AN INTERACTIVE SYSTEM FOR ASSOCIATION RULE DISCOVERY FOR LIFE ASSURANCE

Ahmed A. Rafea
Computer Science Dept., American University in Cairo
113, Kasr El-Aini St,P.O. Box 2511, 11511,Cairo, Egypt
Rafea@aucegypt.edu

and
Suzanne S. Shafik
Chief Executive Communication & information Tech sector,
Misr Insurance Comp.44A El Dokki St.Giza – Egypt
S_shafik@misrins.com

and
Khaled F. Shaalan
Computer Science Dept., Faculty of Computers and Information
Cairo Univ., 5 Tharwat St., Orman, Giza, Egypt
Shaalan@claes.sci.eg

ABSTRACT

This paper 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. A data mining interactive rule induction algorithm is introduced to mine rules at micro levels. The mined rules describe the impact of different insurance policies attributes, customer profiles, and market channels on company portfolio growth. A system was built based on this algorithm and was tested and verified on real data set in Misr Insurance company, which is the leading insurance company in Egypt.

Key words : Drawbacks of rule induction, Discretization, Rule Discovery For Life Assurance, clustering, variable support matrix.

1.INTRODUCTION.

Many problems face the insurance companies due to the particularities of insurance industry. These problems are related to economic developments, social demographic changes, competition, and companies' internal factors [27]. Insurance companies try to discover patterns that can help, direct and focus marketing efforts to achieve their goals in increasing new policies emission and decreasing policies withdrawal. Discovering relevant knowledge from available large volume of historical data and experts domain knowledge, could support the decision makers in achieving their goals. Life assurance data contains quantitative and categorical attributes. Two types of data are identified to be used in mining rules: 1) global data holding time series of economic and demographic indicators, and 2) life assurance policies , and many relevant historical transactions. Each historical transaction has a transaction type associated with time stamp of occurrence[26], 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 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 [21]. 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.

In order to help the decision maker a system should be built to discover knowledge at the macro and micro levels. The discovered rules, at the macro level, should 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, should describe the impact of different policies attributes, customer profiles, and market channels on company portfolio growth that is described in terms of the dynamic changes of the episodes of events. In this paper we will concentrate on discovering knowledge at the micro level.

The rest of this paper is organized as follows. In section 2, we discuss related work and the limitations of some of the existing techniques used in discovering association rules. Section 3 presents the architecture of a system for knowledge discovery in life assurance. Section 4, describes briefly the preprocessing and data warehouse building phases. Section 5 introduces the proposed data mining algorithm. Section 6 presents a case study, while section 7 concludes with suggestion for future directions.

2-RELATED WORK

Knowledge discovery in database (KDD) process is dynamic, interactive and iterative. It consists of preprocessing, data mining and post processing phase [1], [5], [19], [3], [23], [4], [25], [12], [14]. Association rules are one of the most used techniques in data mining to discover relationships among attributes in a database [11], [16], [17]. Mining association rules in large relational tables, containing both quantitative and categorical attributes, have been introduced in [13], [8]. Some problems are encountered in mining association rules in large relational tables. These problems are mainly clarity and loss of information. The clarity problem is manifested in discovering obscure rules holding too many obvious patterns that overwhelm the user. Many techniques are introduced based on human involvement in data mining to overcome the clarity problem [18], [22]. Some of these techniques are constraint-based mining, and post-processing the generated rule base by domain expert. Constraint-based mining presents architecture for exploratory association mining. This architecture consists of two phases [15], [7]. In phase one, the user initially specifies constrained association query by using ad hoc mining query language to define the part of database to be mined, the type of pattern/ rule to be mined, and the properties that the pattern should satisfy, and provides a sophisticated mining-query optimizer. Phase two represents mining phase where its output consists of all associations' relationships that satisfy the defined conditions.

