A Purchasing Sequences Data Mining Method for Customer Segmentation

Hai Wang, *Member, IEEE,* and Shouhong Wang

***Abstract*—Purchasing behavior serves a base for online customer segmentation. Online purchasing behavior is characterized by purchasing sequences. This paper reviews the existing three major techniques of sequence data analysis, and discusses their limitations in online purchasing sequences analysis for customer segmentation. The study proposes a new data mining method for online customer segmentation, and applies this method for an online nutrition product store. The data mining results indicate that the proposed data mining method is novel and effective for online customer segmentation.**

***Index Terms*—Data Mining, Sequence Data Mining, Customer Segmentation.**

I. Introduction

Customer segmentation is to identify target groups of customers and develop effective marketing strategies for the business organization [8]. There have been many customer segmentation approaches [1]. However, purchasing sequences based customer segmentation methods are underdeveloped [6]. In fact, the purchasing sequences based customer segmentation approach is particularly powerful in e-commerce where demographic factors are hidden (e.g., income and age) or simply irrelevant (e.g., geographic).

Sequence data analysis has been studied in the field of data mining for years [4]. In the data mining field, sequence data mining is to discover patterns of sets of sequences. For example, the purchasing history of an online shopper is a chronic sequence of the purchase records. The sequence represents the customer's online shopping behavior. If we have thousands such sequences, we can discover the several major patterns of these sequences. These patterns can be useful for us to understand the customer segments based on the online customers' purchasing sequences.

The first author is supported by Grant 312423 from the Natural Sciences and Engineering Research Council of Canada (NSERC).

H. Wang is with Sobey School of Business, Saint Mary's University, Halifax, NS, Canada B3H 3C3 (phone: 912-496-8231; fax: 912-496-8101; email: hwang@smu.ca).

S. Wang is with the Department of Marketing/Business Information Systems, Charlton College of Business, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA (email: swang@umassd.edu).

This paper proposes a data mining model for analyzing online purchasing sequence data, and applies this model to a case of online purchasing sequences based customer segmentation. The remainder of the paper is organized as follows. First, we provide a brief overview of the commonly used analytical techniques on sequence data. Next, we discuss the limitations of these techniques in purchasing sequences data mining, and develop a model for online customer segmentation. Then, we present an application of the proposed model to a case of customer segmentation for an online nutrition product store. Finally, we conclude with a summary of the study.

II. Analytical Methods for Sequence Data

A sequence is a set of ordered elements. There have been many techniques for sequence data analysis. Here we give a brief overview.

A. Time Series

Time series are sequence data measured strictly against the time dimension. Because of this property of time series, approaches to time series analysis are structured, in contrast to data mining techniques for general cases of sequences where time is not a key factor (e.g., DNA sequences). Although there have been a variety of methods for time series analysis, the time window method is typically used to describe the pattern of the time-dependent variable (e.g., stock price). A classical time series analysis method is the linear autoregressive moving average (ARMA) model [2] expressed as follows:

$$x_t = a + \sum_{i=1}^{p} b_i x_{t-i} + \sum_{j=1}^{q} c_j x_{t-j} + e_t \qquad (1)$$

where t is time; x_t is the value of the time-dependent variable at time t; i and j are window indices; a, b_i, and c_j are regression parameters, and e_t is the residual term.

In general cases of sequence data where time is not necessarily a useful static reference dimension, the time series analysis methods, such as the ARMA method, are not particularly useful for sequence data analysis. Also, existing time series analysis methods are powerless in dealing with cases of multiple dependent variables, non-numerical data, and sequence pattern analysis, compared with the methods discussed below.

1-4244-0318-9/06/$20.00 ©2006 IEEE

B. Association Distance Measure

According to the associate distance measure (ADM) method, a sequence is a vector [3]. In the ADM method, the difference between two sequences S_1 and S_2 is measured by the Euclidean distance between the two vectors.

$$d = \sum_{i=1}^{n} f_i \quad (2)$$

where i is the position index of the longest sequence, n is the number of the elements of the longest sequence, f_i is the dissimilarity of the elements at position i:

$$f_i = 1 \quad \text{if } S_1(i) \neq S_2(i)$$
$$f_i = 0 \quad \text{otherwise.} \quad (3)$$

Missing values are treated as dissimilarity. In terms of the equivalence between a sequence and a vector, the order of the elements in a sequence is not important. This marks the limitations of the ADM method.

