PREDICTION OF CUSTOMER BEHAVIOR USING CMA

F. Fenita
Department of Computer Science and Engineering
Karunya University, Coimbatore, India
Email id: fenita87@gmail.com

Sumitha C.H
Department of Computer Science and Engineering
Karunya University, Coimbatore, India
Email id: sumithach@karunya.edu

ABSTRACT:

The process of extracting certain sequential pattern that exceeds the minimum support threshold is called Sequential Pattern Mining. It has been used to predict various aspects of customer buying behavior for a long time. The sequence that is discovered using sequential pattern mining algorithm reveals the chronological relation between the items and provides valuable information which will help in developing the marketing strategies. There are various approaches proposed for sequential pattern mining. But in the existing sequential pattern mining approaches, the cyclic behavior of the buying pattern and the interval between the two consecutive items is hardly known. The proposed algorithms can be implemented and supplied to test against real time customer data from the consumer goods company. The experimental results illustrate how the model can be used to predict likely purchases within a certain time limit.

Keywords: Apriori, FreeSpan, MEMSIP, PrefixSpan, SPIRIT, Trend Modeling, Trend distribution function.

1 INTRODUCTION

A pattern is a type that predicts the manner in which the set of events occur. The process of extracting the information about such pattern is called as Pattern Mining. Sequential pattern is a sequence of item sets that occurs frequently in a specific order. All the customer transactions viewed together as a sequence is called as customer-sequence where each transaction is represented as an item sets in that sequence. Given a set of sequences, a complete set of Frequent subsequences can be mined using Sequential
Pattern Mining. The applications of Sequential Pattern Mining are customer shopping sequences, Telephone calling patterns, natural disasters like earthquakes etc. The process of knowledge discovery in database consists of three steps. The first one is called preprocessing which includes data cleaning, integration, selection and transformation. Secondly, data mining which is the main process where different algorithms are applied to produce the knowledge which are hidden. Thirdly, the post processing process which evaluates the mining results according to the user’s requirements. The knowledge can be presented only if the results are satisfactory. Different knowledge comes out as a result when various data mining techniques are applied to the data sources. Depending on the evaluation results obtained, mining can be redone or the user can modify the requirements.

2 SEQUENTIAL PATTERN MINING

The process of finding out the inter-transaction pattern in a given set of sequence is called as sequential pattern mining. The sequential pattern mining algorithms are used in finding out the complete set of frequent sub-sequences in the given set of sequences. These algorithms are efficient in the case of sequential patterns which are too long. The discovery of sequential pattern in transaction data is very important in many application domains especially in customer analyses where certain buying patterns such as follow-up purchases can be determined. Thus, discovered sequential patterns from the super markets are used to develop the marketing strategies. Sequential pattern is a sequence of item sets that occurs frequently in a specific order. All the customer transactions viewed together as a sequence is called as customer-sequence where each transaction is represented as an itemset in that sequence. The customer support a sequence s, if s is contained in the customer sequence, then the support of the sequence s is defined as the fraction of customers who support the sequence.

\[
\text{Support}(S) = \frac{\text{Number of support customers}}{\text{Total number of customers}}
\]  

(1)

The problem of discovering the sequence was first introduced by Agrawal and Sirkant [2]. Many algorithms were developed later and successfully improved the efficiency of the task of mining sequential patterns. The most basic and earlier algorithms are based upon Apriori algorithm [1]. The core of the Apriori property is that any sub
pattern of a frequent pattern must be frequent. Based on this heuristic, a series of Apriori-like algorithms such as AprioriAll, AprioriSome, DynamicSome [8], and GSP [5] were developed.

Later on, different kinds of algorithms are proposed by different researchers. For example, FreeSpan [3] and PrefixSpan [7] were developed by the data projection approach. Han et al. [6] proposed two algorithms for mining partial periodic pattern single period and multiple periods. Yanget al. [11] proposed distance-based pruning to find the periodic patterns, which may contain a disturbance of length up to a certain threshold. The mining of frequent partial periodic sequential patterns in a time series is to find possibly with some restriction or disturbance. Ozden et al. [9] proposed the sequential algorithm and interleaved algorithm to determine cyclic association rules. Associate rules capture interrelationships between various items.

