J Intell Inf Syst (2011) 36:385–402
DOI 10.1007/s10844-009-0113-8

String analysis technique for shopping path in a supermarket

Katsutoshi Yada

Received: 24 March 2009 / Revised: 3 August 2009 / Accepted: 13 November 2009 /
Published online: 3 December 2009

Abstract The sensor network technology developed in recent years has made it possible to accurately track the in-store behavior of customers which was previously indeterminable. The information on the in-store behavior of customers obtained by using this technology, namely information on their shopping path, provides us with useful information concerning the customer's purchasing behavior. The purpose of this research is to apply character string analysis techniques to shopping path data so as to analyze customers' in-store behavior, and thereby clarify technical problems with them (the character string analysis techniques) as well as their usability. In this paper we generated character strings on visit patterns to store sections by focusing exclusively on customers stopping by these sections in order to elucidate the visiting patterns of customers who made a large quantity of purchases. In this paper, we were able to discover useful information by using the character string analysis technique EBONSAI, thereby illustrating the usability and usefulness of character string analysis techniques in shopping path analysis.

Keywords String analysis · Shopping path · Supermarket · Marketing · Consumer behavior · EBONSAI

1 Introduction

Thanks to technological advances and a lowering of implementation costs, radio frequency identification, commonly known as RFID, has come to be used in a variety of businesses. In 2005, the Ministry of Economy, Trade, and Industry conducted an experiment entitled the "Future Japanese Store Project (METI 2005)." In the experiment, shopping carts equipped with RFID devices were used to grasp customers'

K. Yada (✉)
Faculty of Commerce, Kansai University, Suita, Osaka, Japan
e-mail: yada@ipcku.kansai-u.ac.jp

behavior within stores by gathering data on customer behavior and purchasing activity at the store counter. In this project, data on customers' movements within the store was gathered electronically, and thus it was possible to obtain detailed data on customer purchasing behavior within the store, something which had previously remained totally unknown. The focus on using RFID to gather detailed data on customer behavior within the store is a trend observed not just in Japan but in Europe and America as well.

Until now, in order to understand consumer behavior in the retail industry, historical data on customers, like point-of-sale data (POS data) has traditionally be used. Using such data, one can determine which customer purchased what and where, and that data can then in turn be analyzed in greater detail. For example, in the field of marketing, Guadagni and Little (1983) and Gupta (1988) proposed consumer purchasing behavior models using such data. More recently, in order to handle large volumes of data, data mining was conducted in many industries (Hamuro et al. 1998; Ohsawa and Yada 2009), and this was helpful for improving sales promotion activities or brand strength. However, while customer purchasing history data is able to record the purchasing results for a given customer, it is not able to shed any light on how customers moved through the store or how they came to purchase them. In previous studies, in other words, the route traced by customers within a store was treated as a form of black box, and only the data on resulting purchases was made the subject of subsequent analysis.

Progress in RFID technology in recent years brought a complete about face to that situation. The "Future Store Initiative (Loebbecke 2005)" by Metro in Germany leads many companies to apply RFID to various situations in retail and distribution (Curtin et al. 2007). In the research field of supply chain management, many articles (Angeles 2005; Jones et al. 2004) have already been published in many international journals, focused on the effectiveness of logistics to earn profits. Also, some researchers (Tajima 2007) have dealt with the strategic value of RFID in supply chain management from the viewpoint of corporate strategy. Furthermore, research on consumer perception of RFID (Chen and Pfleuger 2008; Smith 2005) has been done in the field of consumer psychology. These results indicate that RFID has wide applicability in various business fields. In particular, in marketing applicability studies on RFID technology, the greatest emphasis was placed on providing RFID devices for customers or their carts, and analyzing customer routes within the store by tracing their movements and determining their behavior (Loebbecke 2005; Sorensen 2003). Tracing customers' movements within a store makes it possible to have a better understanding of what and why customers make purchases than is the case when simply noting the product purchases, as was the case with previous marketing studies. There have been very few studies based on customer data that describes customer movements within the store. The reason for this is that until now it was exceedingly difficult to obtain such data. Accordingly, customer movement data obtained using RFID will be a springboard for new avenues of research in the field of marketing.

Among studies that have employed RFID-based customer movement data analysis, there is a study by Larson et al. (2005). They employed a clustering method that improved on the k-means algorithm, and thereby discovered a number of customer groups. By exploring these customer groups in data, they were able to suggest a number of hypotheses. However, until now there have been no studies

of applied implementation or research focusing on classification issues or abstraction of characteristics based on customer movement data.

In the retail industry, those targeting customers for a given marketing strategy need to grasp these characteristics and understand purchasing behavior. Accordingly, application studies that focus on classification problems and characteristic abstraction, and not just clustering, are thought to have important business implications.

The RFID data used in the present study is typically referred to as stream data or a data stream. Stream data is data in which changes in a subject are recorded electronically and continuously over time. In the distribution and communications fields, there is a tremendous need to obtain useful information based on such data. Moreover, such data has attracted the attention of many researchers as an important domain of application for data mining. However, because the volume of data tends to be huge and because the data tends to be unstructured, it is difficult to directly apply methods that target the sort of tabular data that in past studies were largely ignored.

We introduce knowledge expressions in the form of character strings for stream data including information about customer movements, and have proposed the adoption of EBONSAI (Hamuro et al. 2002b; Yada et al. 2007), a character parsing application used in the field of business. In other words, by abstracting information on the paths that customers trace within a store and expressing that information in the form of character strings, we thought to implement rule-based abstraction using existing character string parsing algorithms. The application of this existing technology to a new field not only demonstrates the usefulness of that technology but also clarifies new technological issues at the same time. In this study, by applying this approach to actual stream data, we hope to lay open discussions of technological issues and the feasibility of applying it to stream data to which character parsing methods are applied.

2 Analysis of customer movements and character strings

2.1 Analysis of customer movements and character strings

Customer movement analysis is a store management method that makes it possible to improve the efficiency of store layout design and sales promotional plans by analyzing the routes that customers take within a store. Figure 1 shows the movement of a customer within a store superimposed over the store layout. The paths of customer movements and their directions are shown using linked lines with arrows. Moreover, sections where a customer stops are shown as nodes, whereas red nodes indicate locations where the customer purchased something. As can be seen from the figure, customers move in extraordinarily complex manners when doing their shopping.

