Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Life Insurance Recommender System Based on Association Rule Mining and Dual Clustering Method for Solving Cold-Start Problem

Abdhesh Gupta
Department of Information Technology
Medicaps Institute of Science and Technology-Indore, India

Anwiti Jain
Department of Computer Science
Medicaps Institute of Science and Technology-Indore, India

Abstract— Recommender system based on web data mining is very useful, more accurate and provides worldwide services to the user. Recommender systems are becoming very popular in recent years. In this paper a web recommender system is proposed for life insurance sector based on web data mining using association rule which supports the insurance needy as well as life insurance representative to select best suitable life insurance plan for any particular person. Traditional recommender systems have being replaced by web mining techniques, but for new profile customers these recommender systems are not suitable, this is known as cold-start problem. In this paper we also proposed a solution for cold-start problem.

Keywords— Web data mining, Association rule mining, Apriori Algorithm, Cold-start, First-rater and Life insurance recommendation system.

I. INTRODUCTION

Data mining can be defined as the process of selecting, exploring and modelling large amount of data to uncover previously unknown patterns. In the insurance industry, data mining can help firms grow business advantage. For example, by applying data mining techniques, companies can fully develop data about customers buying patterns and behaviour- as well as gaining a greater accepting of their business to help minimize fraud, improve underwriting and increase risk management. In this paper we explore the data mining technique for recommendation system using association rule mining with some improvement in traditional recommendation system. In this paper we also discuss and proposed solution for new customers that how to acquire new customers information into system and get best recommendation for new customers. This is known as Cold-start problem [3].

Traditional recommendation methods can be classified into two main categories [4], Collaborative filtering and content-based approach. Collaborative filtering techniques guess product preferences for a user based on the opinions of other users. The opinions can be obtained openly from the users as a rating score or by using some inherent measures from purchase records as timing records [5]. There are two approaches for collaborative filtering, user-based also known as nearest-neighbours and item-based also know as model based algorithms. In the user-based method were the earliest used [6]. They treat all user items by means of statistical techniques in order to find users with analogous preferences. The advantage of these algorithms is the quick incorporation of the most modern information, but they have the inconvenience that the search for neighbours in large databases is slow [7]. Item-based collaborative filtering algorithms use data mining techniques in order to develop a model of user ratings, which is used to predict user preferences. In the content based filtering method is based on content learn from the target items i.e. items are recommended by comparing between their contents and user profile.

II. BACKGROUND

Association rule mining is one of the most important and well researched techniques of data mining, [1]. It aims to extract interesting correlations, frequent patterns, associations or informal structures in the middle of sets of items in the transaction databases or other data repositories. Association rules are broadly used in various areas such as telecommunication, networking, market and risk management, inventory control etc. Association rule mining is to find out association rule that assure the predefined least support and assurance from a given database. Association rule mining is usually divided into two parts. One is to find those item sets whose occurrences exceed a predefined threshold in the database, those item sets are called frequent or large item sets. The second part is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is Ln, Ln= {I1,I2,I3,….In}, association rules with this item sets are generated in the following way: the first rule is {{I1,I2,I3,….,In-1}}=>{In}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. The second part is simple, so our main research is on first part.

III. ASSOCIATION RULE MINING

Insurance companies can use association rules in market analysis. Here the data analyses consist of information about what policies customer purchases. The insurance company can generate association rules that show what different

policies are purchased with a specific policy. Based on these facts, company tries to capitalize on the association between different policies with the same company is much more likely to renew than is a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding a single less than three. By offering quantity discounts and selling bundled packages to customers, such as life security and investment policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

A database in which an association rule is to be found is viewed as a set of tuples, where each tuple contains a set of items. Here there are some transactions and five items;

1. Life security,
2. Market based,
3. Tax benefit,
4. Investment and
5. Retirement plan.

We need to find the frequent product set with strong relation by Association Rule Mining and recommend these products to web users to increase the cross-sales of life insurance sites.

According to Association Rule Mining, life insurance e-commerce can be displayed by following steps:-

A. *Prepare data source for mining*

In life insurance, there are basically three type's data sources for mining: User information database, Transaction database and Insurance plan database.

