Consumer Behavior Analysis by Graph Mining Technique

Katsutoshi Yada[1], Hiroshi Motoda[2], Takashi Washio[2], and Asuka Miyawaki[1]

[1] Faculty of Commerce, Kansai University, 3-3-35 Yamate-cho, Suita, Osaka 564-8680, Japan
{Yada, da10645}@kansai-u.ac.jp
[2] Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
{motoda, Washio}@ar.sanken.osaka-u.ac.jp

Abstract. In this paper we discuss how graph mining system is applied to sales transaction data so as to understand consumer behavior. First, existing research of consumer behavior analysis for sequential purchase pattern is reviewed. Then we propose to represent the complicated customer purchase behavior by a directed graph retaining temporal information in a purchase sequence and apply a graph mining technique to analyze the frequent occurring patterns. In this paper we demonstrate through the case of healthy cooking oil analysis how graph mining technology helps us understand complex purchase behavior.

1 Introduction

In Japan, "Health" related products have become a major focus of attention among consumers and industries in recent years. Because of the fact that health related products have a high added value it is understandable that food retailers who are groping for sales promotions that do not rely on price reduction to attract customers find that the consumer group that purchases said health related products is a very attractive one indeed. However, these retailers who are attempting to lure ever greater numbers of high value customers at their shops by proposing effective food menus find themselves faced with the difficult dilemma of trying to extract the characteristics of these consumers from their complex purchasing behavior. Simply analyzing the contents of these consumer's shopping baskets is not enough to shed light on their purchasing pattern and lifestyles. It is also necessary to draw out the characteristics of the relationships existing amongst groups of products and relationship amongst products when a multitude of products are bought at one time. In this paper we present the application by using graph mining technique to understand chance, health food boom to analyze purchasing historical data from the view point of consumer behavior and to create the trigger to discuss and communicate the future scenario.

Up to the present we have developed a variety of analysis methods that can be used to analyze the purchase history data of consumers. For example, we developed

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 800–806, 2004.

E-BONSAI [1][2] which makes use of sequence analysis technique as a way to carry out temporal sequential analysis of categorical data and were able to successfully extract the characteristics found in consumer purchasing patterns. However, although it is possible to analyze the sequential patterns of limited product groups for which E-BONSAI is designed, it is not possible to resolve the above-mentioned problems by sequential pattern analysis. It became necessary for us to extract the characteristics from the products purchased in multiple purchase patterns along with their purchasing sequence.

In our research we make proposals using characteristic patterns extracted from temporal sequences of purchased product groups that are represented as graph structured data. Graph structure is effective and useful to express complicated forms of data and phenomena. There have already been several algorisms for mining graphs and they are utilized to analyze chemical compounds and medical data. We believe that by applying these graph mining technique to the marketing field it will be possible to discover new implication that were not possible to detect by the traditional forms of technology. In this paper we apply graph mining technique to the FSP data of supermarkets in an attempt to discover new opportunities through interactions of retailers and experts employed by a variety of manufacturers.

2 Analyzing the Behavior of Consumers Using Graph Mining Technology

2.1 Graph Mining

Graph mining is a technique used to extract characteristic patterns from a variety of graph structured data [4]. The graph structure is a nice way of representing and explaining complex data forms and phenomena but because of its strong expressiveness its computational complexity has been a problem to extract specific patterns. However, recent development has made it possible to perform a complete search in extracting all the subgraph in a reasonable computation time. AGM algorithm [3] is one of the most advanced algorithms for graph mining and is able to deal with directed/undirected and colored/uncolored graphs. While graph mining research is still in the developmental stage there is a fair amount of research being carried out already concerning its practical applicability.

For example, graph mining is applied to extract patterns from chemical compound data. In chemical domain molecular structure of chemical substances has always been expressed using graph structures, and thus it is natural that molecular analysis is one of the most frequently challenged application area of graph mining. In fact graph mining technique successfully extracted meaningful substructures that cause carcinogenicity in organic chlorine compounds.

Although various types of work are currently being carried using graph mining, to the best of our knowledge, we are the first to apply graph mining technique to the business and marketing field. We have applied graph mining technique to POS data,

which includes customer ID that has been accumulated in the retailing industry, in order to investigate the possibilities of applying graph mining to the marketing research field.

2.2 Gh Structures and Consumer Behavior

Based on the POS data in Table 1 that include customer ID information, we have attempted to express the purchase behavior of consumers using graph structure. The data shown in Table 1 is a detailed sales record of a single consumer following the purchase of salad oil. After purchasing salad oil, the customer ID:1 visited the store 2 times and purchased multiple products from differing categories.