The approaches presented above primarily target the orders of items purchased, or cyclic patterns occurring within a particular time frame defined by users. However, by these works, the cyclic behavior and the interval of the items purchased is hardly known. Hence the Data mining skills and the fundamentals of statistic are combined to introduce an algorithm Cyclic Model Analysis (CMA) to find out the model of recurring purchasing. The modeling process starts with the discovery of sequential patterns from the transactional database. From the discovered sequence, the trend distribution function (TDF) is calculated. Then the existence of periodicity and the interval of successive events are identified by the Generalized Periodicity Detection (GPD) and Trend Modeling (TM) computed, which will be explained in more detail later. Next, the CMA algorithm is used to obtain the period and trends of quantities of purchasing. Consequently, marketing people can recommend the right products to the right customers at the right time.

3 TREND DISTRIBUTION FUNCTION

The distribution function is a time series representation of the transactions associated with the sequence discovered. The traditional sequential formulation reveals the chronological order of purchase only. If the plot of the function is sketched, the movement along the curve shows the tendency of the purchase. For any given subset of the domain of the function, a simple linear regression is used to construct the regression
line of the function within the sub-domain. If the slope of the obtained straight line is a positive number, the purchase of the item increases in a certain rate. If the slope is a negative number, the sales volume of the item decreases after a certain point is reached. If f(x) is linearly decreasing distribution function and reaches zero within the domain, the point x is called the degeneration point of the function. The degeneration point signifies the fact that the customers tend to stop purchasing. Hence, the business owner must be alerted before the degeneration point is reached.

Once the frequent itemsets are generated the trend distribution function is calculated. First the trend distribution function for one itemset that is itemset containing only one item is calculated. The entire database is divided into four tables. The first item from the generated frequent item set is taken and the count of occurrence of that particular item is checked in the tables. The difference in the count value is calculated and it is incremented by one. The distribution calculated can be of three types. If the obtained value of the distribution function is positive then it is called as ascending type of distribution. If the obtained value of the distribution function is negative then it is called as descending type of distribution. If the obtained distribution function is zero, then the trend is constant that occurs more frequently in real world scenarios. Similarly the trend distribution function is calculated for two itemsets and three itemsets and so on. Once the trend distribution is calculated, these values can be given as inputs for further procedures.

4 GENERALIZED PERIODICITY DETECTION

The scheme presented in this paper takes a two-phase approach to cope with all periodicity-related problems, which occur in the analysis process of sequential pattern mined from transactions. It is well known that a host of algorithms have been developed for efficient mining of sequential patterns. To solve the periodicity problem, a mathematical model constructed to portray the sequential pattern mined from the database.

The scheme comprises the sequential pattern mining technique and the algorithms presented in this section. Once the result of sequential pattern mining is given, the primary concern is to know where there are regularities that can be found. Thus, the value of trend distribution function is computed and then the GPD is introduced to detect the periodicity of the function. If the periodicity can be identified, the analysts can have a
better understanding of the patterns and decide on the next appropriate action. After the value of trend distribution function is determined the GPD is introduced to detect the periodicity of the function.

The trend distribution function values, the number of elements, the minimal period which is the smallest possible value of the period and the maximum error which is the error threshold used to judge if the investigated value is subject to the generalized periodicity are given as input to the GPD algorithm. The output generated from this algorithm will be the list of periods and empty if the periodicity does not exist. The first step is to find the linear model $y=ax+b$ of the given function $f$. If the error $(x_i)$ is less than the maximum error threshold predefined by the user, then the value $x_i$ is the period prospect. The list of periods and the trend distribution function values are given as the input to the Trend Modeling (TM) algorithm.

$$\text{Error}(\lambda) = \frac{\sum_{i=1}^{n-1} |h(x_{i+1})-h(x_i)|}{\sum_{i=1}^{n} |h(x_i)|}$$

5 TREND MODELING

Once the existence of periodicity has been identified, the characteristic of the sequence can be captured more precisely. Hence, a mathematical structure, which symbolizes and describes the transactions occurring in the real world, is needed. The results of the modeling process must map the relationships between relevant attributes in the transaction databases to those relationships between relevant attributes of the function obtained by computing rules. Broad movements evolve more gradually than other movements which are evident. These gradual changes are called trends. The changes which are transitory in nature are described as fluctuations.

Since regression analysis is frequently used for fitting equation to data, regression techniques are applied to construct Trend Modeling procedure. TM algorithm is straightforward and begins with the establishment of simple linear regression model $y=ax+b$. Combined with the result of GPD, the complete model characterizing the given input distribution function is obtained. The polynomial gained by the TM process is an aid to identify the nature of the sequence mined. To illustrate how Trend Modeling is used to find the approximating model of a given input, two typical linearly increasing/decreasing periodic functions are used as sample input.
6 CYCLIC MODEL ANALYSIS

The purpose of the establishment of the mathematical model is to help analysts obtain a better understanding of the whole picture of what happened and predict what is likely to happen. With GPD and TM it can be determined if customers tend to repeat buying at regular period and an equation can be formulated to approximate the distribution obtained from the sequence. In short, the mathematical model established by GPD/TM is used to describe characteristics of the sequential pattern mined from designated time frame. Next, CMA is proposed to analyze and describe the characteristics of the sequential pattern mined from the transactional databases. The user must determine the value of parameters minimum period, maximum error, degree and terminating condition.

