

**APPLYING ASSOCIATION RULE MINING TO DETERMINE LOSSES
OCCURRENCES ON SISTEM OPERASI TERPADU (SOT) DATA**

By

Frieda Putri Aryani
2-2012-205

A thesis submitted to the Faculty of
ENGINEERING AND INFORMATION TECHNOLOGY

in partial fulfillment of the requirements
for the
MASTER'S DEGREE

in

INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY



SWISS GERMAN UNIVERSITY
EduTown BSD City
Tangerang 15339
Indonesia

January 2014

Revision after the Thesis Defense on February 18th 2014

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Frieda Putri Aryani

Student

Date

Approved by :

Dr. Ir. Moh. A. Amin Soetomo, M.Sc.

Thesis Advisor

Date

Alva Erwin, S.Kom., M.Sc

Thesis Co-Advisor

Date

Dr. Harya Widiputra, S.Kom.,M.Kom.

Thesis Co-Advisor

Date

Dr. Ir. Gembong Baskoro, M.Sc.

Dean of Engineering and Information Technology
Faculty

Date

ABSTRACT

APPLYING ASSOCIATION RULE MINING TO DETERMINE LOSSES OCCURRENCES ON SISTEM OPERASI TERPADU (SOT) DATA

By

Frieda Putri Aryani

Dr. Ir. Moh. A. Amin Soetomo, M.Sc., Advisor

Alva Erwin, S.Kom., M.Sc., Co-Advisor

Dr. Harya Widiputra, S.Kom., M.Kom., Co-Advisor

SWISS GERMAN UNIVERISTY

SKK MIGAS as a supervisor to control the activities of upstream oil and natural gas made by PSC (Production Sharing Contractor) need some method to perform data analysis from Sistem Operasi Terpadu (SOT) data to discover knowledge. Based on that, this research will focus on analyzing loss production opportunity of SOT data using Association Rule Mining (ARM) Techniques. ARM is one of the most significant and well researched techniques of data mining for discovering interesting correlation or association among sets of data to maximizing the knowledge discovered in SOT data. This research utilizes Rapid Miner as data mining open source software to produce ARM rules. However, the ARM result produces thousands of rules that are redundant. The post processing analysis needs to be done to reduce redundant rules and to sort the rules by their priority. Since production loss can be determined as a risk, then risk probability and impact matrix analysis used to prioritize rules. The objective interestingness measure also used by using support and confidence value to sort rules. The final results are encouraging and also produced valuable information to identify sets of loss that cause highest impact on financial value for every occurrence of this set.

Keywords : Association Rule Mining, Loss Production Opportunity, Objective Interestingness Measure, Rapid Miner, and Risk Matrix.



DEDICATION

I dedicate this thesis to my family, my parents and faculty of engineering and information technology in SGU (Swiss German University), Tangerang - Indonesia and Kelompok Kerja Pengolahan Data & Informasi Sub Kelompok Kerja Pengelolaan SKK MIGAS.



ACKNOWLEDGMENTS

The author wishes to acknowledge with gratitude to Allah SWT for the abundance of grace and gift, so the author can finish this work without experiencing significant obstacles.

The author also wishes to express most gratitude to my husband Ryan Bangun Raharja, my son Daffa Kazuo Arkana and my daughter Dena Keisha Azzahra. The author thanks them for all encouragement, the strength, the patience and the supports they had given that enabled the completion of this work.

The author thanks to Dr. Ir. Moh. A. Amin Soetomo, M.Sc. for acting as the advisor. Also thanks to Alva Erwin, M.Sc for acting as co-advisor, Dr. Harya Damar Widiputra, for acting as co-advisor. They have provided guidance, suggestions, constructive feedbacks, and supports that helped to shape this work.

Thanks to Kelompok Kerja Pengolahan Data & Informasi Sub Kelompok Kerja Pengelolaan SKK MIGAS for supporting to complete this research.

Last but not least, the author also would like to thank all of MIT students and friends, special mention to Bobby Suryajaya for his knowledge sharing and act as our Subject Matter Expert, Uma Bala Devarakonda for her great comments and discussion, and also those great experts who published their works on the internet, for their support in providing information and references that helped in the completion of this work.