Another technique involves expert driven validation to post processing generated rule base by using several types of validation operators including rule grouping, filtering, browsing, and redundant rules elimination [5], [2]. Loss of information problem occurs due to discretization of quantitative attributes. If the number of intervals is large, the support for any single interval can be low and some rules involving this attribute may be lost because they lack minimum support. If the size of interval is large, rules may be lost because they lack minimum confidence. Previous works concentrated on solving problem of loss of information by automatic adjusting the minimum and maximum interval thresholds for mining association rules on descritized quantitative attributes. In [13], the user defines parameters of minimum support, minimum confidence, and maximum support, the introduced algorithm merges the adjacent intervals until the summation of the support of all merged intervals is equal to maximum support threshold. Another approach is stated in [8], where the merge of adjacent intervals is performed in a bottom-up manner by using modified b-trees algorithm on the basis of maximizing the statistical interestingness of a set of association rules. These algorithms proceed as a black box, admitting a little user interface, they suffer from lack of user control and ignore valuable prior domain knowledge to generate understandable rules.

3. A PROPOSED SYSTEM FOR KNOWLEDGE DISCOVERY IN LIFE ASSURANCE

The proposed system consists of two main phases: preprocessing and building a data warehouse phase, and data mining phase. These phases are interconnected through a repository management system [10] that contains an ontology for life insurance industry, data warehouse definitions, and preprocessing and transformation algorithms definitions

3.1 Preprocessing Phase And Building a Life Assurance Data Warehouse

3.1.1 Preprocessing phase

Preprocessing phase handles extraction of source data from operational data, transformation and loading these data into data warehouse. It reconciles the syntactic and semantic differences between operational data sources and data warehouse. [11], [24]. In the following, we will present the most important preprocessing procedures.

-Generate Boolean attributes

This procedure is used to convert the continuous attributes holding some national economic and demographic time series into Boolean attributes denoting the direction of ratio of variance of these indicators. Three attributes are generated to define the direction of ratio of variance(increasing , constant, decreasing) .

-Discretization of continuous attributes

In order to induce association rules, some continuous attributes are transformed to discrete categorical values to be closer to a knowledge-level representation than continuous values [6]. Attributes are discretized through interacting with the user to acquire his/her interesting focus area. Discretization methods are provided to produce regular or irregular intervals. In regular interval method, the user defines the start value for an attribute and the interval value, or the user defines number of intervals and the discretization method generates interval table based on the values of the given attribute in the data. In irregular interval method, the user defines the lower and upper limits of each interval to focus on specific interval of the attribute values.

- **Building a composite record from related historical policy transactions.**

  This procedure is used to overcome the dynamic nature of life assurance policy. The procedure concatenates a life assurance policy record, and related historical transactions records to generate a composite record. A composite record consists of two parts. Part one holds life assurance policy attributes namely insurance-code, tariff-code, customer demographic information, marketing channel information, and policy financial information. Part two holds the historical ordered episode of events (transactions) stored as an array. Each cell in the array consists of some of the transaction attributes namely type code, effect on portfolio status, and transaction date.

3.1.2 Life Assurance Data Warehouse

Acquiring the factors affecting the growth of life insurance business at the national level and the portfolio growth of a pacific company, an entity relationship model of life assurance data warehouse is designed. Figure 3 depicts life assurance data warehouse schema Life assurance data warehouse consists of three types of entities: replicates of operational database entities, a derived data cube, and external entities.

- **Replication of operational database:** It consists of life policy entity, customer personal insurance policies, customer, producer, agent, coding table, valid episode of events table, composite life policy entity

-Data cube : Data cube entities consist of two types of historical aggregated life transactions entities. The first type is an aggregation of transactions related to transaction type such as emission, re-emission,etc., and date. The second type is an aggregation of transaction according to the elapsed period between policy' emission date and transaction date.

- **External entities**: External entities consist of demographic indicators, economic indicators and saving channel indicators used to enrich assurance data.

3.2 -LIFE ASSURANCE DATA MINING PHASE

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 the composite records and valid episodes of events records from the replicates of operational database entities in the data warehouse to partition the generalized composite records based on the pattern of valid episode of events. The rule induction phase works on the appropriate partition to generate association rules related to the requested mining task namely customer loyality, expansion of market channel, and insurance product evaluation. The following sections describe these two phases.

