organize. Malicious detection is the way toward checking and examining the events occurring in a PC system so as to distinguish signs of security issues (Kesarula *et al*., 2011).

In data mining technology, the association rules can be generally utilized as a part of MDS to get a matching between the normal behavior pattern and the abnormal pattern. In data mining, there are a lot of benefits among them, we can utilized one technique to a different data sources. The essential issue in malicious detection is how we can successfully and effectively separated the malicious patterns and ordinary patterns from a huge number of data and how feasibly creates malicious rules automatically after assembled raw data (Parekh *et al*., 2012).

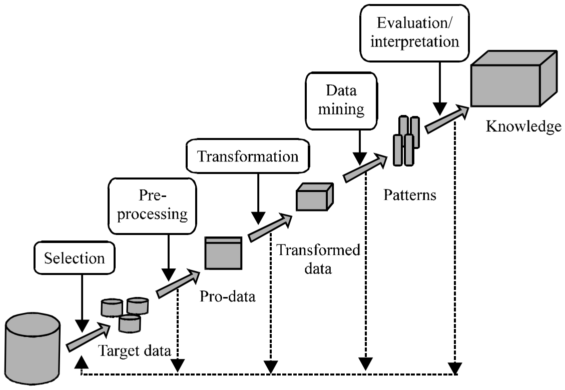

Fig. 1: The process of knowledge discovery in database

Interesting, supportive data can be withdraw from huge data. Professionals consider mining of data as the central method to discover the knowledge in database. Figure 1 illustrates method of knowledge discovery at database.

It is generally called extraction of data, data pattern analysis. The association and correlation among patterns in a huge sets of data or value-based will be disclosure during the process of mining frequent pattern (Al-Maqaleh and Shaab, 2013).

**Literature review:** Krishnan and Balasubramanian (2017) proposed method in association rule mining is an

**Corresponding Author:** Nawfal Turki Obeis, College of Information Technology, University of Babylon, Babil, Iraq

adaptive rule-based multiagent intrusion detection system to detect the anomalies in the real-time datasets (Krishnan and Balasubramanian, 2017).

Meng and Ren (2016) given a model to hybrid P2P networks for discovering outlier malicious node based on mining. In their model, the behavior patterns of node in a subnets will be used to discover the frequent patterns utilizing the method of mining frequent pattern at first and after that the global frequent patterns of nodes will be produce and update through gradually broadcasting and grouping the local frequent patterns (Meng and Ren, 2016).

Aung and Nyein (2015) provided traditional FP growth algorithm, one of the association algorithms is modified and used to mine itemsets from large database. The required statistics from large databases are gathered in to a smaller data structure (FP-tree) (Aung and Nyein, 2015).

Elhag *et al*. (2015) give a model to using genetic fuzzy systems and fuzzy association rule for enhanced detection rates on intrusion detection systems.

Usha and Rameshkumar (2014) clarify the ideas of frequent pattern mining and three critical methodologies which are: the generation of candidate method without candidate generation and vertical design method. Additionally, they explain how the several frequent pattern algorithms can be used in a various domains and specifically in the pattern detection of crime (Usha and Rameshkumar, 2014).

Bhavsar and Waghmare (2013) proposed a system to detect an intrusions using fuzzy class-association rule mining and support vector machine.

Hanguang and Yu (2012) give a model to using Apriori algorithm to classic of association rules in web-based on intrusion detection system and applies the rule base generated by the Apriori algorithm to identify a variety of attacks.

## MATERIALS AND METHODS

**Organization of datasets:** Dataset can be organized in two ways: vertically or in horizontally. For quite, a few years and especially before the advances in the relational database systems, the data is arranged in records horizontally and then processed in a vertical fashion. At the point when the data organization horizontal each exchange contains just things determinedly associated with a user services. When the layout is horizontal, the database is composed as rows from items with every line representing client's transaction to the extent the things that are obtained in the exchange. There is an alternative approach to manage this information organize for instance, vertical design. It includes each things related with a column of items value representing to the exchange in which it is accessible.