TABLE I: WEB DATABASE

User information Database		**Transaction Database**		**Insurance plan Database**	
Field Name	**Type**	**Field Name**	**Type**	**Field Name**	**Type**
UserId	Char	TransactionId	Numeric	PolicyId	Char
UserName	Char	PolicyId	Char	PlanName	Char
Gender	Char	UserId	Char	PlanDetials	Char
DateOfBirth	Date				
Income	Numeric				
Contact	Char				

B. *Apply Joins in Databases*

To apply Association Rule Mining, we need to join database as shown in table

TABLE III: WEB DATABASE

UserId	**InterestedPlan**	**TransactionId**
User1	1,3,4	01
User2	1,2,4	02
User3	2,3,5	03
User4	1,2,4,5	04
User5	1,2,3	05

C. *Find the frequent plan set and generate results*

A program is set to automatically mine the frequent item set by Apriori algorithm [2].

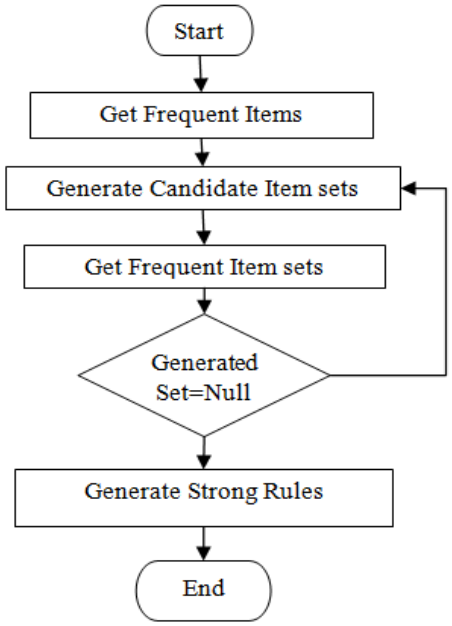

Fig.1Apriori Algorithm

Algorithm I: Apriori Algorithm

The Apriori Algorithm is the most well known association rule algorithm and is used in most commercial products.

Input:

Li-1 //Large itemsets of size i-1

Output:

Ci //Candidates of size i

Algorithm:

Ci=Ø;

for each I € Li-1 do

for each J L≠ € i-1 do

if i-2 of the elements in I and J are equal then

Ck=Ck U{I U J};

1) Scan the reorganized database to find the support of items in table III and calculate the first-level frequent item set.

TABLE III: FIRST LEVEL FREQUENT ITEM SET

Plan	1	2	3	4	5
Support	4	4	3	3	2
Frequent Plan	Yes	Yes	Yes	Yes	Yes

2) Calculate second level frequent plan set from second- level candidate set as show in table IV

TABLE IV : SECOND-LEVEL FREQUENT PLAN SET

Plan Set	{1,2}	{1,3}	{1,4}	{1,5}	{2,3}	{2,4}	{2,5}	{3,4}	{3,5}	{4,5}
Support	3	2	3	1	2	2	2	1	1	1
Frequent Plan	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No

3) Calculate third- level frequent plan set from third-level candidate set as show in table V

TABLE V : THIRD-LEVEL FREQUENT PLAN SET

Plan Set	{1,2,3}	{1,2,4}	{1,2,5}	{2,3,4}	{2,3,5}	{3,4,5}
Support	1	2	1	1	1	1
Frequent Plan	No	Yes	No	No	No	No

Result shows that {1, 2, 4} is the frequent plan set. If we keep calculating next level candidate plan set, a null set appears, the data mining process finishes. Plan 1, 2 and 4 exist strong association rules.

IV. COLD-START PROBLEM AND PROPOSED SOLUTION

The first-rater or early-rater problem arises when it is impossible to offer recommendations about a policy that was just incorporated in the system and, therefore, has few, or even none, evaluations from users. Analogously, such drawback also occurs about him, it would be impossible to determine his behaviour in order to provide him recommendations. Actually, this variant of the first rater problem is also referred as the cold-start problem [3].

An important business problem is the acquirement of new customers. In traditional approaches sales department make stratagem and put their efforts to increase customer base, we can also improve results of this sales efforts by more quantitative data mining approaches can lead to more focused and more successful results. A traditional sales approach we target a group of persons and try to convince them for taking policies. The main problem with this approach is that much of the marketing effort may not give good return. At the same point, sales become more difficult and greater marketing budgets lead to lower and lower returns. Hence in that case it is important to identify population group among already insured customers through which uninsured customers could be targeted. A statistical technique called "cluster analysis," sometimes used in the private sector to identify various market segments, was used to identify target groups of uninsured persons based on the previous available data of policy holders. Clustering is a technique of partitioning or segmentation the data into groups that might or might not be incoherent. The clustering usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters. Since clusters are not predefined, a domain expert is often required to interpret the meaning of the created clusters.