3.2.1 Generalized Composite Record Clustering

The clustering process consists of two main tasks. The first task is to generate a generalized composite records by dropping specific information. The second task is to cluster the generalized composite records according to the effect of the episode of events on the company portfolio as shown in figure1. The clustering process is described in figure 2

Seq	Episodes of events	Effect on portfolio	P. No.
1	Emission, annulation	Withdrawal by annulation	1
2	Emission, annulation, rie-emission	Enforce by re-emission	2
3	Emission, amendments increase, annulation	Withdrawal by annulation	1
4	Emission, amendments decrease, annulation	Withdrawal by annulation	1
5	Emission, amendments, annulation, reemission	Enforce by re-emission	2
6	Emission	Enforce by Emission	3
7	Emission, annulation, re-emission, early-death	Withdrawal by earlydeath	4
8	Emission, amendments, annulation, re- Emission, early-death	Withdrawal early-death	4
9	Emission, liquidation	Withdrawal byl iquidation	5
10	Emission, liquidation, re-emission	Enforce by re-emission	2
11	Emission, amendments, diminution	Enforce and diminution	6
12	Emission, amendments, loans	Enforce with loans	7
13	Emission,amendments,loans , liquidation	Withdrawal after loans by liquidation	8
14	Emission, death	Withdrawal by death	9
15	Emission, maturity	Withdrawal by maturity	10
16	Emission, amendments, loans, maturity	Withdrawal by maturity	10

Figure 1 Valid episodes of events.

* *Create generalized composite record**

for each composite record do

{

Drop specific information which is policy number customer code, producer code, from composite records.

If this record exist in the generalized composite table

Then add 1 to -weight field,

Else create a new record in the generalized composite table

}

3- for each generalized composite record do

{

If the episode of events not exist into valid episode of events table (table 1),

then if the logical sequence of historical transactions of generalized- composite record is valid

then , add the new episode of events to valid episode of event table,

else write the erroneous code of event into invalid episode of event partition to be examined by the expert user.

else get the effect on portfolio from valid episode of events table and write the generalized record in the appropriate cluster table

}

Figure 2 Clustering process algorithm.

In case that new transactions are added to the operational database, the data warehouse composite record will be appended accordingly. The clustering algorithm is designed to maintain the rules generated incrementally, i.e. not to repeat the clustering procedure from scratch; a process that takes long time. The incremental updating method takes into consideration the occurrence of one of the following events: A composite record is moved from one group (generalized composite record) to another in the same cluster, which affects weights of both groups within the cluster. A composite record is moved from one cluster to another. This will affect the number of the generalized composite records in the two updated clusters. A new generalized composite record is added to a cluster, which increase the number of records in this cluster by the weight of the generalized record. An existing group is deleted from a cluster, which decrease the number of generalized records within the cluster by the weight of this record. There is also a possibility that a complete new cluster is created and hence all needed adjustments will be done.

3.2.2 Rule Induction

Rule induction algorithms are based on well known existing techniques to obtain association rules as Apriori algorithm. These algorithms are modified to enable a user to control and impose his area of focus during knowledge discovery steps in order to overcome the loss of information problem and to enable him/her to generate rules that he/she is interested in. Loss of information problem occurs as result of discretization. The proposed algorithm solved this problem by allowing the user to define the relative weight or support of each attribute interval category such that the mining algorithm could generate rules using this attribute interval category only if this support is satisfied. For example if the total number of composite records is 100 in a certain cluster and the user specifies the support as 10%, for a certain interval category of an attribute, rules can only be generated from the generalized composite records that contain this attribute category interval, if the total weight of these records is at least 10. This technique will enable the user to choose the relevant attribute value by giving it a small relative weight to enforce the mining algorithm to generate rules having this attribute value in their premises even if the number of occurrence of records having this attribute value is small. Less relevant attribute value is to be given higher support such that it will be pruned if the number of occurrence of records having this attribute value is small.