Table 1: Transactions dataset

| TID | List of items |
|---|---|
| 1 | A, B, C, E, G |
| 2 | B, E, F, G, H |
| 3 | A, B, C, D |
| 4 | A, B, C, D, E, F, G, H |
| 5 | E, F, G, H |
| 6 | D, E, F, G, H |
| 7 | E, F, G |

Table 2: Horizontal layout vertical layout

| | | List of items | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TID | List of items | A | B | C | D | E | F | G | H |
| 1 | A, B, C, E, G | 1 | 1 | 1 | 3 | 1 | 2 | 1 | 2 |
| 2 | B, E, F, G, H | 3 | 2 | 3 | 4 | 2 | 4 | 2 | 4 |
| 3 | A, B, C, D | 4 | 3 | 4 | 6 | 4 | 5 | 4 | 5 |
| 4 | A, B, C, D, E, F, G, H | | | | | 5 | 6 | 5 | 6 |
| 5 | E, F, G, H | | | | | 6 | 7 | 6 | |
| 6 | D, E, F, G, H | | | | | 7 | | 7 | |
| 7 | E, F, G | | | | | | | | |

Table 3: Comparative analysis of the different existing frequent patterns mining approaches

| Approaches | Search direction | Data structure/layout | First researchers name | Years of publications |
|---|---|---|---|---|
| Apriori | BFS | Hash tree (Horizontal) | Agrawal | 1993 |
| Viper | BFS | Vertical | Shenoy | 2000 |
| Eclat | DFS | Vertical | Zaki | 2000 |
| FP-growth | DFS | Prefix tree (Horizontal) | Han | 2000 |
| Mafia | BFS | Vertical | Burdick | 2001 |
| PP-mine | DFS | Prefix tree (Horizontal) | YabuXu | 2002 |
| COFI | DFS | Prefix tree (Horizontal) | El-Hajj | 2003 |
| Diffset | BFS | Vertical | Zaki | 2003 |
| TM | DFS | Vertical | Song | 2006 |
| TFP | Hybrid | Prefix tree (Horizontal) | Show-Jane Yen | 2009 |
| SSR | DFS | Horizontal | Show-Jane Yen | 2012 |

There are some database measures, lower space of the dataset and better support of dynamic dataset. Recently, many studies have been conducted on the organization of data to see how the performance of disclosure frequent pattern can be effected. The revelation of interesting connections concealed in enormous datasets is the target of the data mining in frequent pattern. Tables 1-3 illustrate examples of data organization (Mashoria and Singh, 2013).

## RESULTS AND DISCUSSION

**Association rule mining:** The way toward extricating or mining knowledge from a huge dataset is called data mining. This procedure first comprehends the current data and after that predicts the new data. Usually, descriptive and predictive represent the main categories into which

Table 4: Comparison of different frequent patterns mining algorithms for malicious detection

| Algorithm | Advantages | Disadvantages |
|---|---|---|
| Apriori (Agrawal and Srikant, 1994) | The main advantage is that k items candidate itemsets can be produced by joining extensive itemsets having k-1 items and expelling those that contain any subset that is not expansive which infers that the itemsets are not frequent in that subset. This outcomes in the creation of less number of candidate sets | It is especially expensive to deal with an extensive number of candidate sets. It is monotonous to examine the database recursively and check a colossal arrangement of candidates by pattern matching |
| Hash based technique (Agrawal *et al*., 1996) | The number of itemsets decreased as compared to the Apriori algorithm. Less execution time in comparison with Apriori | A little higher cost in the main cycle because of the creation of hash table |
| Transaction reduction (Agrawal *et al*.. 1993) | Decrease the quantity of transactions to be examined later on cycles | Does not consider the itemsets as the priority |
| Partitioning (Savasere *et al*., 1995) | Reduction in I/O overheadas well as in CPU overhead in comparison with previous algorithms | Partition size and the quantity of partitions can't be taken extensive. The size to be considered, so that every segment can fit into main memory and in this manner be read only once in every stage |
| Sampling (Toivonen, 1996) | Advantageous when the proficiency is of much significance | Potentially is that a portion of the global frequent itemsets can be omitted. Accuracy might not be metdue to the consideration of sample instead |
| Dynamic itemset counting (Brin *et al*., 1997) | Fewer database scans when compared to the traditional approaches for finding all the frequent itemsets. Item reordering conceptis added that improves thelow level efficiency of the algorithm | The process of candidate set creation will be costly when there exists long patterns |
| Frequent pattern growth (Han *et al*., 2000) | Avoids costly, repeated database scans. Precludes the expensive generation of a huge number of candidate sets | When the database is huge, it is some of the time unreasonable to develop a main memory based FP-tree |
| Parallel frequent pattern (Li *et al*., 2008) | Segmentation applied in this algorithm removes computational dependencies amongst machines and hence, forth the communication between them. Computational time is linear | PFP doesn't take into account load adjust which is very imperative for extensive scale data processing |
| Balanced parallel frequent pattern (Zhou *et al*., 2010) | Because of the load balancing feature this improves parallelization and hence, improves performance | The precision for balancedgrouping strategy is not taken into consideration |

