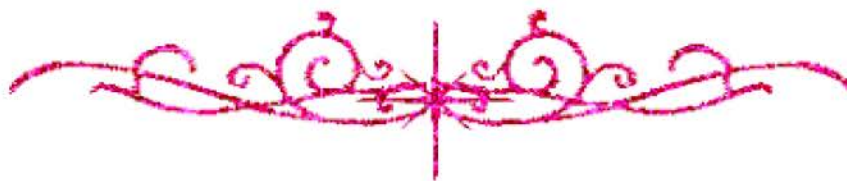


بسم الله الرحمن الرحيم





شبكة المعلومات الجامعية التوثيق الالكتروني والميكرو فيلم



جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

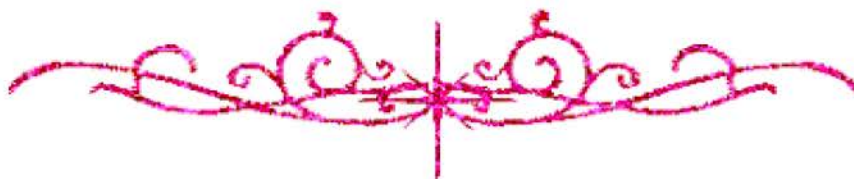
قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات



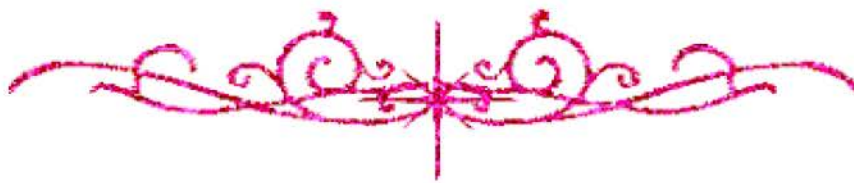
يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



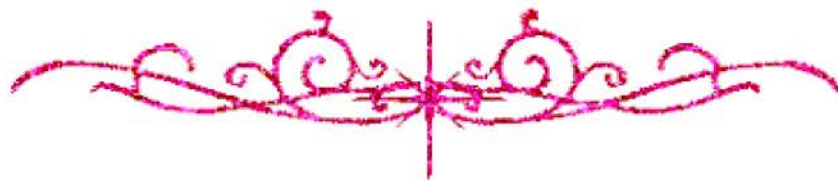


بعض الوثائق الأصلية تالفة





بالرسالة صفحات
لم ترد بالأصل



B1211A

Cairo University

Institute of Statistical Studies and Research

Department of Computer and Information Science

A Tool For Knowledge Discovery In Life Assurance Data

By

Suzanne Soliman Ali Shafik

**Thesis submitted in partial fulfillment
of the requirements for Ph.D. in
Computer and Information Science**

Supervised by

Prof. Dr Ahmed Rafea

Computer science Dept.,
Faculty of Computers and Information
Cairo University

Dr Khaled Shaalan

Associate prof.
Computer science Dept.,
Faculty of Computers and Information
Cairo University

2003

Cairo University
Institute of Statistical Studies and Research
Computer Science and Information department.

Approval sheet

A tool for knowledge discovery In life assurance data

**A thesis submitted in partial fulfillment of the
Ph.D. in Computer science
By**

Suzanne Soliman Aly Shafik

**Computer Science and Information dept.
Institute of Statistical Studies and Research
Cairo University**

2003

This thesis for PH.D. degree in computer sciences, department of computer science and information of statistical studies and research, Cairo University, has been approved by:

Name	Signature
Prof. Dr. Ahmed Rafea	 <i>Ahmed Rafea</i>
Prof. Dr. Fawzi Amer	 <i>M. F. Amer</i>
Prof. Dr. Mervat Gheith	 <i>Mervat Gheith</i>
Dr. Khaled Shaalan	<i>K. Shaalan</i>

Date : 20/7/2003

Acknowledgment

I would like to express my deep thanks and sincere appreciation to Prof. Dr. Ahmed Rafea to his valuable effort, continues support and expert guidance through all phases of this work.