TABLE OF CONTENTS

STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES.....	10
LIST OF TABLES.....	11
CHAPTER 1 - INTRODUCTION.....	12
1.1 Background.....	12
1.2 General Statement of Problem Area.....	13
1.3 Research Problem.....	18
1.4 Research Limitations.....	19
1.5 Research Questions.....	19
1.6 Research Objectives.....	19
1.7 Significance of Research.....	19
CHAPTER 2 - LITERATURE REVIEW.....	20
2.1 Data Mining.....	20
2.2 Data Mining Techniques.....	21
2.3 Association Rule Mining.....	24
2.3.1 Apriori Algorithm.....	26
2.3.2 FP Growth Algorithm.....	26
2.3.3 Comparison of Algorithm.....	28
2.4 Research on Association Rule Mining.....	29
2.5 Evaluation of Association Rule Patterns.....	30
2.5.1 Objective Interestingness Measure.....	30

2.5.2	Subjective Arguments.....	32
2.6	Sistem Operasi Terpadu (SOT) Data	33
2.7	Data Mining Methodology	36
2.7.1	SEMMA.....	36
2.7.2	CRISP-DM	37
2.8	Theoretical Framework	40
CHAPTER 3 - METHODOLOGY		42
3.1	Business Understanding	42
3.2	Literature Review	44
3.3	Data Collection.....	45
3.4	Data Preparation.....	46
3.5	Modeling	46
3.6	Analysis.....	47
3.7	Conclusion.....	47
CHAPTER 4 - RESEARCH DESIGN & EXPERIMENT		49
4.1.	Research Groundwork.....	49
4.1.1.	Environment Setting	49
4.1.2.	The Tools – Rapid miner.....	49
4.2.	Data Collection.....	50
4.2.1.	Data Attributes.....	51
4.3.	Data Preparation.....	53
4.4.	Data Modeling.....	54
CHAPTER 5 - RESULT & ANALYSIS		59
5.1.	Existing Data Analysis	59
5.2.	Threshold Changing	61
5.2.1.	Threshold-1.....	61
5.2.2.	Threshold-2.....	62

5.2.3. Threshold-3.....	63
5.3. Post Processing Analysis.....	63
5.3.1. Probability Determination	64
5.3.2. Impact Determination	64
5.3.3. Risk Determination.....	66
5.3.4. Rules Selection	67
5.4. Final Rules Analysis.....	77
CHAPTER 6 - CONCLUSION & RECOMMENDATION.....	80
6.1. Conclusion.....	80
6.2. Recommendation.....	80
6.3. Future Work	81
GLOSSARY	82
REFERENCES	83
APPENDIX.....	86
CURRICULUM VITAE.....	92



SWISS GERMAN UNIVERSITY

LIST OF FIGURES

Figures	Page
Figure 1. SOT Implementation Stages.....	13
Figure 2. General Gap Analysis.....	15
Figure 3. Fishbone Diagram.....	17
Figure 4. Main Problem.....	18
Figure 5. Data Mining as a Step in KDD Process.....	21
Figure 6. Sample of Classification Modeling.....	22
Figure 7. Sample of Clustering.....	23
Figure 8. Sample of Association Rule Mining Process.....	25
Figure 9. Example of FP-Tree Construction.....	28
Figure 10. Contingency Table for Variables X and Y.....	30
Figure 11. Probability and Impact Matrix.....	33
Figure 12. CRISP-DM Process Model.....	38
Figure 13. CRISP-DM Tasks.....	39
Figure 14. Theoretical Framework.....	40
Figure 15. Research Methodology.....	43
Figure 16. Data Collection Process.....	50
Figure 17. Data Preparation Process.....	54
Figure 18. Data Modeling Process.....	55
Figure 19. Analysis Process.....	59
Figure 20. Pareto Diagram using Losses Volume.....	60
Figure 21. The Consequent Result of Threshold-1.....	62
Figure 22. The Consequent Result of Threshold-2.....	62
Figure 23. The Consequent Result of Threshold-3.....	63
Figure 24. Risk Matrix.....	66
Figure 25. Losses Risk Mapping.....	67
Figure 26. Final Rules.....	77

LIST OF TABLES

Table	Page
Table 1. SOT Production Monitoring Data.....	14
Table 2. Basic Data Mining Techniques.....	21
Table 3. Data Mining Techniques Comparison	24
Table 4. Comparative Table.....	28
Table 5. Losses Category based on PRODML	33
Table 6. Losses Category Based on PUPO-PPAM.....	35
Table 7. Conditions of Conclusion	48
Table 8. Losses Data Dimension	51
Table 9. Losses Data Attributes	51
Table 10. Metadata of Numerical to Binominal Operator Output.....	55
Table 11. The Highest Losses Volume	60
Table 12. Financial Impact for Top Loss Category	61
Table 13. Threshold Changing.....	61
Table 14. Risk Probability Level	64
Table 15. Risk Impact Level.....	65
Table 16. Grouping of Antecedent.....	68
Table 17. Grouping of Consequent.....	69
Table 18. Group of Support Value.....	70
Table 19. Group of Support Value 0.101955.....	72
Table 20. Consequent with The Most Top Losses.....	74
Table 21. The Process of Association Rules.....	74
Table 22. Final Result.....	76
Table 23. Top Losses Analysis	78
Table 24. Final Rules Analysis	79