A particularly important influence on purchasing behavior is the rate at which a customer stops in a particular section of the store; in other words, what is key is whether the customer actually passes by and stops in any given section. This is expressed in the data as product section stops. Naturally, there are cases where customers stop by a section but elect not to buy anything. Whether the customer chose to buy something can be easily determined by comparing the movement data

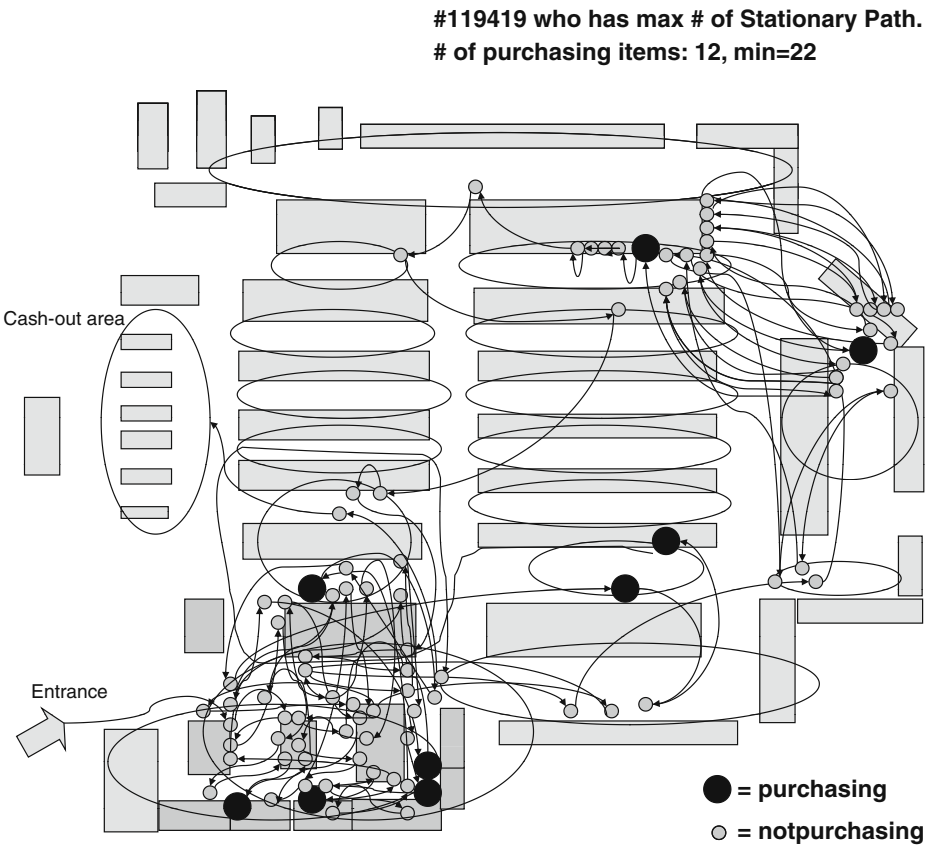
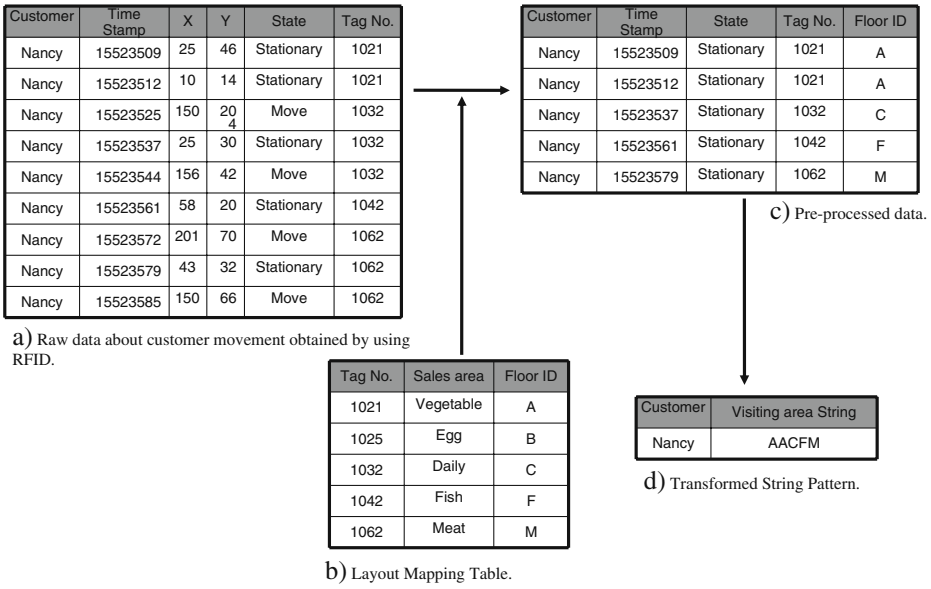


Fig. 1 An example of customer movement data

with the purchase data. This information is particularly important to those in the store merchandizing industry. For this reason, when looking at customer movement data in this paper, we focus on customer stops at product sections and abstract characteristics of the routes taken within the store by the customer.

Because it is difficult to process the stream data obtained by RFID as is, some additional processing measures must be undertaken. For that reason, in this paper, we employ character strings as knowledge representations that can be used to analyze customer movement data. We shall explain this transformational process with reference to Fig. 2. Figure 2a shows the raw data obtained using RFID. The data includes a wide variety of items, including RFID tag number, shopping cart state, and acceleration in the X and Y directions as a function of time and customer ID. This raw data is transformed using the layout mapping table shown in Fig. 2b. This layout mapping table has been provided with floor section IDs by joining the RFID tags with the store location points. Each RFID record is transformed into a character that uniquely identifies each floor. At this point, this narrows the data to one thing, namely, which in-store section the customer is currently located. Then, by linking up the succession of floor IDs based on the order in which the customer visits different sections of the store, we obtain a character string pattern like that shown in



Customer	Time Stamp	X	Y	State	Tag No.
Nancy	15523509	25	46	Stationary	1021
Nancy	15523512	10	14	Stationary	1021
Nancy	15523525	150	204	Move	1032
Nancy	15523537	25	30	Stationary	1032
Nancy	15523544	156	42	Move	1032
Nancy	15523561	58	20	Stationary	1042
Nancy	15523572	201	70	Move	1062
Nancy	15523579	43	32	Stationary	1062
Nancy	15523585	150	66	Move	1062

a) Raw data about customer movement obtained by using RFID.