My gratefully thanks, appreciation, and gratitude to Dr. khaled Shalaan for his fruitable effort, encouragement, support, and cooperation during all phases of the work.

My deep thanks to Prof. Dr. Mawaad Hasaneen Chairman of Misr Insurance Co. for his approval of using company life assurance historical data during the implementation phase.

My sincere thanks to IT department in Egyptian Insurance Supervisory Authority for providing me with life assurance market statistics in electronic form.

All my appreciation, thanks and gratitude to ISSR Computer and information science department for all kind provided facilities and cooperation.

My deep thanks, appreciation and gratitude to Prof. Dr. Adeeb Riyad Ghonemy to his support through valuable discussions, and who made all his library available during the research time.

Finally my deep sincere thanks to all my colleagues in IT sector in Misr Insurance Co. for their efforts and cooperation during implementation phase of this work.

Statement

I certify that this work has not been accepted in substance for any academic degree, and is not being concurrently submitted in candidature for any other degree. Any portions of this thesis for which I am indebted to other sources are mentioned and explicit references are given.

Student name :

Suzanne Shafik

Table of Contents

Chapter1 INTRODUCTION

1-1	General procedures to build a knowledge discovery tool	3
1-1-1	Preprocessing and Building a data warehouse Phase	3
1-1-2	Data mining and post processing	4
1-2-	Motivation of building knowledge discovery tool for life assurance	4
1-2-1	Problems related to environmental factors	4
1-2-2	Problems related to companies' internal factors	6
1-2-3	Problems related to technical issues	6
1-2-4	Problems related to life assurance data characteristics	7
1-3	Thesis objective	7
1-4	Proposed methodology for life assurance knowledge discovery	8
1-5	Thesis layout	8

Chapter2 Knowledge Discovery in Business Organizations

2-1	Overview	10
2-2	Drivers for Knowledge discovery in business organizations	11
2-3	Customer Relationship Management (CRM) systems	12
2-3-1	Customer Relationship Management (CRM) definition	12
2-3-2	One-to-one profiling system	13
2-4	Direct Marketing Knowledge discovery systems	17
2-4-1	Direct marketing Knowledge discovery types	17
2-4-2	Direct marketing. Applications	17
2-5	Financial Investment Knowledge Discovery systems	18
2-5-1	Financial Investment Knowledge discovery characteristics	18
2-5-2	Genetic Learner Overlaid with Entropy Reduction (GLOWER)	19
2-5-3	R-Mini system	20
2-6	Banking Knowledge Discovery in Database systems	20
2-6-1	Application of Knowledge Discovery in Banking	20
2-6-2	Barclays Information Warehouse review	20
2-7	Knowledge Discovery in Insurance System	22
2-7-1	Automatic Analysis of Deviations: Health-KEFIR	23
2-7-2	ALDER mining system	26
2-7-3	UPA (IBM's Underwriting Profitability Analysis) System	28
2-8	Conclusion	31

Chapter3 Knowledge Discovery in Database (KDD) Techniques

3-1	Definitions of Knowledge Discovery in Databases (KDD)	32
3-2	Review of preprocessing phase previous work	34
3-2-1	Acquisition of operational data bases	34
3-2-2	Preprocessing operations	34
3-3	Review of data warehousing techniques	47
3-3-1	Design of a Data Warehouse	38
3-3-2	Meta-data for Warehousing	39
3-4	Review of data mining techniques	41
3-4-1	Decision tree model representation	44
3-4-2	Definition of Neural Network model representation	46
3-4-3	Clustering and Nearest Neighbor techniques	50
3-4-4	Rule Induction	54
3-4-5	Evaluation of data mining techniques	60
3-4-6	Summary of data mining techniques evaluation	65
3-5	Review of post processing techniques	68
3-5-1	Visualization techniques	68
3-5-2	Interaction technique	69

CHAPTER 4 Life Assurance Domain Ontology Analysis

4-1	Ontological analysis of life assurance domain.....	70
4-2	Development of life assurance domain ontology	72
4-2-1	Specification of life assurance ontology goal and scope	72
4-2-2	Conceptualization phase	72
4-3	Management of ontology	79