Table 1. Example of POS data including ID information

ID	Date	Category	Product code	Price zone
1	Jan. 27th, 2004	Salad oil	Healthy Okona	Expensive
1	Feb. 2nd, 2004	Eggs	Organic eggs	Expensive
1	Feb. 2nd, 2004	Milk	Unmei milk	Expensive
1	Feb. 2nd, 2004	Mayonnaise	Half&Herb	Average
1	Feb. 12th, 2004	Eggs	Organic eggs	Expensive
1	Feb. 12th, 2004	Bread	Brown rice bread	Expensive
1	Feb. 12th, 2004	Milk	Honeboso milk	Inexpensive
:	:	:	:	:
2	Feb. 3rd, 2004	Salad oil	Nisshin oil	Average
:	:	:	:	:

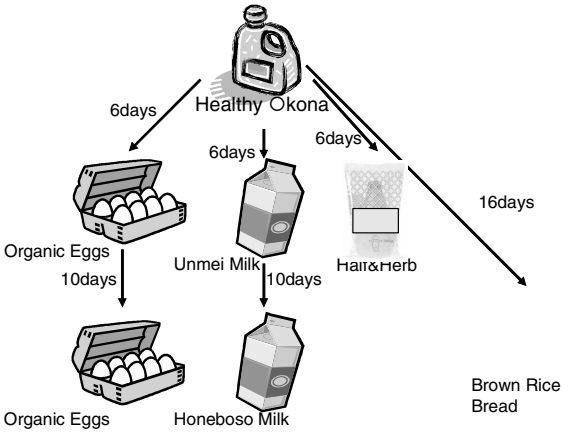

Fig. 1. Graph structured data and purchasing behavior

Using a graph makes it possible to express in extreme detail purchasing information concerning the composition of products as to when and where multiple products ware purchased at one time. Figure 1 shows the purchasing behavior of the customer ID:1 using a graph structure. The root of the graph shows the purchase of "Healthy Okona" salad oil. On the following visit to the store (6-days later) the consumer purchased "Organic eggs," "Unmei milk" and "Half & Herb." The arrows connecting these items represent the purchasing sequence. Next to the edges linking the respective products, labels are attached that indicate the intervals between said purchases. The "Organic eggs" and "Honeboso milk" which were purchased on the following visit to the store are connected by arrows with the products purchased in the same category on the previous visit, and these also have labels attached which indicate the number of days that have elapsed since the previous purchase date. Also, the categories for products such as "Brown rice bread" which were not purchased on the preceding visit are linked to the salad oil.

The data contained in this graph includes not only the information about simultaneous purchases on the product level but also time-sequence information about multiple purchases and information on groups of purchased products, making it possible to extract characteristic patterns from this information which can lead to the discovery of new knowledge which were unobtainable by traditional methods.

3 Analyzing the Behavior of Consumers in the Salad Oil Market

3.1 Salad Oil Market

The keyword, "Health" is one of the most powerful consumer and industrial oriented attention-getters in the Japanese food and food-related markets. In the midst of ongoing deflationary economy, products marketed under the keyword of "Health" are bought at high prices and it is believed that growth in this sector of the market will continue to expand for the time being. The item that played a large role in spurring on this growth was a salad oil product that was marketed under the image of "Health" by a manufacturer. Although this product was a new-comer to the salad oil market, it has occupied over a 10% share of the market, and we are continuing to see a constant influx of products that play up the "Health" aspect from almost all concerned manufacturers making this an important segment of the overall market. In this paper we refer to this group of products that are marketed under the keyword of "Health" as health-oriented salad oils.

The salad oil market is composed of 17 health-oriented salad oil products and 16 normal salad oil types and the market share of health-oriented salad oil products in the overall salad oil market is over 30%. While there are obviously some products with strong brand-name appeal there is a big switchover taking place basically to these health-oriented salad oil products and we feel that it is appropriate to consider the users of these products as making up a single consumer segment. Fundamental analysis reveals that when compared with consumers who purchase other types of more

ordinary salad oils, users of this segment possess several special characteristics including a tendency to use oil itself in smaller quantities than their counterparts and other types of imbalances were also seen in the other products they purchased as well.

We also discovered that among the segment of consumers who purchase health-oriented salad oil products there exists a consumer segment that only purchase these health-oriented salad oil products and one that purchases the said products in conjunction with other products. In this paper we will refer to the former group as "Healthy users" and the latter group as "Dual users." In general "Healthy users" purchase salad oils at high price while "Dual users" show an extremely high level of response to salad oils that are on sale. However, when compared with normal users, both of these segments tend to purchase products at high prices excluding salad oil products and the ratio of the purchase of products on sale is low. Hereafter we focus our attention of analysis on these three consumer segments consisting of "Healthy users," "Dual users" and "Normal users."

3.2 Preprocessing and Transformation of Data

We made use of FSP data of the Kanto area's GMS (General Merchandizing Store) covering a 1-year period running from July 2002 to June 2003. During this period salad oil purchases were higher than the average with 2979 "Healthy users" (Comprising more than 66% of the category's total) purchasing mainly health-oriented salad oil products, 3437 "Dual users" (Comprising less than 66% of the health-oriented salad oil total) purchasing health-oriented salad oil products and other products, and 12,088 "Normal users" not purchasing any health-oriented salad oil products. We carried out analysis to discover the characteristics of the purchasing behaviors of "Healthy users" and "Dual users" in the 1-month period following their health-oriented salad oil product purchases. From the principal food product categories made up of 50-product groups we extracted data for eggs, milk, bread and mayonnaise because of their strong relationships with salad oils. We divided the price zones of the market prices of each individual product associated with these product groups into the ranks of a high price zone, a middle price zone and a low price zone. We also broke the 1-month analysis period into 3 10-day periods and analyzed the purchased products groups within the said periods.