C. Sequence Alignment Method

According to the sequence alignment method (SAM), a sequence is a set of elements arranged in a certain order [7]. The SAM method also uses a distance measure to calibrate the similarity between sequences. However, in the SAM method, the similarity between two sequences is measured by the necessary operations to covert one sequence to the other. A general formula to calculate the distance between two sequences is

$$d = w_d D + w_i I + w_r R \quad (4)$$

where D is the number of deletion operation, I is the number of insertion operation, R is the number of reordering operation, and w_d, w_i, and w_r are predetermined weights for deletion, insertion, and reordering operations respectively. For instance, suppose that we have the following two sequences:

S_1: {a, c, d, m, p, y}
S_2: {a, d, c, m, t, y}

We can convert S_1 into S_2 by applying the following operations:

(1) Reorder the second and the third elements.

(2) Delete the fifth element.

(3) Insert "t" to the fifth element.

If we choose $w_d=w_i=2$ and $w_r=1$, the distance between S_1 and S_2 is 5.

Compared with the ADM method, the SAM method takes the order of the elements into account in measuring the distance between the sequences. Nevertheless, the general formula for distance calculation might be over simplified for a real case. Also, the time measure in the sequence is missing in the SAM method.

In summary, there have been several popular techniques for sequence data analysis. In the business field, time series analysis methods have been used for sequence analysis with a single time sensitive variable during the past three decades, but have their limitations in general business sequence pattern analysis. Although there have been applications of the SAM and ADM methods for mining Web navigation sequence patterns [5] and customer segmentation [6], the ADM and SAM methods do not represent the time measure explicitly. Accordingly, data mining techniques beyond the existing methods must be developed to analyze sequence data where patterns are the major concern and time is a critical factor.

III. MINING PURCHASING SEQUENCE DATA

A. Online Purchasing Sequences

Internet provides a rich environment for data collection for online transactions. Online purchasing sequences (OPS) are the major business intelligence sources of online consumers' purchasing behaviors, and can be used for marketing purposes. OPS are not typical time series, because time is a spontaneous reference factor here. On the other hand, the time aspect is important for us to understand the online shoppers' behaviors, and ought to be modeled in the analysis.

An OPS is a set of transactions <P, t>, where P is a purchase event, and t is the time of the purchase event. The formal expression of OPS is

$$OPS_k = \{ \ <P_1, t_1>, \ \ldots <P_i, t_i>, \ \ldots <P_n, t_n> \ \} \quad (5)$$

where k is the recorded customer's number, $<P_i, t_i>$ is the purchase transaction. Using a formal description,

$$OPS ::= \varnothing \mid <P, t> \mid \{<P, t>, OPS\} \mid \quad (6)$$

For example, using symbols and numbers to encode the event occurrences, a purchase sequence of an online shopper can be described in the OPS form such like

$$OPS_{234} = \{ <35, 100>, <40, 105>, \ldots <38, 124> \}$$

Here, P is purchase money value, and t is day. P can also be a package of items purchased, and can be further decomposed into purchased items in the purchasing event.

$$P = \{ \ p_1, \ \ldots p_j, \ \ldots p_m \ \} \quad (7)$$

Decomposition of P would make the sequence analysis more meaningful. An approach to decomposition of P for sequence analysis is to decompose one event into several simultaneous events, as follows.

$$<P_i, t_i> = <p_{i1}, t_i>, \ \ldots <p_{ij}, t_i>, \ldots <p_{im}, t_i> \quad (8)$$

B. OPS Mining for Customer Segmentation

Since they provide the patterns of customers' purchase behaviors, OPS are sources of data mining for customer segmentation. To start the data mining, the miner designs a set of seed OPS (SOPS). Each SOPS is considered to be an ideal or desired online shopper with unique purchasing patterns. Typical patterns for ideal regular customers are discussed below.

Regular customer:
$SOPS_1$ = { <p, t>, <p±a, t±b>, <p±a, t±2b > }
where a and b are the pre-determined increments of p and t respectively.

Regular customer with lower purchasing frequency:
$SOPS_2$ = { <p, t>, <p±a, t±2b>, <p±a, t±4b > }

Regular customer with higher purchasing values:
$SOPS_3$ = { <p+d, t>, <p+d±a, t±b>, <p+d±a, t±2b > }

Customer who is responsive to promotions:
$SOPS_4$ = { <p, t>, <p±a, c_1±b >, <p±a, c_2±b > }
where c_i is coupon expire times.

Clearly, SOPS is usually not just one, but should not be too many. They may originally come from the miner's *a priori* knowledge, and could be updated based on the real OPS.

The parameters (i.e., p, a, b, c, and d in the above typical patterns, as well as the number of purchases contained in the SOPS) are also highly depending on the specific application. Practically, the setting of the values of parameters a and b would affect the result segment sizes. The larger the value of a or b is, the larger the size of the result segment would be. The data miner can also set a segment size for a group by adjusting parameters a and b to observe the magnitude of these parameters.

The SOPS updating process itself is important, and can be viewed as a data mining task. However, this paper places the focal point on mining OPS against a given set of SOPS, and leaves the issue of SOPS migration to future studies.