Tag No.	Sales area	Floor ID
1021	Vegetable	A
1025	Egg	B
1032	Daily	C
1042	Fish	F
1062	Meat	M

b) Layout Mapping Table.

Customer	Time Stamp	State	Tag No.	Floor ID
Nancy	15523509	Stationary	1021	A
Nancy	15523512	Stationary	1021	A
Nancy	15523537	Stationary	1032	C
Nancy	15523561	Stationary	1042	F
Nancy	15523579	Stationary	1062	M

c) Pre-processed data.

Customer	Visiting area String
Nancy	AACFM

d) Transformed String Pattern.

Fig. 2 RFID data and product-section visiting pattern strings

Fig. 2d. For example, if we use the mapping table, we can express the store-section visiting pattern for the customer identified as Nancy in Fig. 2 as "AACFM."

2.2 Purpose of this research

The purpose of this research is to propose a knowledge discovery system that can abstract useful information from character strings representing store-section visiting patterns for both positive and negative purchasing events. This is accomplished by applying character string parsing technologies on stream data pertaining to customer purchasing behavior within the store. At the time that we devised this system, we made use of a previously existing system known as EBONSAI. EBONSAI (Hamuro et al. 2002b; Yada et al. 2007) is a time series analysis technique adapted from the BONSAI character parsing approach employed for the genome project. Up to now, EBONSAI had been used for time series analyses of sales data, web log data, and the like, but it had never been used for the kind of stream data that is generated by RFID. In this paper we hope to demonstrate that it can be applied to the kind of stream data found in the field of marketing. We shall do this by clarifying technological issues, showing the method's usefulness, and applying it to character parsing for this kind of stream data.

2.3 EBONSAI

EBONSAI is an adaptation of the BONSAI character parsing system that was originally developed in the field of molecular biology (Arikawa et al. 1993; Asai et al. 2004; Hirao et al. 2003; Shimozono et al. 1994). It is a system whereby positive and negative events are expressed as character strings, and using those partial

character strings or partial sequences highly refined decision trees are generated. We shall begin by first explaining the BONSAI algorithms that form the core of the EBONSAI system.

Let P be positive data set, N be negative data set, and |P| and |N| be the numbers of records in P and N, respectively. Given a substring α, let P_T Sand n_T be the numbers of records containing α in P_T and n_T, respectively, and let P_F and n_F be the numbers of records not containing α in P and N, respectively. Defining entropy function ENT(x,y) in the following manner,

$$\mathrm{ENT}(\mathrm{x},\mathrm{y}) = \begin{cases} 0 & \mathrm{x}=0 \text{ or } \mathrm{y}=0 \\ -\mathrm{x}\log\mathrm{x} - \mathrm{y}\log\mathrm{y} & \mathrm{x},\mathrm{y} \neq 0 \end{cases} \tag{1}$$

we define in the following expression the entropy obtained after classifying the original data into two subsets depending on whether data contains α as a substring or not.

$$\frac{\mathrm{p_T}+\mathrm{n_T}}{|\mathrm{P}|+|\mathrm{N}|}\mathrm{ENT}\left(\frac{\mathrm{p_T}}{\mathrm{p_T}+\mathrm{n_T}},\frac{\mathrm{n_T}}{\mathrm{p_T}+\mathrm{n_T}}\right)+\frac{\mathrm{p_F}+\mathrm{n_F}}{|\mathrm{P}|+|\mathrm{N}|}\mathrm{ENT}\left(\frac{\mathrm{p_F}}{\mathrm{p_F}+\mathrm{n_F}},\frac{\mathrm{n_F}}{\mathrm{p_F}+\mathrm{n_F}}\right) \tag{2}$$

We compute α which minimizes this value. Namely, we choose α for which the information gain is maximized. After partitioning the original data based on α, BONSAI continues to proceed in a recursive manner.

Like BONSAI, EBONSAI incorporates an alphabet indexing mechanism. This mechanism is achieved by substituting the smallest possible character string for a given characteristic character set for positive events. This makes it possible to abstract high-level rules that can interpret relatively small character strings while reducing the search space. From the total alphabet set Σ, we convert the original character string using the mapping (image) ϕ for the smallest collection of letters generated randomly Γ, and in the above-mentioned manner generate a decision tree. Next, we search until the neighborhood of ϕ cannot be further refined, and output a

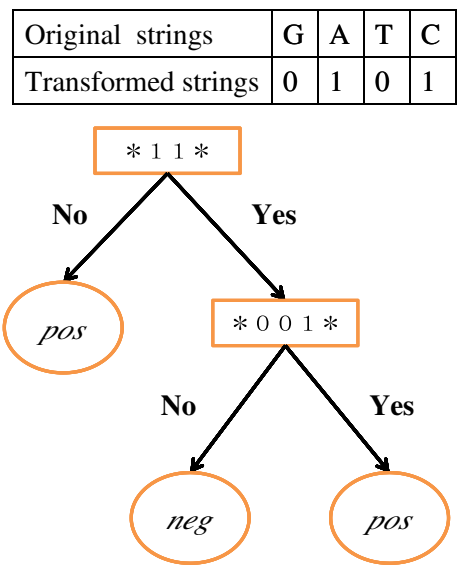


Fig. 3 Example of EBONSAI output

decision tree that has the greatest discrimination. By using the appropriate alphabet listing, it is possible to refine the classification and simplify one's hypothesis.

EBONSAI functions are easy to appreciate by looking at the output. Figure 3 gives an example of EBONSAI output. Based on the upper mapping table, EBONSAI converts the four given character strings into 1s and 0s. For positive events that have been converted, it is possible to check whether they conform with character strings abstracted from the root of the decision tree. For example, we follow along the "yes" arrows in the case that a character string of "11" is included, and follow along the "no" arrows for cases where it is not. In this way, by using only a few converted character strings, EBONSAI can generate decision trees having a relatively simple predictive ability. Because EBONSAI was for the most part used for purchasing pattern character strings, it can be applied to character string data comprising 100 character strings or more. In addition, other areas where EBONSAI was improved are described below.

