



AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Database Intrusion Detection Using Sequential Data Mining Approaches

A Thesis

Submitted in partial fulfillment of the requirements of the degree of
Master of Science in Electrical Engineering

Submitted by

Pakinam ElAmein Abd ElAziz Hussein

B.Sc. of Electrical Engineering
(Computer and Systems Engineering)
Ain Shams University, 2008

Supervised By

Prof. Dr. Hoda Korashy Mohamed
Dr. Mohamed Ali Ali Sobh

Cairo, 2015

Examiners' Committee

Name: **Pakinam ElAmein Abd ELAziz Hussein**
Thesis: Database Intrusion Detection Using Sequential Data Mining Approaches
Degree: Master of Science in Electrical Engineering (Computer and Systems Engineering)

Title, Name and Affiliation

Signature

Prof. Dr. Fatma Abd Elstar Hassan Omara
Faculty of Computer and Information,
Cairo University.

.....

Prof. Dr. Mohamed Watheq El Kharashy
Computer and Systems Engineering Dept.
Faculty of Engineering,
Ain Shams University.

.....

Prof. Dr. Hoda Korashy Mohamed
Computer and Systems Engineering Dept.
Faculty of Engineering,
Ain Shams University, Cairo (Supervisor)

.....

Date:

STATEMENT

This dissertation is submitted to Ain Shams University for the degree of Master of Science in Electrical Engineering (Computer and Systems Engineering).

The work included in this thesis was carried out by the author at the Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt.

No part of this thesis was submitted for a degree or a qualification at any other university or institution.

Name: Pakinam Elamein Abd ElAziz Hussein

Signature:

Date: April, 2015

Curriculum Vitae

Name of Researcher	Pakinam ElAmein Abd ElAziz Hussein
Date of Birth	6/9/1985
Place of Birth	Saudi Arabia
First University Degree	B.Sc. in Electrical Engineering (Computer and Systems Department)
Name of University	Ain Shams University
Date of Degree	June 2008

ABSTRACT

*Pakinam ElAmein Abd ElAziz Hussein,
Database Intrusion Detection Using Sequential Data Mining
Approaches,
Master of Science dissertation, Ain Shams University, 2015.*

The awareness of the normal behaviors and practices with a transactional database is an essential factor to have an efficient detection of the data violations. This awareness enables an easier identification of any new behavior for the transactions. Data Mining plays an important role in providing this awareness. The main role of the data mining is to study and analyze raw data to generate useful information. In other words, Data Mining can be used to study database transactions and generate behaviors.

This thesis mainly demonstrates how the sequential data mining algorithm can be enhanced and used in identifying database intrusions. It starts with providing a background about the fields studied close to be able to provide this research. Afterwards, the selection of a famous the sequential data mining algorithm called Apriori Algorithm that fit the criteria of the research. There are three implementations of this algorithm that share the same idea but differ in the implementation. The three versions of the algorithm were implemented and studied closely. The main purpose of the study is to identify how they can be improved. The close study resulted in identifying key points that can be enhanced. The

study directed the research in trying to improve the algorithm in two directions which are the performance and accuracy of generated patterns.

On this thesis, an enhancement has been introduced to the performance and accuracy of Apriori algorithms. The research proposed some modifications on the Apriori model to improve it. On this research, the modifications were applied on the three different versions of the algorithm and studied carefully. There are significant improvements but also there are some trade-offs. The results of the implementations are discussed in details. Also, it is clarified where the enhancements lies. As for the trade-offs, there are some suggestions to overcome them.

The thesis also explains how the enhanced form of apriori algorithm can be used in intrusion detections. A simple Database Intrusion Detection Model is proposed. This model is formed of three phases. It is based on the enhanced form of the apriori Algorithm introduced on the thesis. The model also treats some of the trade-offs discussed after the enhancement.

Key words: Sequential data mining, apriori algorithms, intrusion detection, predictive data mining, descriptive data mining, host-based intrusion detection, network-based intrusion detection, computer security, malicious transaction, misuse detection, anomaly detection.

SUMMARY

This dissertation demonstrates how sequential data mining algorithms can be enhanced. It also demonstrates how they can be used in detecting intrusion in the database. The dissertation is in eight chapters organized as follows:

Chapter One: It begins with an introduction on the main fields covered on the thesis which are Data Mining and Intrusion Detection. First, it starts with defining the data mining and its importance. Afterwards, it provides an introduction to the intrusion and intrusion detection. Also, the chapter provides a summary of the efforts done in the area data intrusion detection. Furthermore, the chapter provides a summary of the work done on the thesis. Finally, it provides a summary of the rest of the thesis is organized.

Chapter Two: In this chapter, a summarized background is provided on all the fields and topics that the research interacts with. The main topics that are covered on the chapter are database, intrusion, intrusion detection, data mining and sequential data mining.