3.3 Special Characteristics of the Extracted Consumer Behavior Patterns

When we analyzed the above-mentioned data using graph mining technique we discovered several characteristic aspects of purchasing behavior regarding product purchasing and price zones.

1) Tendencies of "Healthy users" to purchase high price zone products

In each of the 3 periods "Healthy users" simultaneously purchased high price zone products spanning over differing categories at a ratio of around 10 to 15% which was relatively higher than the 5% found with "Dual users". Further, when compared with "Dual users" it was discovered that many more "Healthy users" tended to make consecutive purchases of high price zone products in the categories of bread, eggs and

milk as shown in Figure 2. The 10.8% ratio of "Healthy user" consumers who purchased high price zone products 6 times or more from among the 3 categories in the 3 periods greatly exceeded the 4.7% ratio found with "Dual users."

"Healthy users" did not only show a tendency to purchase products that have a low frequency of purchase such as salad oils in the high price zone, but also tended to purchase items with a high frequency of purchase such as bread, eggs or milk in the high price zone. This makes the "Healthy user" segment an extremely attractive one to retailers and manufacturers. In our meetings with specialists a proposal to create a new consumer segment based on these product groups was suggested and we have since launched a project to do so.

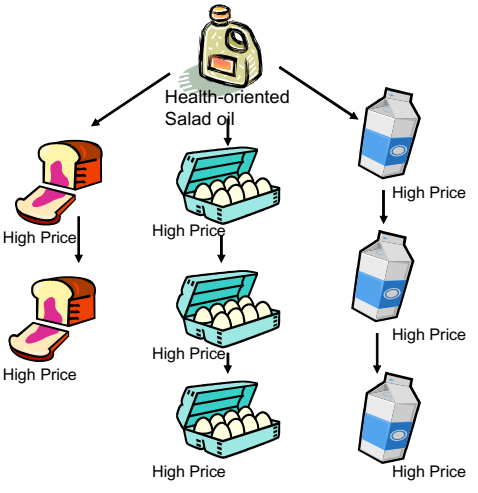

Fig. 2. Tendencies of healthy users

2) Importance of detailing these categories

We discovered that both types of users exhibited a pattern of continuing to purchase products in the high price zone categories in which they made purchases in the first 10-days following the purchase of a health-oriented salad oil product. In the previous section we came across a rule which showed that "Healthy users" possessed a strong tendency to purchase products in the high price zone in all categories, and we also verified that "Dual users" also possess tendencies to make purchases in particular high price zones and that they tend to continue to make these purchases.

This tendency of "Dual users" does not stem from the purchase of a health-oriented salad oil product but is rather assumed to be more like a consumer who originally possesses a loyalty to a particular category making trial purchase of a health-oriented salad oil product that has been marketed at a sale price. In particular a large ratio of "Dual users" continues to purchase milk in the high price zone and this can be said to be true of "Healthy users" as well. The purchasing frequency of salad oil is lower than

that of milk or eggs and thus it may be that consumers who purchase these products often may also tend not to pay too much attention to salad oil products.

4 Conclusion

In this paper we carried out an initial attempt that involved applying graph mining to the behavior analysis of consumers in the marketing field, converting the purchase history of real world industrial data into graph structures. Using graph structured data to represent consumer behavior makes it possible to effectively convey information that possesses temporal sequence property and in particular relationships existing between multiple product groups purchased in multiple purchase settings. By extracting patterns discovered by graph mining in analyzing consumer behavior in the salad oil market, we were able to define several characteristic patterns. Graph mining has so far not been applied to analyze the behavior of consumers. Our results are encouraging and we hope that this would be a valuable initial step toward a new type of consumer behavior analysis to understand a chance in the purchasing historical data. The future directions of our work are to present scenario communication process among participants after understanding the specific events by using graph mining application and to evaluate the performance of business action emerged from various processes.

References

1. Hamuro, Y., Kawata, H., Katoh, N., Yada, K.: A Machine Learning Algorithm for Analyzing String Patterns Helps to Discover Simple and Interpretable Business Rules from Purchase History. Progress in Discovery Science. LNAI Vol.2281. Springer-verlag, Berlin Heidelberg New York (2002) 565-575.
2. Hamuro, Y., H., Katoh, N., Ip, E. H., Cheung, S. L., Yada, K.: Combining Information Fusion with String Pattern Analysis: A New Method for Predicting Future Purchase Behavior. V. Torra(ed.): Information Fusion in Data Mining. Studies in Fuzziness and Soft Computting. Vol.123, Springer-verlag, Berlin Heidelberg New York (2003) 161-187.
3. Inokuchi, A., Washio, T., Motoda, H.: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. (2000) 13-23.
4. Inokuchi, A., Washio, T., Nishimura, Y., Motoda, H.: General Framework for Mining Frequent Structures in Graphs. Proc. of the International Workshop on Active Mining. (2002) 23-30.