- In business, in order to handle a number of cause-and-effect relations, it is necessary to contend with a various attributes simultaneously. For this reason, EBONSAI is able to employ a number of character string attributes. In addition, just like general decision tree algorithms, EBONSAI can handle category attributes and numerical attributes in one model simultaneously, and not just character string attributes.
- EBONSAI can handle data structured in the form of a table described using XML, and if used in conjunction with the MUSASHI open source platform, a viable system can be easily constructed.

2.4 System overview

Figure 4 shows a concept diagram of the knowledge discovery system employing RFID which we developed. Three databases were used for the raw data, and each of these is associated with a preprocessing system. The preprocessing systems handle

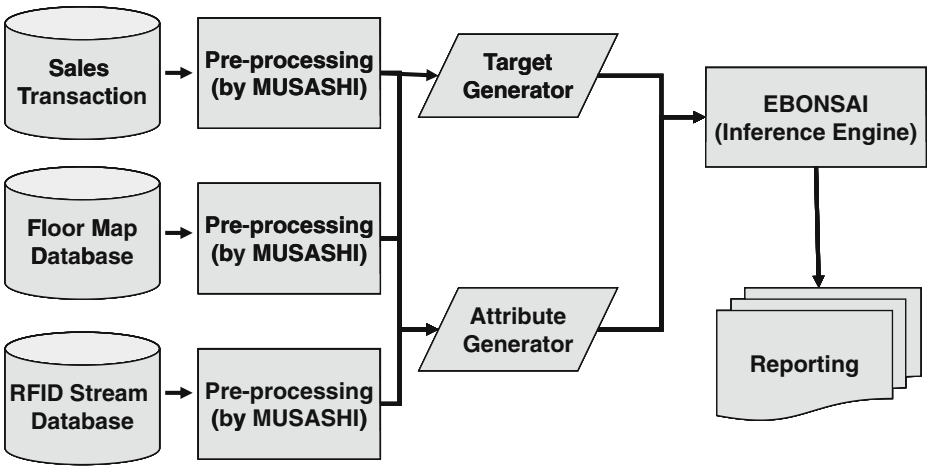


Fig. 4 Overview of the knowledge discovery system for customer movement data

data in XML form, and then transfer this data to the target generator and attribute generator in the next stage. Next, the data is combined and a classification model is constructed based on the mining engine. All of this was put together using the MUSASHI open source platform for data mining (Hamuro et al. 2002a).

We will now explain all of these major subsystems in greater detail. This system employs three databases. The first database houses data on customer purchasing history, and includes information on customer ID, purchase price, product information, and the like. The second database contains store layout information. This database contains a database of products together with RFID sensor location information. This makes it possible to track customer position information and the sections in the store where purchased products were obtained. The third database contains RFID sensor logs. By pooling all of these databases together, it is possible to determine how a given customer moved within the store, as well as where the products that a customer purchased were located within the store.

Next, let's examine the target attribute generator. What we have developed here is a system that, by pooling various databases together, can construct classification models for customers. Accordingly, it is necessary to generate target attributes to be subject to classification from the above-mentioned databases. Using the RFID sensor log database and the purchasing history database, this component generates attributes freely defined by the user. For example, we might want to contemplate the ideal customer for a particular shop or the characteristics of buyers of a particular product. In the same way, we can prepare components that generate explanatory attributes that use classification models from the above database.

From this, we can derive explanatory attributes relating to purchasing information on a particular product or a product category together with information on customer movement within the store. For example, using data on the customer's movements within a store, we can generate a sequence attribute that elucidates the order in which a customer visited various product sections of a store.

Finally, the mining engine can construct a classification model based on target attributes and explanatory attributes. The mining engine for this system was developed using EBONSAI as its foundation. For that reason, decision trees can be output using numerical figures, categories, and string attributes that were generated by the above described databases.

3 Experimental results

3.1 Explanation of the data

We will now demonstrate how this system can be used with actual customer movement data, and will perform an experiment on rule abstraction. We used customer movement data gathered at a mid-sized super market in Japan. In this project, the shopping carts that customers used were equipped with RFID receivers, and each product section had RFID tags. This made it possible to track customer movements within the store precisely. The experiment was conducted in September 2006. In addition to passenger movement data, floor layouts and purchasing history data were also gathered. The floor layout within the store was divided into seven sections. Each of those sections had subsections, and in total there were 17 subsections.

The purpose of the analysis in this experiment was to use the system proposed in this paper to clarify what characterized the movements of customers who bought a relatively large number of items. In this case, our data was somewhat restricted, as we simply measured the number of purchased items at the time the customer visited the store. However, we did not consider purchasing power (the total amount a particular customer spent per month on purchases), nor did we consider the intervals between shopping trips or frequency with which customers shopped. For our clustering method we used k-means, and we defined customers purchasing a relatively large number of items as "high-volume" (HV) customers, with the rest of the customers being deemed "low-volume" (LV) customers, that is same result as using other clustering methods such as chi-square based clustering. The average number of items purchased by HV customers was 19 per store visit; the same average for LV customers was 7.86. The ratio of HV customers was 33.3%, and LV was 66.7%.

In this experiment, we used two kinds of attributes, numerical attributes and character string attributes, and these were output from the component that generates explanatory attributes. In terms of character string attributes, we used two kinds of product section visiting pattern strings, those for product sections and those for product subsections.

Moreover, each area's percentage of total staying time is used as a numerical attribute. Since initial consumer behavior research was done (Feldman and Hornik 1981; Jacoby et al. 1976; Hornik 1984), shopping time has been found to be one of the most important factors affecting purchases. Many researchers have studied shopping time for purchasing in store, such as research on the relationship between shopping time and the situation (Yalch and Spangenberg 2000) and sales promotions (Marmorstein et al. 1992) when purchasing. Also in recent research on internet shopping (Morganosky and Cude 2000), shopping time is one important index for understanding consumer behavior. In particular, time staying in sales areas in stores is one of the most important determining factors for purchasing behavior (Baker 2000). If staying time comprises a large share of customer shopping time, it is

Fig. 5 Sample calculation of component ratio of staying times

Table 1 Sample numerical attributes of customer i

Customer	Sec. A	Sec. B	Sec. C	Sec. F	Sec. M	Sec. R	Sec. V
i	0.5	0.2	0.3	0	0	0	0

conjectured that purchase value will increase. Shopping time varies widely depending on the individual, thus in this paper, the percentages of staying times in each section are used as explanatory attributes.