The induction algorithm uses the data stored in each cluster to generate rules that describe customer loyalty, company's market expansion for products in different geographical areas, and insurance types evaluation. Mining each category of these rules is defined as a mining task. The generated rule premises are a subset of instance of factors affecting portfolio status and the conclusion part is a an episode of events that cause this effect. Figure 6 depicts the proposed algorithm

1- Select the mining task and consequently the appropriate cluster
2- Get the confidence threshold for generating a rule (this means that the rule will only be generated if the number of occurrences of records described by this rule divided by the total number of records in the cluster greater than the given confidence threshold)
3- Construct a matrix (calculated relative weight) with number of rows equal to the number of attributes (m) and number of columns (n) equal to the maximum number of categories of a certain attribute
4- Using the appropriate cluster, fill in the calculated relative matrix with the relative weight of each attribute category in this cluster
5- Compare the calculated relative weight with the user given support and mark irrelevant attributes categories.
6- For each generalized composite record do
7- For each generalized composite record attribute do
{ if the attribute category is irrelevant
then mark it as irrelevant
copy relevant attributes category into a new table}
8- Group similar rows in the new table and calculate a confidence value for this grouped records
9- Generate rules

Figure 3 Rule induction proposed algorithm

4. CASE STUDY

A sample of knowledge discovery at micro level system is implemented. The data sources were operational life assurance data extracted from different geographical locations from Misr Insurance Company . The number of policy records was 177446. The number of historical transactions records was 352681. The number of the generated composite records was 283365. The number of generalized composite records was 131353. In the following paragraphs we will demonstrate two experiments. The first experiment is to show the impact of our approach in acquiring variable support for each attribute category from the user. The second experiment demonstrates how the generated rules can assist the decision maker in taking strategic and tactic decisions.

4.1 Impact of The Proposed Approach on Rules Generated

At the beginning of any mining task, the system acquires the support for each attribute category defined at discretization step during preprocessing phase of a generalized composite record in the corresponding cluster. Figure 4 depicts the user interface screen that acquire these supports.

In order to show how our technique has enhanced the rule generated, we conducted the following experiment steps:
Run the system and give variable support for each attribute category based on the user interest.

1) Count the number of rules generated and the number of used premises in these rules
2) Rerun the system and give equal support for all attributes categories.
3) Count the number of rules and the premises used in these rules
4) Examine the quality of rules generated in each case by comparing the number of rules and premises used

Running this experiment to generate rules related to the customer loyality mining task, the number of generated rules using the variable support matrix was 27 rules and the number of used premises is 10. In case of fixed support matrix to generate rules for the same mining task, the generated number

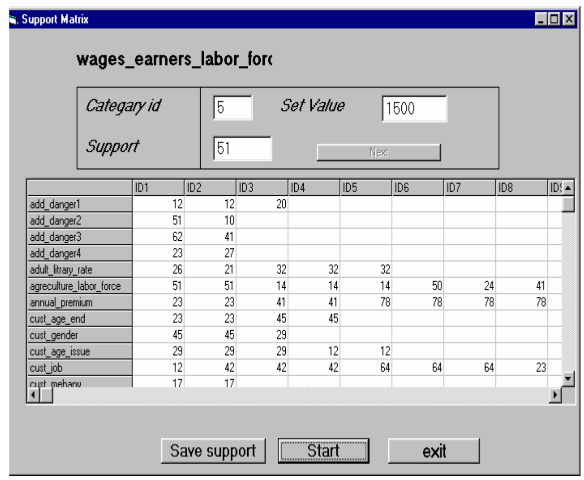

Figure 4 Screen layout of accepted variable support matrix used

of rules was 9 and number of used premises was 7.