Chapter Three: This chapter discusses the algorithm that was selected for study by the thesis, Apriori algorithm. The chapter begins with explaining why the algorithm is selected by the thesis. The chapter mainly explains the generic model of the algorithm and how it works. Finally it discusses three models or versions of Apriori algorithm. They share the same idea but differ in the implementation mechanism.

Chapter Four: In this chapter, implementations of the three versions of the apriori algorithm are presented. A study has been performed of the results of these implementations to determine how they can be improved.

Chapter Five: This chapter explains the modifications proposed by the thesis on the algorithm for enhancement. It explains what are the modification actions, where and how to apply them. It also explain where the enhancement exactly is expected.

Chapter Six: In this chapter, implementations of the three versions of the apriori algorithm (after applying the modifications proposed) are presented. A study has been performed of the results of these implementations and how the performance has improved. Also the results are studied closely to see if there is any trade-offs.

Chapter Seven: In this chapter, the results of implementations before and after applying the modifications are presented. A comparison between the results is presented. Also the points of improvements are discussed as well as the trade-offs.

Chapter Eight: Finally, this chapter includes the conclusions extracted from the research and the research results. It also includes the future work that might be done based on this work.

.

ACKNOWLEDGEMENT

First, I would like to thank ALLAH for his great support to me in accomplishing this work.

I would like to express my gratitude to Prof. Dr. Hoda Korashy Mohamed for her encouragement to me on continuing this work.

I would like to express my gratitude to Dr. Mohamed Ali Sobh for his leading effort in the development of different parts of this work from the technical development to the documentation work.

I would like to express my deepest gratitude for my dear mother for her great support and encouragement. Without her this work would have never developed or come to real. I also want to thank her for saving my time even if this was at the expense of her comfort.

I would like to thank my father for everything. Without his efforts I would have never became what I am today.

At last but not least, I would like to thank my brother for his help and support during our father's health condition.

CONTENTS

LIST OF FIGURES.....	XVII
LIST OF TABLES.....	XXI
LIST OF ABBREVIATIONS	XXII
LIST OF PUBLICATIONS	XXIII
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: BACKGROUND	5
2.1 INTRODUCTION TO DATABASE	5
2.1.1 Database Users.....	6
2.1.2 Database Transaction logs	6
2.1.3 Vulnerabilities within Database.....	8
2.2 INTRUSION.....	10
2.2.1 Intrusion Definition:	10
2.2.2 Types of intrusion and intruders	10
2.3 INTRUSION DETECTION	15
2.3.1 Intrusion Detection Systems.....	16
2.3.2 Database Intrusion Detection.....	17
1.4 DATA MINING	20
2.3.3 Data Mining Tasks.....	20
2.3.4 Data Mining Approaches History in DB IDS	20
2.4 SEQUENTIAL DATA MINING	22
CHAPTER 3: APRIORI ALGORITHMS	25
3.1 APRIORI BASED ALGORITHMS.....	25
3.1.1 Definitions.....	25
3.1.2 Apriori Algorithms Principles.....	26

3.1.3	<i>Generic Model of Apriori algorithm</i>	28
3.2	APRIORIALL ALGORITHM	31
3.2.1	<i>AprioriAll Algorithm</i>	31
3.3	APRIORISOME	33
3.3.1	<i>Apriorisome Algorithm</i>	34
3.4	DYNAMICSOME ALGORITHM	35
3.4.1	<i>Dynamicsome Algorithm</i>	36
CHAPTER 4: IMPLEMENTATION OF APRIORI ALGORITHMS		38
4.1	IMPLEMENTATION MODEL DESCRIPTION.....	38
4.2	DATABASE TRANSACTION LOG USED.....	43
4.3	IMPLEMENTATION MODEL VS ALGORITHM TYPE	45
4.3.1	<i>Implementation Principles</i>	46
4.3.2	<i>Implementation of Aprioriall Algorithm</i>	46
4.3.3	<i>Implementation of Apriorisome Algorithm</i>	46
4.3.4	<i>Implementation of Dynamicsome Algorithm</i>	47
4.4	IMPLEMENTATION ANALYSIS OF ORIGINAL ALGORITHMS.....	48
4.4.1	<i>Analysis of The Implementation Performance</i>	48
CHAPTER 5 THE PROPOSED ENHANCEMENTS		52
5.1	ANALYSIS OF ORIGINAL ALGORITHMS	53
5.1.1	<i>Performance Analysis</i>	53
5.1.2	<i>Analysis of The Algorithm's Accuracy</i>	54
5.2	LEVELS OF ENHANCEMENTS	55
5.2.1	<i>Level 1: Performance Enhancement</i>	55
5.2.2	<i>Level 2: Enhancing Quality of The Output</i>	56
5.3	MODIFICATIONS PROPOSED FOR ENHANCEMENT	56
5.3.1	<i>First Modification: Transaction sorting</i>	56
5.3.2	<i>Second Modification: Modifying the pattern generation for Candidate set (C_k)</i>	58
5.4	ENHANCED APRIORI ALGORITHM.....	59