The supermarket used in this example is comprised of seven sales sections (A,B,C,F,M,R,V). Customer i stayed 50 s in Section A, next moved to Section B and stayed 20 s, then moved to Section C and stayed 30 s (refer to Fig. 5). There were 100 s total staying time in Customer i's shopping.

Thus, if customer i remained in section x t_{it} seconds, then the component ratio r_{ix} of time spent by customer i in area x was expressed as follows:

$$r_{ix} = \frac{t_{ix}}{\sum t_{ix}} \tag{3}$$

Customer i of Fig. 5 above has the seven attributes in Table 1.

3.2 The effectiveness of character string expressions in the form of visiting patterns

The purpose of this research is to demonstrate the applicability of character string expressions in analyses of visiting patterns. In this section we would like to point out that visiting pattern character string attributes may potentially contain valuable information. First, we compared the predictive accuracy of a model which uses the component ratio of time spent in each section r_{ix} prepared above with a model that uses a visiting pattern character string (walking). Information related to the customer's visiting pattern and sequence was not included in the component ratio of time spent by customers in each section, while at the same time there was no information concerning staying time in the visiting pattern character string. However, going by the opinions of business people and marketing researchers, it was inferred that the time spent in each section and the component ratio of time spent in each section are related to the customer's merchandise purchases.

We thought that it would be possible to conjecture on whether or not there was any important information included in the visiting pattern character string by comparing the accuracy of models using both of these. For the component ratio of time spent in each section model, EBONSAI was used on the logistic regression, C4.5, and visiting pattern character string, and the overall accuracy, precision, recall, and F-Measure (Ohsawa and Yada 2009; Witten and Frank 2000) were

Table 2 Comparison of the predictive accuracy of the component ratio of time spent in each section and the visiting pattern character string

	Attributes used	Overall accuracy	Precision (HV)	Recall (HV)	F-Measure
Logit	r_{ix}	0.727	0.623	0.615	0.619
C4.5	r_{ix}	0.75	0.62	0.795	0.697
EBONSAI	Walking	0.764	0.615	0.923	0.738

calculated as averages of cross-validation (10 fold). EBONSAI employs the same pruning algorithm (Quinlan 1993) as C4.5, and the experiment was conducted with a parameter of cf = 0.25 and the smallest sample number per leaf as 15 for both of them. We used WEKA (Witten and Frank 2000) for the experiment. As can be seen in Table 2, the EBONSAI model which uses visiting pattern character strings had a classification accuracy that was roughly the same as or higher than the two models for the component ratio of time spent in each section. It was not possible to generate and compare all of the attributes due to constraints with the data, but it can be inferred that important information concerning the customer's purchase quantity is included in the visiting pattern, the same as with the component ratio of time spent in each section.

When actually trying to detect important information, it is essential to evaluate the predictive accuracy of the methods. Not only this, but it is also essential to evaluate these from the perspectives of whether or not the abstracted rules offer new information to specialists, and whether they can produce business action in reality (Hirao et al. 2003). In order to demonstrate the usefulness of visiting pattern character strings it is important to make clear what sorts of information specialists can obtain. For this, we used a decision tree created by using attributes from the component ratio of time spent in each section and the visiting pattern character string, then had specialists compare the rules obtained from the component ratio of time spent in each section and those obtained by following the visiting pattern character string, and examined the information which was thereby abstracted. In order to increase the specialists' interpretation potential, the calculations were performed with the pruning parameter as cf = 0.10, and the smallest sample number per leaf as 30. The following summarizes the interpretation of the rules by the specialists.

The decision tree that uses the component ratio of time spent in each section in Fig. 6 represents a rule that is consistent with the existing conventional wisdom of the specialists. This is the rule which says that for the top node, if the component ratio of time spent in the fish section is 0.23 or higher, then the customer is classified as an HV

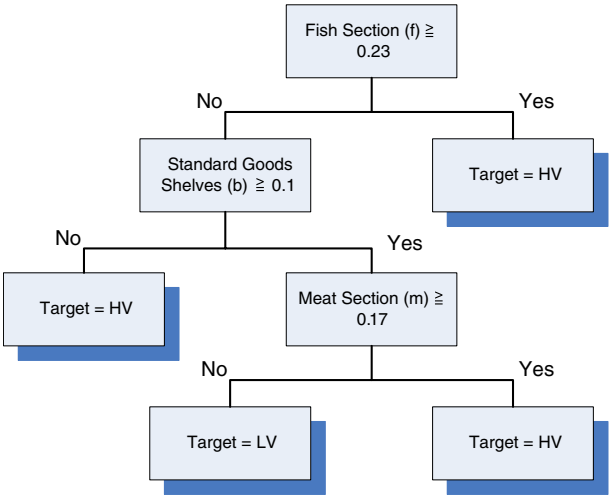


Fig. 6 Example of a decision tree that uses the component ratio of time spent in each section

customer. The importance of the three types of fresh goods (fruits and vegetables, fish, and meat) has traditionally been widely known in the supermarket industry. The claim has been made that fish in particular has vital significance to stores as a product that attracts customers irrespective of its price. According to them, the fish section was considered the store's most important section for attracting customer loyalty. This conventional wisdom can be interpreted to mean that HV customers will have a high component ratio of time spent in this section. In the next node, when the standard goods shelves (b), which include goods that are delivered daily and have a short shelf life, is less than 0.1, then the customer is classified as an HV customer. The fact that the component ratio for daily delivered goods is low is thought to indicate that not only are necessary items being selected, but so are a wide variety of products as well. For the final node, when the meat counter (m) is 0.17 or higher, then the customer is classified as an HV customer. This is similar to the aforementioned fish in that it is a category which customers focus on, and therefore was interpreted in a manner consistent with the conventional wisdom.