the tasks of data mining can be grouped. The general properties are described of the data in the dataset are portrayed by descriptive mining (Obeis and Wesam, 2016). Predictive mining performs inference on the present data keeping in mind the end goal to make expectations of association rule mining. Particularly, two data mining methods have been using for malicious detection: frequent pattern and association rules. Association rule approaches find the correlations between attributes or features that used to describe a dataset. Association rules mining began as a procedure for finding fascinating rules from transactional datasets. The formal of the problem can definition as: let "I" = $\{i_1, i_2, i_3,..., i_n\}$ is a set of literals, it's called items. Let database, "D" is a group of transaction records in which every transaction T represent a group of items such that $T \subseteq I$. Every transaction is related with a one of a kind identifier it is called Transaction ID (TID). The transaction T contains A a set of items in I, if $A \subseteq I$. An association rule can be representation of the form $A \rightarrow B$ where $A \subseteq I$, $B \subseteq I$ and $A \cap B = \emptyset$. The rule $A \rightarrow B$ has support s in the transaction set D if s of transactions in D contain AUB:

$$\text{Support}(AB) = \frac{\text{Support sum of AB}}{\text{Overall records in the database D}} \quad (1)$$

The rule $A \rightarrow B$ holds in the transaction set D with confidence c if c of transactions in D that contain A also contain B:

$$\text{Confidence (A/B)} = \frac{\text{Support}(AB)}{\text{Support}(A)} \quad (2)$$

Suppose a set of the items as $I = \{I_1, I_2, I_{3,}, \ldots, I_m\}$ and the database of transactions as $D = \{t_1, t_2, t_3, \ldots, t_n\}$ if $t_i = \{I_{i1}, I_{i2}, I_{i3}, \ldots, I_{ik}\}$. The problem of association rule is to identify all association rules as $A \rightarrow B$ with a confidence and minimum support. The percentage of transactions that contains both A and B in all transactions is called the support of the rule and is calculated as |AB|/|D|. The rule measures support the significance of the correlation between itemsets in database. The confidence is defined as the percentage of transactions that include B in the transactions that include A. The rule measures confidence the degree of correlation between the itemsets in database and is calculated as |AB|/|A|. The support measure is the measure of a rule frequency and the confidence measure represents the measure of the strength of the relation between the groups of items (Dhanabhakyam and Punithavalli, 2011; Aung and Nyein, 2015).

**Frequent patterns mining:** The professionals reveal of frequent pattern from huge data using data mining represent an active research area. Frequent pattern mining is a topic for some core research in data mining for previous years. Agrawal *et al*. (1993) proposed a frequent pattern mining for analysis of market basket as association rule mining. Frequent itemsets expect a fundamental part in various data mining actions that try to find interesting patterns from databases for example, association rules, classifiers, clusters and numerous a greater amounts of which the mining of association rules is a standout among the most well-known issues.The main motivation for looking association rules started from the need to investigate supermarket exchange data to survey client conduct as terms of the bought items. Frequent patterns are itemsets or substructures that exist in a dataset with recurrence that no not as much as a client indicated threshold (Agrawal and Srikant, 1994).

The main objective of frequent patterns mining is to discover the frequently happening items in a substantial database. Frequent Patterns (FPs) are itemsets, substructures or subsequences showing up in a dataset with frequency. They can be classified in two methods: pattern growth algorithms and candidate generation algorithms. This could be classified into a few structures and they are as Data Structures and Traversal Strategy, i.e., Depth First Strategy (DFS) or Breadth First Strategy (BFS). Table 3 shows the brief history of the different Frequent Pattern algorithms in vertical and horizontal data layouts on the literature survey (Agrawal *et al*., 1993).

**Frequent patterns algorithms for malicious detection:** Asin literature survey, researchers use distinctive techniques to detect malicious. Table 4 explains the pros and cons of different algorithms.

## CONCLUSION

Malicious detection technology is a powerful way to deal with the issues of network security. In this study, we exhibit the literature survey on various algorithms belonging to utilizations of association rule mining and frequent patterns mining with regards to malicious detection. Association rule mining and frequent patterns mining give knowledge about various frequent pattern mining algorithm and association rule mining. This study can encourage the researchers about to get information and uncovers the benefits of applying frequent pattern mining algorithm alongside association rule mining in malicious recognition.

## REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Databases, September 12-15, 1994, Morgan Kaufmann Publishers, Santiago, Chile, pp: 487-499.

Agrawal, R., M. Mehta, J.C. Shafer, R. Srikant and A. Arning *et al*., 1996. The quest data mining system. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining Vol. 96, August 02-04, 1996, ACM, Portland, Oregon, pp: 244-249.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Acm. SIGMOD. Rec., 22: 207-216.

Al-Maqaleh, B.M. and S.K. Shaab, 2013. An efficient algorithm for mining association rules using confident frequent itemsets. Proceedings of the 3rd International Conference on Advanced Computing and Communication Technologies (ACCT) 2013, April 6-7, 2013, IEEE, Rohtak, India, ISBN:978-1-4673-5965-8, pp: 90-94.

Aung, K.M.M. and N.O. Nyein, 2015. Association rule pattern mining approaches network anomaly detection. Proceedings of 2015 International Conference on Future Computational Technologies (ICFCT 2015), March 29-30, 2015, Conal, Singapore, ISBN:978-93-84468-20-0, pp: 164-170.

Bhavsar, Y.B. and K.C. Waghmare, 2013. Intrusion detection system using data mining technique: Support vector machine. Intl. J. Emerging Technol. Adv. Eng., 3: 581-586.

Brin, S., R. Motwani, J.D. Ullman and S. Tsur, 1997. Dynamic itemset counting and implication rules for market basket data. Proc. 1997 ACM SIGMOID Int. Conf. Manage. Data, 26: 255-264.

Dhanabhakyam, M. and M. Punithavalli, 2011. A survey on data mining algorithm for market basket analysis. Global J. Comput. Sci. Technol., Vol. 11, 10.17406/gjcst

Elhag, S., A. Fernandez, A. Bawakid, S. Alshomrani and F. Herrera, 2015. On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems. Expert Syst. Applic., 42: 193-202.

Han, J., J. Pei and Y. Yin, 2000. Mining frequent patterns without candidate generation. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 1-12.

Hanguang, L. and N. Yu, 2012. Intrusion detection technology research based on apriori algorithm. Phys. Procedia, 24: 1615-1620.

Kesavulu, R.E., V.N. Reddy and P.G. Rajulu, 2011. A study of intrusion detection in data mining. Proceedings of the World Congress on Engineering (WCE 2011) Vol. 3, July 6-8, 2011, Sri Venkateswara University, London, UK., ISBN: 978-988-19251-5-2, pp: 1-6.

Krishnan, S.D. and K. Balasubramanian, 2017. A fusion of multiagent functionalities for effective intrusion detection system. Secur. Commun. Netw., 2017: 1-15.

Li, H., Y. Wang, D. Zhang, M. Zhang and E. Chang, 2008. Pfp: Parallel fp-growth for query recommendation. Proceedings of the ACM Conference on Recommender Systems, October 23-25, 2008, Lausanne, Switzerland, pp: 107-114.

Mashoria, V. and A. Singh, 2013. Literature survey on various frequent pattern mining algorithm. IOSR. J. Eng., 3: 58-64.

Meng, X. and S. Ren, 2016. An outlier mining-based malicious node detection model for hybrid P2P networks. Comput. Netw., 108: 29-39.

Obeis, N.T. and B. Wesam, 2016. Review of data mining techniques formalicious detection. Res. J. Appli. Sci., 11: 942-947.

Parekh, S.P., B.S. Madan and R.M. Tugnayat, 2012. Approach for intrusion detection system using data mining. J. Data Min. Knowl. Discovery, 3: 83-87.

Savasere, A., E. Omieccinski and S. Navathe, 1995. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very Large Databases, September 11-15, 1995, Zurich, Switzerland, pp: 432-443.

Toivonen, H., 1996. Sampling large databases for association rules. Proceedings of the 22th International Conference on Very Large Databases, Septempber 3-6, 1996, Bombay, India, pp: 134-145.

Usha, D. and K. Rameshkumar, 2014. A complete survey on application of frequent pattern mining and association rule mining on crime pattern mining. Intl. J. Adv. Comput. Sci. Technol., 3: 264-275.

Zhou, L., Z. Zhong, J. Chang, J. Li and J.Z. Huang *et al*., 2010. Balanced parallel fp-growth with mapreduce. Proceedings of the 2010 IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT), November 28-30, 2010, IEEE, Beijing, China, ISBN:978-1-4244-8883-4, pp: 243-246.