CHAPTER 6: IMPLEMENTATION OF PROPOSED ENHANCEMENTS ...	61
6.1 ENHANCEMENTS ON THE IMPLEMENTATION MODEL	61
6.1.1 <i>Enhancement on Performance Level</i>	61
6.1.2 <i>Enhancement on The Accuracy Level</i>	62
6.2 PROPOSED INTRUSION DETECTION MODEL	64
6.2.1 <i>Initialization Stage</i>	66
6.2.2 <i>Periodic Follow-up and Detection Stage</i>	66
6.2.3 <i>Action Stage</i>	67
CHAPTER 7: RESULTS	69
7.1 IMPLEMENTATION RESULTS OF ORIGINAL ALGORITHMS	69
7.2 POINTS OF STRENGTH AND DEFECTS	71
7.3 IMPLEMENTATION RESULTS OF ENHANCED ALGORITHM	72
7.4 POINTS OF STRENGTH AND DEFECTS	74
CHAPTER 8: CONCLUSION AND FUTURE WORK.....	77
8.1 CONCLUSION	77
8.2 FUTURE WORK	78
REFERENCES	79
APPENDIX A.....	83
A.1 APRIORIAL SOURCE CODES.....	83
A.2 APRIORISOME SOURCE CODES	85
A.3 DYNAMICSOME SOURCE CODES.....	88
A.4 SAMPLES OF MODIFICATIONS.....	91
A.5 SOURCE CODES FOR OTHER FUNCTIONS IN USE	92
ملخص	94
مستخلص الرسالة.....	96

List of Figures

<i>Figure 2-1: Transaction Log Record.....</i>	<i>7</i>
<i>Figure 2-2: IP Address Spoofing</i>	<i>11</i>
<i>Figure 2-3: Denial-of-Service Attack</i>	<i>12</i>
<i>Figure 2-4: Man-in-the-Middle Attack</i>	<i>13</i>
<i>Figure 2-5: Sniffer Attack.....</i>	<i>13</i>
<i>Figure 3-1: Flow of apriori algorithm.....</i>	<i>27</i>
<i>Figure 3-2: Apriori Generic Model</i>	<i>28</i>
<i>Figure 3-3: Apriori Algorithm steps</i>	<i>30</i>
<i>Figure 3-4: Example on Apriori algorithm</i>	<i>30</i>
<i>Figure 3-5: Aprioriall Example</i>	<i>32</i>
<i>Figure 3-6: Apriorisome Example</i>	<i>33</i>
<i>Figure 3-7: Dynamicsome Example</i>	<i>35</i>
<i>Figure 4-1: DB structure of Model.....</i>	<i>39</i>
<i>Figure 4-2-: Apriori Implementation model</i>	<i>40</i>
<i>Figure 4-3: Implementation Model Fn</i>	<i>40</i>
<i>Figure 4-4: Apriori & AprioriGen Model</i>	<i>41</i>
<i>Figure 4-5: ApriorGen fn Mechanism</i>	<i>42</i>
<i>Figure 4-6: Sample of implementation output</i>	<i>42</i>
<i>Figure 4-7: Sample of the transaction log</i>	<i>43</i>
<i>Figure 4-8: Apriorisome Phases</i>	<i>47</i>
<i>Figure 4-9: Dynamicsome Phases</i>	<i>48</i>
<i>Figure 4-10: Analysis recorded data set</i>	<i>49</i>
<i>Figure 4-11: Time Taken vs support value</i>	<i>50</i>
<i>Figure 5-1: Sample of operations per one Tx</i>	<i>57</i>
<i>Figure 5-2: Example on sorting transactions</i>	<i>58</i>
<i>Figure 5-3: Example on Second Modification.....</i>	<i>59</i>
<i>Figure 5-4: Flow of Enhanced Algorithm</i>	<i>60</i>
<i>Figure 6-1: Model Enhancement.....</i>	<i>62</i>

<i>Figure 6-2: AprioriGen Pattern Generation</i>	63
<i>Figure 6-3: Model Enhancement 2</i>	63
<i>Figure 6-4:Proposed detection model</i>	65
<i>Figure 6-5: Initialization stage</i>	66
<i>Figure 6-6: Periodic follow-up & detection</i>	67
<i>Figure 6-7: Flow chart for action stage</i>	68
<i>Figure 7-1: Implementation Results 1</i>	70,71
<i>Figure 7-2: Implementation results 2 (enhanced algorithm)</i>	73,74

.

List of Tables

Table 4-1. Fields with in transaction log.....44-45

Table 4-2: Illustrative Example.....47

Table 4-3: Illustrative Example.....48

Table 4-4: Frequent Pattern Analysis.....51

Table 5-1: Frequent Pattern Analysis.....54

Table 5-2: Modifying pattern generation..... 58

Table 7-1:Accuracy results.....74