Next, a classification model was constructed via EBONSAI using the component ratio of time spent in each section and visiting time character string, and the abstracted rules (Fig. 7) were evaluated by the specialists. The top node is the same as with Fig. 6, but for the next node when a visiting time character string is used and there is no visiting pattern for V (vegetable section) to F (fish section), then the customer was classified as an LV customer. For the final node, a customer was classified as an HV customer if the meat section (m) was 0.035 or higher. What is distinctive of the obtained rules is that which store counter the customer proceeded to from the vegetable section (standard goods shelves or the fish section) indicated the number of items they would purchase. In other words, this represents the point of divergence between HV customers and LV customers (refer to Fig. 8). Conventionally, store managers had held the simple supposition that the larger a customers' component ratio of time spent in the fish section the higher their purchasing score would be. However, important characteristics in the visiting patterns to sales areas were observed through this research. Therefore, even if customers could be drawn to

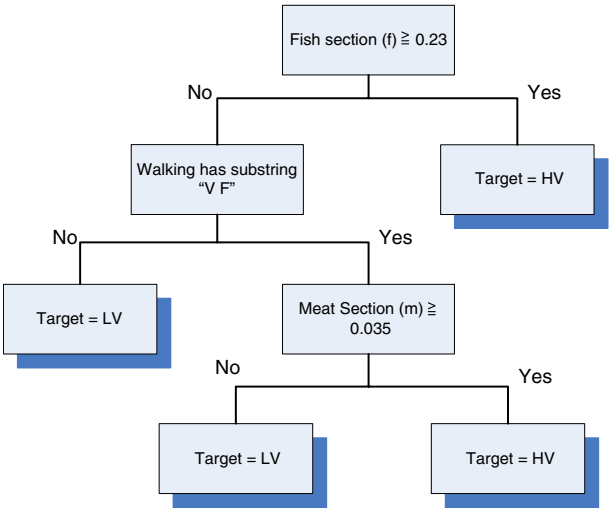


Fig. 7 Rules abstracted via EBONSAI using the component ratio of time spent in each section and visiting pattern character string

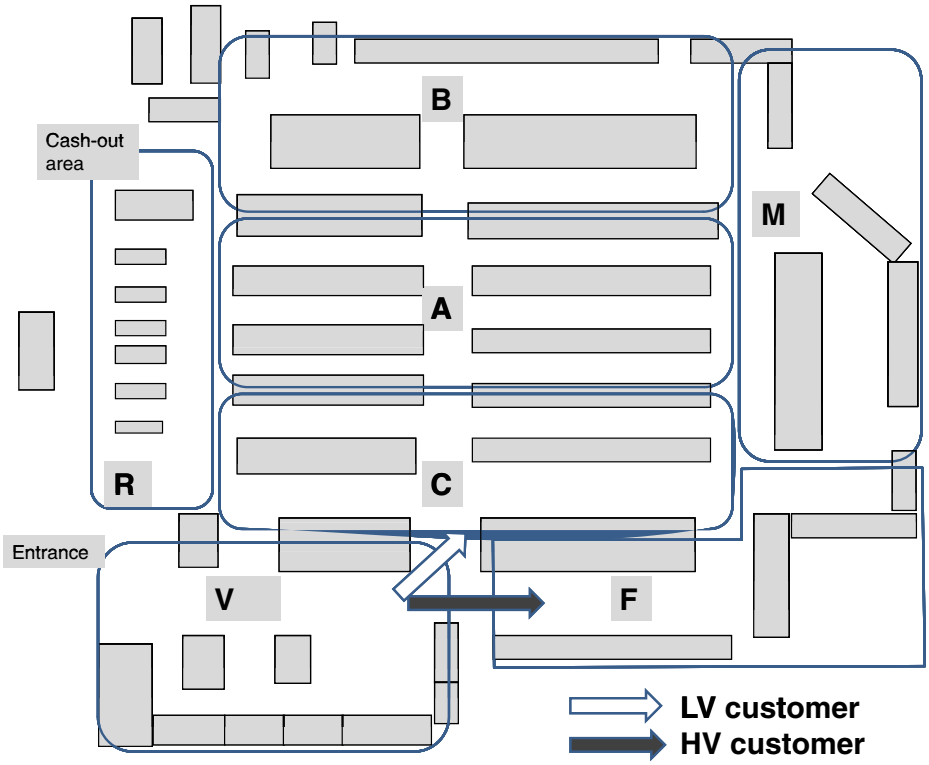


Fig. 8 Patterns of movement between store sections by HV and LV customers

the fish section unless they stopped by the adjoining meat section, in other words, if they went from the fish section to the area of general goods shelves (c) or otherwise passed by the meat counter, then they would be LV customers. It was conjectured that this was because sufficient demand was not successfully created at previous vegetable and fish sections.

What business people like store managers were most interested in were the rules concerning the movement from the vegetable section to the fish section. When we discussed this information with store managers, we obtained the hypothesis that perhaps the majority of HV customers received an impetus of some sort at the vegetable section that prompted them to decide on the products they would purchase (fish). What is more, they also thought that the stimulation of demand for fish at the vegetable section governed the customers' purchasing inclination within the store. Based on this hypothesis, the question of how to induce customers down the shopping path from the vegetable section to the meat section can be considered an important issue to retail stores in terms of in-store layout.

Finally, items pointed out by the specialists in terms of the usefulness of the knowledge representation of the visiting time character string have been compiled into the two points below.

- It is possible to provide basic data which examines in-store layout from a cross-sectional perspective. At stores, the product managers (mainly buyers) for each

section can essentially determine the sales section composition and product assortment based on the actual sales performance. The rules for visiting patterns serve as basic data for creating sales sections based upon the in-store behavior of customers from a cross-sectional and compound point of view. In this case, the fact that a "story between sections" was needed in order to induce customers from the vegetable section to the fish section was explored.

- The abstracted visiting patterns have produced suggestions for business action implementation sites. For example, the implementation of efforts like related menu proposals between the vegetable section and fish section can be considered in order to promote movement between these areas. Furthermore, instead of sales promotion measures related to the vegetable section, the placement of products on sale from the general goods shelves (c) at the boundary between the vegetable section and the general goods shelves (c) was considered in order to contain customer movement.

Since suggestions such as these can be provided to businesses, it is believed that information on the in-store behavior of customers is to be found in the visiting time character string, which is an important knowledge representation regarding shopping path analyses.