3 **, IF** cust_age_issue Between (21 , 30) **and** cust_age_**end** Between (40 , 50) **and** cust_gender = Male and sum_assured <= 5000 **and** annual_premium <= 500 **and** policy_period Between (15 , 20) **and** policy_issue_date Between (1995 , 2000) **and** life_expecting_at_birth Between (28 % , 64 %) **and** gdp_per_capital Between (3000 , 4000) THEN Tran1_Type = Emission **and** Cust_Age_Tran1 Between (30 , 35) **and** Tran2_Type = Annulation **and** Past_Period2 = 1 **and** Cust_Age_Tran2 Between (30 , 35)

9) **IF** cust_age_**end** Between (50 , 60) **and** cust_gender = Male **and** sum_assured <= 5000 **and** annual_premium <= 500 **and** policy_period Between (15 , 20) **and** policy_issue_date Between (1995 , 2000) **and** life_expecting_at_birth Between (28 % , 64 %) **and** labor_force_population Between (25%,28%) THEN Tran1_Type = Emission and Tran2_Type = Annulation and Past_Period2 = 1 **and** Cust_Age_Tran2 Between (50 , 60)

Figure 5 Sample of generated rules by using variable support matrix

2) **IF** cust_age_issue Between (21,30) **and** cust_gender = Male **and** policy_period Between (15, 20) **and** policy_issue_date Between (1995, 2000) **and** life_expecting_at_birth Between (25%,28%) **and** labor_force_population Between (25%,28%) THEN Tran1_Type = Emission **and** Tran2_Type = Annulation **and** Past_Period2 = 1

3) **IF** cust_age_issue Between (35 , 40) **and** cust_age_**end** Between (50 , 60) **and** policy_issue_date Between (1995 , 2000) **and** life_expecting_at_birth Between (25 % , 28 %) **and** THEN Tran1_Type = Emission **and** Cust_Age_Tran1 > 40 **and** Tran2_Type = Annulation **and** Cust_Age_Tran2 > 40

Figure 6 Sample of generated rules by using fixed support matrix

As we can see in figures 5 and 6 the rules generated in the first run was more specific and focus on the area of interest of the user. Whereas the rules generated in the second run was more general and not necessary cover the user needs. Some of the information was lost such as annual_premium, sum_assured

4.2 How Generated Rules Could Serve Strategic and Tactic Decisions

Induced rules support both strategic and tactical decision making for each mining task in the system. An example of strategic decision is to decide characteristics of customers who are unloyable and annulated his/her policy less than three years since the emission date. Hence, the decision maker can decide not to exert efforts in marketing to acquire new policies from this category of customers. Figure 7 depicts two typical rules that have been generated from the system. The conclusion of those two rules may be interpreted by the top management as "Do not target age groups from (21 – 30) and (50-60) for with profit life assurance type included in this data set.

These rules can also be used to take tactic decision . For example, Rule 1 tells the decision maker that customers who issued a policy at age(21 years - 30 years), and customer age at maturity date between 40 years and 50 years , customer gender is male, and sum assured les than 5000 pounds,….etc , are expected to conduct this episode of events: Emission , annulation and elapsed period between emission and anuluation is one year, and customer age_at annulation between (21,30). Taking this episode of events and excluding the last event (annulation) and query the operational database, a set of customers could be generated to support the decision maker to take a preventive action to keep these customers policies in the company portfolio

Micro Level
Customer Loyalty
Annulations
==
confidence : 15
----
8) **IF** cust_age_issue Between (21 , 30) **and** cust_age_**end** Between (40 , 50) **and** cust_gender = Male **and** sum_assured <= 5000 **and** annual_premium <= 500 **and** policy_period Between (15 , 20) **and** policy_issue_date Between (1995 , 2000) **and** life_expecting_at_birth Between (28 % , 64 %) **and** gdp_per_capital Between (3000 , 4000) THEN Tran1_Type = Emission **and** Cust_Age_Tran1 Between (21 , 30) **and** Tran2_Type = Annulation **and** Past_Period2 = 1 **and** Cust_Age_Tran2 Between (21 , 30)

9) **IF** cust_age_**end** Between (50 , 60) **and** cust_gender = Male **and** sum_assured <= 5000 **and** annual_premium <= 500 **and** policy_period Between (15 , 20) **and** policy_issue_date Between (1995 , 2000) **and** life_expecting_at_birth Between (28 % , 64 %) **and** gdp_per_capital Between (3000 , 4000) THEN Tran1_Type = Emission and Tran2_Type = Annulation and Past_Period2 = 1 **and** Cust_Age_Tran2 Between (50 , 60)