The value of the terminating condition is the terminating condition of the process. If the length of the domain is too short, the process should be stopped since its meaningless to investigate the characteristic of repeated patterns. Then the type of the function is determined by finding the local maximum of the function. If the local maximum exists at the end of the domain of the function, the function belongs to ascending type.

The descending type can be determined if the local maximum exists at the beginning of the domain. If the distribution function is ascending or descending type,
then apply GPD/TM directly to get the polynomial approximating the patterns and find the period of the distribution function. If the distribution is neither descending nor ascending type, the whole time frame is partitioned into two sub frames and invokes CMA recursively until the distribution function of the sub frame is simplified. If the length of the inspected sub frame is smaller than the predefined terminating condition trcd, the process will be stopped. CMA performs well in exploring the trends of repeat-buying behaviors and provides practical model for predicting the customer buying behavior.

The process of applying CMA to real-world is displayed in Fig 6.1. Once the mining process is completed, the sequential patterns to be investigated were selected. The distribution function of the designated pattern is computed. Then, the CMA is applied to capture the model of the mathematical model. CMA divides the domain into smaller segments such that the distribution function defined on the segment is of simply increasing or decreasing type.

Then GPD and TM were invoked to obtain the polynomial approximating to the pattern defined on the segment. The result of divide-and-conquer process will be collected and synthesized. Then this synthesized knowledge will be used to help analyze past and likely future behaviors. From the collected knowledge, marketers can predict the customer buying behavior. The consistency of the repeat-buying behavior over time and the item-based feature of the CMA algorithm suggested that a hybrid recommendation system can be formed to provide better prediction. Thus, marketing professionals will have a better tool with which they retain their customers. Consequently, the marketers can take appropriate action to favorably impact customers’ behavior.

7 ENHANCEMENTS

In the cyclic model analysis algorithm, when mining the frequent itemsets, so many frequent itemsets will be pruned. Hence this is the disadvantage in this technique. This can be improved by using fuzzy techniques. Multi-valued logic derived from fuzzy set theory to deal with reasoning that is approximate rather than accurate is called fuzzy logic. In "crisp logic", where binary sets have binary logic, fuzzy logic variables may have a truth value that ranges between 0 and 1. Though fuzzy logic has been applied to many fields, from control theory to artificial intelligence, it still remains controversial
among most statisticians. Fuzzy logics take on a range of values High, medium and low rather than always taking true or false values. This prevents in loss of itemsets using pruning. Fuzzy values are not probabilities. Fuzzy logic and probability are different ways of expressing uncertainty. But both fuzzy logic and probability theory can be used to represent subjective belief, fuzzy set theory uses the concept of fuzzy set membership (i.e., number of variable in a set), probability theory uses the concept of subjective probability (i.e., how probable is a variable in a set). While this distinction is mostly philosophical, the fuzzy-logic-derived possibility measure is inherently different from the probability measure, hence they are not directly equivalent. Hence the fuzzy logics can be used to improve the existing techniques.

8 CONCLUSION

In this paper some of the existing sequential Pattern Mining techniques are compared and analyzed. The cyclic model analysis on sequential pattern mining technique is more advantageous than the other approaches since this the only method which helps in revealing the cyclic property of the sequential patterns. CMA performs well in describing the buying behavior of the customer. However, there are some disadvantages in this technique. First, the periodicity detecting procedure can be improved by applying fuzzy techniques and other statistics tools. Fuzzy logics are used to prevent the loss of frequent item sets during the pruning phase thereby improves the marketing strategies.

REFERENCES

[5] Yu Hirate and Hayato Yamana,” Generalized sequential Pattern Mining with Item

[6] J. Han, G. Dong, Y. Yin, "Efficient Mining of Partial Periodic Patterns in Time Series

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu,
“PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern

Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 487-499,
1994.


[10] J. Yang, W. Wang, P.S. Yu, and J. Han, “Mining Long Sequential Patterns in a

Cyclic Model Analysis on Sequential Patterns” IEEE Transactions on Knowledge and