3.3 Technical issues with EBONSAI as a character string parsing technique

Based upon the considerations mentioned above, it is thought that the visiting time character string includes rich data concerning in-store customer behavior. Accordingly, it is presumable that character string parsing techniques that are capable of directly dealing with such data have the potential to generate useful information. However, it became apparent that there are several problems remaining with EBONSAI, which was used as the character string parsing technique for this paper, which must be resolved in order to apply it to customer shopping path analyses. This section will focus on problems with EBONSAI and character string representations, and elucidate future challenges.

3.3.1 Alphabet indexing in shopping path analyses

We would like to first include our considerations regarding EBONSAI's alphabet indexing. The indexing in EBONSAI is not just for simply reducing search space; it also has the effect of simplifying the semantic interpretation of the rules. For example, in cases where EBONSAI is applied to brand switching patterns, brands that have been released by the same manufacturer, as well as brands which have the same taste and similar targets (low price range products, etc.), are oftentimes substituted with a single letter through indexing. Doing this simplifies the outputted rules, which in turn has the effect of increasing the potential for interpreting said rules. However, the indexing does not function sufficiently from the standpoint of rule interpretation potential when it comes to customer shopping path analyses. This is because when a single letter is substituted for multiple sections it becomes difficult to discern the specific meaning for these substituted section groups. Instead, the indexing invited confusion in the actual discussions with the specialists, and only the decision trees which did not use indexing for rule abstraction through the use of EBONSAI, like the one in Fig. 7 above, drew their attention. For the customer

shopping path analyses, the majority of the specialists were interested in finding particular routes which had a significant impact on the objective variable. Therefore, refraining from the use of indexing, or methods in which the size of the index is made as large as possible, should be considered.

However, making the size of the index larger leads to an increase in calculation time. Figure 9 shows the relationship between the size of the index, which is one of EBONSAI's parameters, and the calculation time before rules are outputted. In this experiment, the inside of the store was divided into 17 sales areas, with the visiting time character string (s-walking) for each sub-area adopted as an explanation attribute. EBONSAI's index size is a specification that is set by parameters, with two as the default size. Viewed from the perspective of predictive accuracy, in most of the cases a high degree of accuracy was usually attained when the index size was two or three. In addition, there was also a tendency for the calculation time to shrink as the index size became smaller. In the graph of Fig. 9, the calculation time increases from an index size of seven, but there was no extreme elongation after that. It is believed that the reason for this is that a large number of sub-areas with an extremely low frequency of visits were included. In the case of stores where an enormous number of samples could be collected and where all of the sub-areas have over a certain visit frequency, there is the possibility that an increase in the index size will lead to an extreme enlargement of the calculation time. Moreover, since EBONSAI's current maximum number of indexes is nine, it must be improved upon so that it can handle indexes with even larger sizes in the future. It will also be necessary to consider methods for reducing the search space by means of new approaches other than indexing.

3.3.2 Sequential pattern for shopping path

EBONSAI is not only capable of generating partial character strings, but can also create decision trees which include sequential patterns, such as those expressed in regular expressions and the like (Hamuro et al. 1998; Yada et al. 2007). However, the specialists did not pay attention to rules which included sequential patterns, but rather requested that rules for sales areas with consecutive movement—namely, decision trees composed of partial character strings—be generated. For sequential

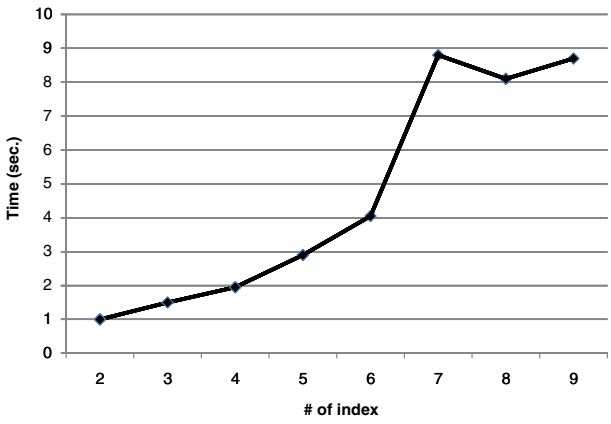


Fig. 9 Calculation time as a function of indexing

patterns, the visiting time character string for sales sections indicates the relative visiting sequence of sales sections, and is not limited to consecutive visits. The specialists acknowledge the usefulness of information concerning consecutive sales counter visits, but their assessment was that there are problems in terms of the potential for interpreting sequential patterns and the suggestions for business action.

Considering that the rules abstracted in the case are relatively short character strings, EBONSAI is not the only means for analyzing visiting time character strings. For example, another conceivable method would be to interpret information abstracted through the use of an algorithm that enumerates frequency patterns, such as LCMseq (Uno et al. 2004), through an existing decision tree like C4.5. For the future, it will be necessary to perform comparative examinations of not only EBONSAI, but also other character string parsing techniques, and to clarify what sort of parsing techniques are suited to the data, environment, and analysis needs.

3.3.3 Missing information on time spent in each section

Significant problems remain with the knowledge representation of the character strings dealt with in this paper regarding customer shopping path analyses. The most important problem is that a substantial amount of time sequence information concerning the in-store behavior of customers disappeared when visiting patterns were converted into character strings. As an example, valuable information such as the time spent moving between sections and the time spent in certain sections was not reflected within the character string-based knowledge representation. In order to resolve these problems, the introduction of new knowledge representations, such as introducing graph structure data (Yada et al. 2006), should be considered. With graph structure data it is possible to include not only visiting patterns for the sales areas, but also time sequence information such as the time spent moving between sections and the time spent in each section. Hereafter, considerations from the perspective of the usefulness of customer shopping path analyses must be appended for these other knowledge representations.

4 Conclusions

In the present study we sough to discover information on customer purchasing behavior by applying existing character string parsing techniques and applying them to stream data describing customer movements and obtained using RFID. In looking at customer movement data we chose to focus on visits that customers made to each product section. By then expressing product section visiting patterns in terms of character strings, we sought to efficiently handle large volumes of stream data. We found that HV customers, who purchase a relative large number of items tended to move from the vegetable section to the fish section. While hypotheses obtained this way are extremely novel and rich in their implications, we suffered from a rather small sample size, and future studies will be needed. Moreover, through this experiment, we were able to appreciate certain issues pertaining to existing character string parsing techniques.