Figure 7 Example Of Life Assurance Mined Rules

5-CONCLUSION AND FUTURE WORK

In this paper we proposed a domain specific life assurance knowledge discovery system, to support decision making. This system can be customized to be used in different insurance types. Our contribution was in the usage of insurance prior domain knowledge in an interactive way to solve rule induction drawbacks as loss of information, , and mining of overwhelmed association rules. . This has been implemented by enabling the user to discretize the attribute values into categories representing his/her focus area and give to each category a support valued according to his interest. Consequently, the rule generated will be more informative and focus on his/her interest instead of a unique support value used in classical Apriori algorithm and its extension. This approach will lead to prune irrelevant attributes categories which not satisfy the support measure, based on user interest.

The proposed system is tested on real life assurance data extracted from different geographical locations of Misr Insurance company covering ten years, and some published national economic and demographic statistics. By using the induced rules, the decision-maker can extract different customer, market channels, and insurance product patterns, to support decision making in strategic and tactic decisions. The clustering phase that run on 283365 composite records and generates 131353, take an overnight on a Pentium III with 1GHz and 256Mb memory. In order to save time, an incremental feature is introduced in the clustering algorithm in order not to start from scratch when the operational data changes. The induction algorithm is very fast. It takes less than a minute on a cluster that contains approximately 13000 records on average.

Possible future work to complete benefit of this work in a business institution could be enhancing system functionality by integrating the system to some available decision support systems, and/or organizational information systems. To study the impact of actions taken by the decision makers as a result of generated rules on portfolio growth. Technical perspective issues are: to use the induced rules to build an underwriting expert system that will help the underwriter in his/her decision by predicting the behavior of an applicant in the future. Improve the time complexity of the data mining algorithms, by using other techniques such as genetic algorithm, neural networks, suitable for financial data characteristics, and compare the induced rules and performance measures of the new technique with the used technique.

REFERENCES

[1] Alex Berson and Stephen J. Smith, **"Data Warehousing, Data Mining, And OLAP"**, MC Graw–Hill, 1997.

[2] Christopher J. Matheus, Gregory Piatetshy–Shapiro and McNeill", **Selecting and Reporting what is Interesting The Kefir Application to Health Care Data**", Advances in Knowledge Discovery and Data Mining, AAA1 Press/The MIT Press, 1996.

[3] David Cheung, Vincent T., Ada W. Fu and Yongjian Fv, "**Efficient Mining of Association Rules in Distributed Databases**", IEEE Transactions on Knowledge and Data Engineering ,Vol 8 , No 6, Dec 1996 .

[4] Graig Silverstein, Sergey Brin and Rajeev Montwani, "**Beyond Market Baskets: Generalizing Association Rules to Dependence Rules**", Data Mining and Knowledge Discovery, Vol. 2, No. 1, Jan 1998, Kluwer Academic Publishers.

[5] Gediminas Adomaviciu and Alexander Tuzhilin, "**Expert driven, validation of Rule-Baed Models in Personalization"** Data Mining and Knowledge - Discovery Vol 5, No 1 / 2 Kluwer Academic Publishers, April 2001.

[6] Huan Liu, Farhad Hussain, Chew Lim Tan and Manoranjan Dash, "**Discretization: An Enabling Technique**", Data Mining and Knowledge Discovery", vol. 6 No. 4, Kluwer Academic Publishers, October 2002.

[7] Jiawei Han, Laks V. S. Lakshmanan and Raymond T.NG, "**Constraint-Based Multidimensional Data Mining**", IEEE Computer, August 1999.

[8] Ke Wang , Soon Hock William Tay, and Bing Liu, " **Interestingness-Based Interval Merger for Numeric Association Rules**", Fourth International Conference on Knowledge Discovery& Data Mining, New York,USA 1998

[9] Krzysztof J. Cios, Pedryez and Roman W. Surniarski, " **Data Mining Methods for Knowledge Discovery**", Kluwer Academic Publishers 1998 Second Printing 2000.