CHAPTER 5 Preprocessing and Building a Data Warehouse

5-1	Life assurance preprocessing phase	83
5-1-1	Ratio table conversion (Generate Boolean feature).....	84
5-1-2	Descritization of continuous attributes.....	85
5-1-3	Unification of customer code.....	86
5-1-4	Construction of new attribute.....	87
5-1-5	Feature selection	87
5-1-6	Building a composite record from related historical policy transactions.....	88
5-1-7	OLAP transformation.....	89
5-2	Build a Life Assurance Data Warehouse	90
5-3	Loading and maintaining life assurance data warehouse	97
5-3-1	Customization and initial user installation of Data Warehouse	97
5-3-2	Refresh-date-warehouse-by-periodical transactions	100
5-4	Preprocessing and building a data warehouse meta data	100
5-4-1	Transformation and loading meta data sub-model	100
5-4-2	Data warehouse meta sub-model	101

Chapter 6 Life Assurance Data Mining

6-1	Data mining module description	103
6-2	Generalized Composite Record Clustering	105
6-3	Macro level rule induction	108
6-4	Micro level rule induction	110
6-5	Experimental results of proposed algorithms.	114
6-5-1	Impact of the proposed approach on Rules generated.....	115
6-5-2	Impact of changing selected mining parameter on rules generated	118
6-6	Data mining meta sub-model.....	119
6-7	Post processing module.....	120
6-7-1	Description of post-processing	120
6-7-2	Comparison of different induced rules levels	121

Chapter 7 Implementation of Repository Management System in Knowledge Discovery Life Assurance Tool

7-1	Tool repository data base system role	122
7-2	Architecture of Life Assurance Repository Management System	122
7-2	Life Assurance System repository management functions	125

Chapter 8 Knowledge Discovery Life Assurance (KDLAT) User interface Implementation(case study)

8-1	Implementation aspects of KDLAT user interface.....	129
8-2	KDLAT case study	131
8-2-1	Case study for Preprocessing phase	133
8-2-2	KDLAT case study for data mining module	135
8-2-3	Case study of KDLAT post-processing	148
8-3	How Generated Rules could serve Strategic and Tactic Decisions.....	152

Chapter 9 Conclusion

Figures

Figure 2-1	Knowledge as strategic factor.....	10
Figure 2-2	The CRM management process	12
Figure 2-3	The 1: 1 Pro system architecture.....	14
Figure 2-4	1: 1 Pro system case study components.....	16
Figure 2-5	Barclays Information Warehouse architecture.....	22
Figure 2-6	Overall design and process flow within KEFIR.....	24
Figure 2-7	KEFIR search space	24
Figure 2-8	ADLER System Architecture	27
Figure 2-9	UPA Architecture	29
Figure 3-1	Architecture of Knowledge Discovery in Databases (KDD).....	33
Figure 3-2	Meta-model levels.....	39
Figure 3-3	Example of decision tree for classification of customer loyalty	44
Figure 3-4	Example of Neural Network.....	47
Figure 3-5	Agglomerative successive clustering	52
Figure 3-6	Comparison of business evaluation measures for data mining techniques	66
Figure 3-7	Comparison of application evaluation measures for data mining techniques	67
Figure 3-8	Comparison of algorithmic evaluation measure for data mining techniques.....	67
Figure 4-1	Dynamic domain event analysis.....	71
Figure 4-2	Ontology Development Phases	72
Figure 4-3	Life assurance domain relation diagram.....	77
Figure 4-4	Ontology sub-model schema.....	79
Figure 5-1	Life assurance preprocessing functions	83
Figure 5-2	Calculation of ratio variance algorithm. pseudo code.....	84
Figure 5-3	Discretization transformation algorithm pseudo code	86
Figure 5-4	Conversion algorithm from intertransactions to composite-file table pseudo code	89
Figure 5-5	Entity Relationship schema of life assurance data warehouse	91
Figure 5-6	Steps for loading / maintaining data warehouse.....	97
Figure 5-7	Loading of Data Warehouse algorithm pseudo code	99
Figure 5-8	Refresh Data Warehouse by Periodical Transactions algorithm pseudo code.....	100
Figure 5-9	Transformation sub-model schema.....	101
Figure 5-10	Data Warehouse meta sub-model schema.....	101
Figure 6-1	Data mining model components.....	104
Figure 6-2	Creation partition-composite DB algorithm pseudo code.....	106
Figure 6-3	Maintaining partitioning composite DB algorithm pseudo code.....	107
Figure 6-4	Macro level rule association mining algorithm pseudo code	109
Figure 6-5	Extraction of generated macro rules.....	110
Figure 6-6	Micro Level rule induction algorithm pseudo code.....	113
Figure 6-7	Extraction of generated micro level rules	114
Figure 6-8	Example of variable support matrix.....	115
Figure 6-9	Generated rules by using variable support matrix	116
Figure 6-10	Generated rules by using fixed support matrix.....	116
Figure 6-11	Numbers of rules at different values of support for equal matrix of support in micro level.....	117
Figure 6-12	Time at different values of support for equal matrix of support in micro level	117
Figure 6-13	Number of rules and time for variable matrix support	118
Figure 6-14	Number of rules at different values of confidence.....	119
Figure 6-15	Number of relevant items at different values of support.....	119
Figure 6-16	Data Mining Sub-model schema.....	120
Figure 6-17	Comparison of macro level rules algorithm pseudo code*.....	121
Figure 7-1	Architecture of meta-model.....	123
Figure 7-2	Interdependencies between different repository sub-models.....	124
Figure 7-3	Functions of system repository management.....	124
Figure 7-4	Building /Maintenance KDALT Ontology Meta Submodel.....	125
Figure 7-5	Define Access Right Code.....	126