Nevertheless, there are fundamental problems with applying the character string parsing techniques used in this study. Namely, time series information with respect to visiting patterns largely vanishes. For example, important information like the

time spent at a particular product section or the amount of time spent moving from one section to another was not reflected in the character-string based knowledge representation. To resolve such issues, it seems that a fruitful approach might be to introduce graphical data. If graphical data were provided, one would be able to include not only product section visiting patterns but also time series information such as the amount of time spent between sections or at a particular section. We hope to address such issues in the future.

Acknowledgements This work was supported by MEXT. KAKENHI 21013032, and "Strategic Project to Support the Formation of Research Bases at Private Universities": Matching Fund Subsidy from MEXT, 2009–2013.

References

Angeles, R. (2005). RFID technologies: Supply chain applications. *Information Systems Management*, *22*(1), 51–66.

Arikawa, S., Miyano, S., Shinohara, A., Kuhara, S., Mukouchi, Y., & Shinohara, T. (1993). A machine discovery from amino acid sequences by decision trees over regular patterns. *New Generation Computing*, *11*, 361–375.

Asai, T., Abe, K., Kawasoe, S., Arimura, H., Sakamoto, H., & Arikawa, S. (2004). Efficient substructure discovery from large semi-structured data. *IEICE Transaction on Information and Systems*, *E87-D*(12), 2754–2763.

Baker, R. G. V. (2000). Towards a dynamic aggregate shopping model and its application to retail trading hour and market area analysis. *Papers in Regional Science*, *79*(4), 413–434.

Chen, J. V., & Pfleuger, P. J. (2008). RFID in retail: A framework for examining consumers' ethical perceptions. *International Journal of Mobile Communications*, *6*(1), 53–66.

Curtin, J., Kauffman, R. J., & Riggins, F. J. (2007). Making the 'MOST' out of RFID technology: A research agenda for the study of the adoption, usage and impact of RFID. *Information Technology and Management*, *8*(2), 87–110.

Feldman, L. P., & Hornik, J. (1981). The use of time: An integrated conceptual model. *Journal of Consumer Research*, *7*(4), 407–419.

Guadagni, P. M., & Little, J. D. C. (1983). A logit model of brand choice, calibrated on scanner data. *Marketing Science*, *2*, 203–238.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research*, *25*, 342–355.

Hamuro, Y., Katoh, N., Matsuda, Y., & Yada, K. (1998). Mining pharmacy data helps to make profits. *Data Mining and Knowledge Discovery*, *2*, 391–398.

Hamuro, Y., Katoh, N., & Yada, K. (2002a). MUSASHI: Flexible and efficient data preprocessing tool for KDD based on XML. In *DCAP2002 workshop held in conjunction with ICDM2002* (pp. 38–49).

Hamuro, Y., Kawata, H., Katoh, N., & Yada, K. (2002b). A machine learning algorithm for analyzing string patterns helps to discover simple and interpretable business rules from purchase history. *Progress in Discovery Science, LNAI, 2281*, 188–196.

Hirao, M., Hoshino, H., Shinohara, A., Takeda, M., & Arikawa, S. (2003). A practical algorithm to find the best subsequences patterns. *Theoretical Computer Science*, *292*, 465–479.

Hornik, J. (1984). Subjective vs. objective time measures: A note on the perception of time in consumer behavior. *Journal of Consumer Research*, *11*(1), 615–618.

Jacoby, J., Szybillo, G. J., & Berning, C. K. (1976). Time and consumer behavior: An interdisciplinary overview. *Journal of Consumer Research*, *2*, 320–339.

Jones, P., Clarke-Hill, C., Hiller, D., Shears, P., & Comfort, D. (2004). Radio frequency identification in the UK: Opportunities and challenges. *International Journal of Retail and Distribution Management*, *32*(3), 164–171.

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing*, *22*, 395–414.

Loebbecke, C. (2005). Emerging information system applications in Brick-and-Mortar supermarkets: A case study of content provision devices and RFID-based implementations. In *PACIS 2005 proceedings* (p. 87). http://aisel.aisnet.org/pacis2005/87/.

Marmorstein, H., Grewal, D., & Fishe, R. P. H. (1992). The value of time spent in price-comparison shopping: Survey and experimental evidence. *Journal of Consumer Research*, *19*(1), 52–61.

METI (2005). *"Japanese-version future store project" (experimental trial of electronic tags for the realization of futuristic store service), news release*. http://www.meti.go.jp/english/newtopics/data/n051108e.html.

Morganosky, M. A., & Cude, B. J. (2000). Consumer response to online grocery shopping. *International Journal of Retail & Distribution Management*, *28*(1), 17–26.

Ohsawa, Y., & Yada, K. (Eds.) (2009). *Data mining for design and marketing*. Boca Raton: CRC.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo: Kaufmann.

Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., & Arikawa, S. (1994). Knowledge acquisition from amino acid sequences by machine learning system BONSAI. *Transaction of Information Processing Society of Japan*, *35*, 2009–2018.

Smith, A. D. (2005). Exploring the inherent benefits of RFID and automated self-serve checkouts in a B2C environment. *International Journal of Business Information Systems*, *1*(1/2), 149–181.

Sorensen, H. (2003). The science of shopping. *Marketing Research*, *15*, 30–35.

Tajima, M. (2007). Strategic value of RFID in supply chain management. *Journal of Purchasing and Supply Management*, *13*(4), 261–273.

Uno, T., Kiyomi, M., & Arimura, H. (2004). LCM ver.2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Proc. of IEEE ICDM'04 workshop FIMI'04* (pp. 1–11).

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with JAVA implementation*. San Mateo: Kaufmann.

Yada, K., Ip, E., & Katoh, N. (2007). Is this brand ephemeral? A multivariate tree-based decision analysis of new product sustainability. *Decision Support Systems*, *44*, 223–234.

Yada, K., Washio, T., & Motoda, H. (2006). Consumer behavior analysis by graph mining techniques. *New mathematics and Natural Computation*, *2*, 59–68.

Yalch, R. F., & Spangenberg, E. R. (2000). The effects of music in a retail setting on real and perceived shopping times. *Journal of Business Research*, *49*(2), 139–147.