[10] Martin Staudt, Anca Vaduva and Thomas c, **"Metadata Management and Data Warehouse"**, Technical Report, Information System Research, Swiss Life , University of Zurich, Department of Computer Science, July 1999

[11] Ming-Syan chen, Jiawei Han and Philip S. Yu, **"Data Mining: An Overview From a Database Perspective"**, IEEE Transactions on Knowledge and Data Engineering Vol. 8, No. 6, Dec. 1996.

[12] Rakesh A. grawal, **"Parallel Mining of Associations Rule**", IEEE Transactions on Knowledge and Data Engineering ,Vol 8 , No 6, Dec 1996.

[13] Ramakrishnan Srikant and Rakesh A. Grawal, "**Mining Quantitative Association Rules in Large Relational Tables**", Proc Sigmod '96, 6/96 Montreal Canada, 1996.

[14] Ramakrishnan Srikant and Rakesh A. Grawal, "**Mining Generalized Association Rules**", Proceedings of The '21st VLDB Conference", Zurich, Switzerland, 1995.

[15] Raymond T. Ng, Laks V. S. Lakshmanan, Jiawei Hon and Alex Pany, "**Exploratory Mining and Pruning Optimizations of Constrained Associations Rules**", SIGMOD98, Seattle, WA, USA.

[16] Ronald J. Brachman, "**The Process of Knowledge Discovery in Database**", Advances in Knowledge Discovery and Data Mining, AAA1 Press / The MIT Press, 1996.

[17] Ronald J. Brachman, Tom Khabaza, Willie Kloegsgen, Gregoy Piatetsky – Shadiro and Evangelos Simoudis, "**Mining Business Databases**" Communications of the ACM, November 1996, Vol. 39, No. 11.

[18] Tomaz imielinski, A Ashu Virmani , **"MSQL : A query language for Database mining, Data mining and knowledge discovery**, Data Mining and Knowledge Discovery", Vol. 3 No. 4, Kluwer Academic Publishers, Dec 1999.

[19] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "**The KDD Process For Extracting Useful Knowledge From Volume F Data**", Communication of ACM, Nov 1996 / Vol. 39, No. 11.

[20] Ussama M. Fayyad, Gregory Piatetsky – Shapiro and Padhraic Smylh, "**From Data Mining to Knowledge Discovery, an Overview**", Advanced in Knowledge Discovery & Data Mining, AAA1 Press / The MIT Press, 1996.

[21] Valery Guralnik, Dumindo Wijesekera and Jaideep Srivastava, "**Pattern Directed Mining of Sequence Data**", 4th International Conference on Knowledge Discovery & Data Mining, August 1998, New York.

[22] Vasant Dhar, Dashin Chou and Foster Provost, "**Discovering Interesting Patterns For Investment Decision Making With Glower–Agenetic Learner Overlaid With Entropy Reduction**", Data Mining and Knowledge Discovery Vol. 4, No. 4, Kluwer Academic Publishers, October 2000.

[23] Wei-Min Shen and Bing Leng, "**A Meta-pattern Based Automated Discovery Loop For Integrated Data Mining–Unsupervised Learning of Relational Patterns**", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, Dec. 1996.

[24] Weiyang Lin, Sergio Alvarez and Carolina Ruiz, "**Efficient Adaptive Support User Association Rule Mining for Recommender System**", Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Jan. 2002, Vol. 6, No.1.

[25] Y. Balaji Padmanabham and Alexander Tuzhili, "**A Belief Driven Method for Discovering Unexpected Patterns**", the 4th International Conference Knowledge Discovery and Data Mining, August 1998.

[26] Y. Gauten Das, King-Ip Lin, Heikki Mannila, Gopal Renganathen and Padhrik Smyth, "**Rule Discovery From Time Series**", Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining, New York, USA November 1998.

[27] Z. S. Leigh, "**Underwriting - A Dying Art**", Journal of The Institute of Actuaries, Vol. 117, part III, The Alden Press Oxford, 1990.