Insurance companies can create special catalogs targeted to various groups based on attributes such as income, occupation and age as physical characteristic of possible customers. The company then can perform a clustering of possible customers based on determined attribute values to create new catalogs. The results of the clustering exercise can be then used by management to create special catalogs for different policies and distributed them to the correct target population based on the cluster for that policy [2].

An insurance company can group its customers based on common features. Company management does not have any predefined for this. Based on the outcome of the grouping they will target marketing and advertising campaigns to the different groups for a particular type of policy.

TABLE VI : SAMPLE DATA

Age	Occupation	Income	Education
35	Employee	20,000/-	Graduate
25	Employee	10,000/-	Graduate
55	Employee	30,000/-	Post-Graduate
45	Employee	40,000/-	Post-Graduate
40	Business	70,000/-	Graduate
35	Business	80,000/-	Graduate

For example (Refer table VI) we have customer following information Age, occupation, income and education. Here we use dual clustering method. There are some limitations in single clustering. For example suppose advertising only for policy of Life security, we could target the customers having less income and occupation as employee. Hence the first group of people, is of younger employees having college degree, is suitable for Life security policies. The second group has higher qualification and also higher income is suitable for tax benefit policies, while last group has businessmen with higher income but low qualification and is suitable for investment policies [2]. As shown in table VI three clusters are created according to their occupation and education. If new customer arrives to the recommendation system then we take salary as a cluster parameter and according to salary we select average value of each cluster and put new customer in nearest cluster, then we determine policy which is mostly preferred by cluster members and same policy will be recommended to the new customer.

But if new customer salary is nearest to the more than one cluster average salary for example cluster 1 average salary is 15,000/- and cluster 2 average salary is 35,000/- and new customer salary is 25,000/-, in that case it is difficult to consider customer in a single cluster. To solve this problem we can use dual cluster method. In dual cluster method on the basis of only one parameter (salary) if we cannot find single nearest cluster then we use second parameter for cluster like in this example we can take age as a second parameter. By taking age as another parameter for clustering we can decide final single cluster for new customer and recommend appropriate policy to the new customer.

V. CONCLUSION

In the insurance industry, web data mining can help firm gain business advantage mainly to support decision making. The insurance companies need to know the essentials of decision making and web data mining techniques to compete in the market of life insurance. In this work a web recommendation framework specially address to overcome critical recommendation system problem. In this work some high level association rule mining method is used to retain existing customer for new policy. Clustering method is used to attract and recommend policy to new customer (cold-start problem). Dual clustering method is used to overcome the limitation of single clustering method which gives more accurate and appropriate recommendation to solve cold-start problem.

REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, *A. Data Mining: A performance Perspective*. IEEE Trans. Knowledge and Data Engineering, vol, 5,6, 1993a, pp. 914-925.

[2] A.B. Devale, Dr. R. V. Kulkarni *Applications of Data Mining Techniques in Life Insurance*. International Journal of Data Mining and Knowledge Management Process Vol.2, No.4 July 2012.

[3] Maria N. Moreno, Saddys Segrera, Vivian F Lopez, Maria Dolores Munoz and Angel Luis Sanchez, *Mining Semantic Data for Solving First-rater and Cold-start Problems in Recommender system* ACM IDEAS 11 2011, September 21-23.

[4] Lee, CH., Kim, Y.H., Rhee, P.K. 2001. *Web personalization expert with combining collaborative filtering and association rule Mining Technique.* Expert System with Applications 21. 131-137.

[5] Sarwar, B., Karypis, G., Konstan, J., Riedl, J. 2001. *Item-based Collaborative Filtering Recommendation Algorithm*. Proceedings of the tenth International World Wide Web Conference, 285-295.

[6] Resnick, P., Lacovou, N., Suchack, M., Bergstrom, P. and Riedi, J. 1994. Grouplens: *An open architecture for collaborative filtering of netnews.* Proc of ACM Conference on Computer Supported Cooperative Work, 175-186.

[7] Schafer, J.B., Konstant, J.A. and Riedl, J. 2001. *E-Commerce Recommendation Applications*. Data Mining and knowledge Discovery, 5, 115-153.