Figure 7-6	KDALT Glossary Maintenance	126
Figure 7-7	KDALT Class Instance Table	127
Figure 7-8	KDALT Relation Table	127
Figure 7-9	List Ontology Components	127
Figure 7-10	List KDALT Ontology Components by Alphabetical Order	128
Figure 7-11	List KDALT Ontology Components by Sequence No.	128
Figure 7-12	List of Attributes and Its Association.	128
Figure 8-1	Dialog flow of KDLAT Screens	130
Figure 8-2	Case study architecture	131
Figure 8-3	KDLAT system main operational menu	132
Figure 8-4	Preprocessing phase menu.	133
Figure 8-5	Discretization parameter	133
Figure 8-6	Condense Rows.....	134
Figure 8-7	Building Aggregation Record.	134
Figure 8-8	KDALT Rule Induction Sessions	136
Figure 8-9	KDALT Rule Induction Levels	136
Figure 8-10	KDALT Macro Level Rules	136
Figure 8-11	Macro Level Rule Induction Scope	137
Figure 8-12	Predefined macro task new emission	137
Figure 8-13	Predefined macro task new emission support and confidence	138
Figure 8-14	Output rules for new emission for task no. 1 for market.....	138
Figure 8-15	Print out of output rules for new emission for task no. 1 for market	138
Figure 8-16	Screen layout of input new emission for task no. 1	139
Figure 8-17	Screen layout of input support and confidence for new emission	139
Figure 8-18	Screen layout of output rules for new emission for task no. 1	140
Figure 8-19	Print out of output rules for new emission	140
Figure 8-20	Predefined macro task – new emission for governorate	141
Figure 8-21	Screen layout of output rules for new emission	141
Figure 8-22	Print out of output rules for new emission	142
Figure 8-23	Screen layout of input new emission for expert user represent	142
Figure 8-24	Screen layout of input new emission support and confidence	143
Figure 8-25	Screen layout of output rules for new emission	143
Figure 8-26	Screen layout of output rules for new emission	143
Figure 8-27	Micro Level Rule Induction	144
Figure 8-28	Micro level main task selection menu	144
Figure 8-29	Micro level sub task selection menu	144
Figure 8-30	Select predefined support matrix or enter new matrix support	145
Figure 8-31	Micro level main task selection menu	145
Figure 8-32	Micro level sub task selection menu	145
Figure 8-33	Screen layout of accepted variable support matrix in micro level	146
Figure 8-34	Part of generated micro rules	146
Figure 8-35	Print out of output rules for annulation	147
Figure 8-36	Selection screen for partition to use rules into prediction	147
Figure 8-37	Screen layout of predicted data	148
Figure 8-38	Selection menu of KDLAT post processing module	148
Figure 8-39	Screen layout of selection of rule scope analysis	149
Figure 8-40	Screen layout of selection of rule scope analyses	149
Figure 8-41	Rule scope analysis report.....	150
Figure 8-42	Macro generated rules	151
Figure 8-43	Print out of generated rules	151
Figure 8-44	Example of life assurance mined rules.....	152