OPS are usually irregular. In mining OPS, it is useful to see the dissimilarity between the actual OPS and the SOPS. This sub-section describes an OPS analysis method that matches an OPS in the OPS database against a SOPS. A simplified high level description of the OPS matching algorithm is listed as follows.

1. Define SOPS and corresponding parameters for each SOPS.
2. For customer k=1 to K
   {
   Identify OPS_k form **S**, the set of OPS for data mining.
   For segment g=1 to G
   {
   Select $SOPS_g$ from **s**, the set of SOPS.
   If ($SOPS_g \subseteq OPS_k$) Then add customer k to group g.
   }
   }
3. The G groups of customers are the segments.

List 1. The OPS Matching Algorithm

This algorithm is rather pragmatic. For instance, the up-to-date version of OPS and the definitions $SOPS_g \subseteq OPS_k$ would include subjective criteria of the data miner, based on the problem context and other information available to the data miner. These subjective criteria reflect the perception of the quality of customers. The above OPS matching algorithm has been implemented in PROLOG which has powerful logic reasoning abilities.

IV. THE CASE OF ONLINE NUTRITION PRODUCTS STORE

In this section we describe the use of the proposed data mining method to discover customer segments for an online nutrition products store. The US based online store has been profitable for the past seven years. It sells food supplement products, herb, diet products, and sport nutrition products. The OPS used in this case study were purchase transaction logs. These transaction logs were recorded by the server of the Web site. In our experiments the customer log-in ID, purchase date, and purchase values (before tax and after shipping costs) were included in OPS. 2235 valid customer OPS occurred in the past 2 years were collected for the data mining experiments for customer segmentation. SOPS were generated in three dimensions: purchase value, purchase frequency, promotion responsiveness. A part of the data mining results is exhibited in Table I.

TABLE I
A PART OF DATA MINING RESULTS

		a= 20 b= 30	a= 40 b= 60
Static (Over 2 years)	Regular Segment	8.1%	23.4%
	Promotion Responsive Segment	18.6%	20.3%
		1st year	2nd year
Dynamic (a=20)	Regular-Higher Frequency Segment (b=30)	7.3%	8.7%
	Regular-Lower Frequency Segment (b=60)	13.1%	19.0%
	Promotion Responsive Segment (b=30)	15.2%	21.5%

Interesting points beyond the specific mining results were

observed about our proposed method. First, the proposed data mining method is able to reveal sensitivity information about the customer segments. For example, the values of parameters a and b show how sensitive of purchase values and purchase times would be for the segmentation. Second, generally, the customer segments were quite dynamic. Significant numbers of customers might change from one segment to another. This purchasing sequences analysis method is able to reveal the dynamics of customer segments, and would be useful for marketing strategy design.

V. Conclusions

This paper discusses the importance of purchasing sequences in customer segmentation. It provides an overview of commonly used analytical techniques for sequence data. The traditional time series analysis methods are useful only for analysis of sequence data with a single time-sensitive variable, but powerless for sequence data with multiple variables and non-numerical data. The ADM and SAM methods are useful for analysis of non-numerical sequence data, but ignore the timing property of sequences. OPS can be numerical or non-numerical, and are all time-sensitive. To facilitate data mining on OPS, the paper has proposed a new data mining method for OPS analysis. This method is to match OPS in the online purchase transaction database against the SOPS, and reveal customer segments. We have applied this data mining method to an online nutrition store for customer segmentation. Based on the experiments, we conclude that the OPS data mining method is promising for e-commerce customer management.

References

[1] R. Bagozzi, *Advanced Methods of Marketing Research*, Oxford, UK: Blackwell Scientific Publications, 1994.

[2] G. E. P. Box and G. M. Jenkin, *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day, 1970.

[3] B. Everitt, *Cluster Analysis*, New York, NY: Halsted Press, 1980.

[4] U. Fayyad, D. Haussler, and P. Stolorz, "Mining scientific data," *Communications of the ACM*, vol. 39, no. 11, pp. 51-51, 1996.

[5] B. Hay, G. Wets, and K. Vanhoof, "Mining navigation patterns using a sequence alignment method," *Knowledge and Information Systems*, vol. 6, pp. 150-163, 2004.

[6] C. H. Jon, H. J. P. Timmermans, and P. T. L. Popkowski-Leszczyc, "Identifying purchase-history sensitive shopper segments using scanner panel data and sequence alignment methods," *Journal of Retailing and Consumer Services*, vol. 10, pp. 135-144, 2003.

[7] D. Sankoff, and J. B. Kruskal, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Reading, MA: Addison-Wesley, 1983.

[8] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, Norwell, MA: Kluwer Academic Publishers, 2000.