Tables

Table 3-1	Differences between the Nearest-Neighbor data mining technique and clustering.	30
Table 3-2	Trade off between coverage and accuracy	54
Table 3-3	Business measures for reviewed data mining techniques	60
Table 3-4	Application measures for reviewed data mining techniques	61
Table 3-5	Algorithmic measures for reviewed data mining techniques	63
Table 4-1	Extraction of life assurance glossary of terms	73
Table 4-2	Class attributes for concept name: Life assurance policy	76
Table 4-3	Extraction of Life Assurance attributes table	77
Table 4-4	Life assurance relations table	78
Table 4-5	Attributes of table: glossary of term definition	80
Table 4-6	Attributes of table: ontology attribute table definition	80
Table 4-7	Attributes of table: class attributes table definition	81
Table 4-8	Attributes of table: ontology relation table definition	81
Table 4-9	Attributes of table: access- right table definition	81
Table 4-10	Attributes of view: Associations of attribute-term definition	82
Table 5-1	Life policy table attributes.	92
Table 5-2	Composite Transaction table attributes description	93
Table 5-3	Grouped Composite Transactions table attributes description	93
Table 5-4	Aggregated Assurance Market Dimension table attributes description	93
Table 5-5	Company Portfolio aging Transactions entity attributes description	94
Table 5-6	Demographic indicators table attributes description	95
Table 5-7	Economic Indicators table attributes description	96
Table 5-8	Saving channels tables attributes descriptions	96
Table 5-9	Comparison of savings to national income concept	97
Table 6-1	Valid episodes of events.	105
Table 8-1	Case study used records	135

Abstract

Many problems face the insurance companies due to the particularities of insurance industry. These problems are related to environmental factors, internal factors, and technical factors. Discovering relevant knowledge from available large volume of historical data and experts domain knowledge, could support the decision makers in achieving their goals.

This work uses prior domain knowledge to guide the mining of association rules in life assurance business environment. This approach is used in order to overcome the drawbacks of data mining using rule induction such as loss of information, discover too many obvious patterns, and mining of overwhelmed association rules. Life assurance data consist of enforced life assurance policies records, and many relevant historical transactional records. Each historical transaction has a transaction type associated with time stamp of occurrence, which represents an event to the assurance policy. These events compose an episode of events that have a certain pattern such as (emission, annulation, re-emission), (emission, amendment, loans, liquidation), or....etc. The episode of events is a subset of the transactions item-list. The final status of enforced policies in the company portfolio is affected dynamically by the sequential order of episode of events occurred during the policy life since its emission. This effect represents a positive or negative status on the company portfolio growth. It is required to handle dynamic nature of data, where appending a new event to an existing episode of events may change the status of life assurance policies, change discovered patterns and affect the support of mined rules. Two types of data are identified and used in mining rules, global data its attributes are time series of economic and demographic indicators, and insurance data aggregated on different levels and dimensions.

In order to help the decision maker a tool was built to discover knowledge at the macro and micro levels. The discovered rules, at the macro level, describe the impact of changes of national economic and demographic indicators on the growth of life assurance business. The discovered rules, at the micro level, describe the impact of different policies attributes, customer profiles, and market channels on company portfolio growth.

Our approach for data mining is a hybrid approach that consists of two data mining techniques namely: clustering and rule-induction. The clustering technique is applied on preprocessed multidimensional quantitative and categorical operational database