Estimation of Personal Factors Affecting Purchasing Behavior and Its Application

August 2018

Chiaki Doi

A Thesis for the Degree of Ph.D. in Engineering

Estimation of Personal Factors Affecting Purchasing Behavior and Its Application

August 2018

Graduate School of Science and Technology
Keio University

Chiaki Doi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis proposes a method to estimate personal factors affecting purchasing behavior from the consumer's purchase history and service usage history in order to make proposals for services and products that can be tailored for the consumer. The effect of the proposed method is clarified by demonstration experiments.

1.1 Background

Consumers choose, purchase, and consume various products and services every day. When consumers purchase products and services hereafter products are taken to cover services as well, they recognize the products from advertisements and word-of-mouth information. In deciding whether to purchase the products, they consider their condition and situation. Through this decision-making process, they purchase and consume products. To summarize these processes, Blackwell [Roger et al., 1995] proposed the consumer decision process model shown in Figure 1.1. This model shows the process of consumer behavior leading to purchase, consumption and disposal. First, consumers receive inspiration from the outside that encourages them to understand products and memorize them. Second, while balancing personal factors/influences such as the budget, consumers purchase if necessity is judged to be sufficiently highly. The personal factors or influences include cultural and social status, families and relatives, and current employment situation. Most are demographic factors and psychographic factors. Demographic factors include gender, age and occupation. Psychographic factors include values, personality, and lifestyle.

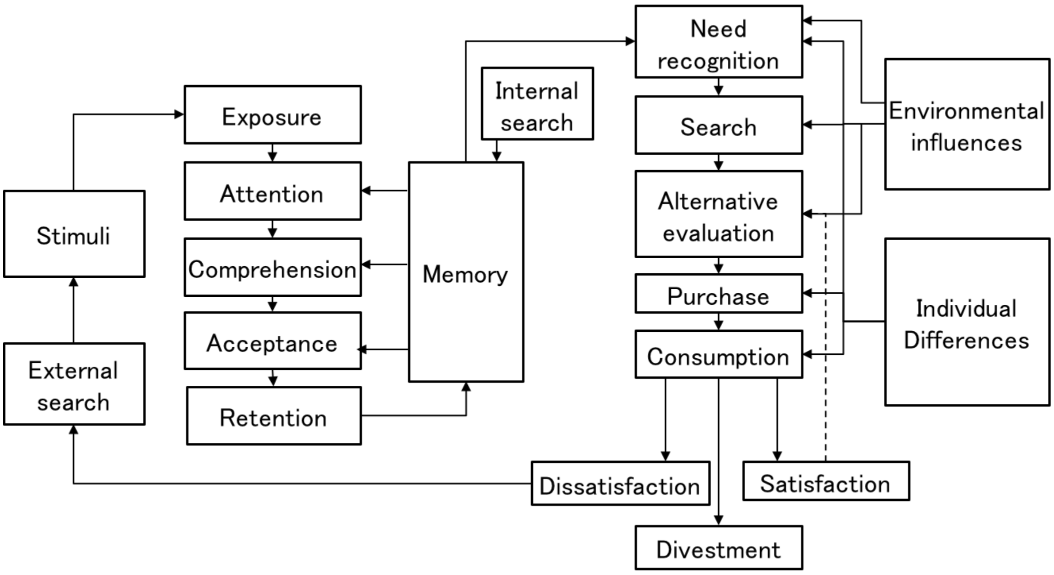


Figure 1.1: Consumer decision process model

From the consumer decision making process model, it is understood that consumers get information on products and services from the outside, store them in memory, and then recall and use those memories according to their needs. For companies wanting to sell their products, it is important that consumers remember their products when they recognize their needs so as to lead them to making a purchase. Lindquist [Lindquist and Sirgy, 2009] detailed the eight approaches to encouraging consumers to memorize products as shown in Table 1.1. This approach involves distribution method and distributed content.

Table 1.1: Eight approaches to encourage consumers to memorize

Method	Type	Detail
Repeated	Distribution method	Messages are more likely to remain in memory if they are repeated.
Information competition	Distribution method	When similar information competes for attention, it is hard to hold in memory.
Information integrity	Distribution method	When all information desired by the consumer is not disclosed, the information tends to remain in memory.
Mood	Distribution method	Positive mood such as happiness and joy will promote memory retention.
Time	Distribution method, Distributed content	Consumers become more likely to forget information as time goes by, so advertisements are more effective flows near the timing of product purchases.
Relevance	Distributed content	Information that is relevant to the consumer is more likely to be memorized.
Nearest	Distributed content	New knowledge added to well-known items are easy to memorize.
Motivation	Distributed content	Consumers who are motivated to purchase products are more likely to retain information.

As for the distribution method, it is necessary to consider the advertisement medium and the timing at which the consumer contacts the information. For example, it is conceivable to distribute information on products and services to a consumer who watches television during the day through a television commercial. For consumers who use social networking services, it is conceivable to distribute information by presenting advertisements when using the service. The contents described should differ to suit the consumer's purchase timing. Therefore, it is necessary to optimize the message content to suit the consumer's situation and condition. For example, information on low calorie foods and diet items are most effect if shown to consumers who are starting to address a weight increase. For consumers who are committed to the origin and materials of products, information distribution that matches or is similar to the conditions preferred by consumers is considered to be effective. In order to understand the situation and condition of consumers, it is common to ask consumers directly through interviews and questionnaires. However, this technique is burdensome to both sides, to the interviewer and the interviewee. To ensure adequate coverage, the interviewer often has to pay the interviewee, which raises costs. The interviewee takes time to respond to the queries. Furthermore, the disclosure of personal information becomes a psychological burden. For these reasons, it is difficult to obtain answers from most consumers.

Therefore, it is expected that a way to understand the situation and state of consumers by using the purchase history acquired from the ID-POS system. ID-POS systems can automatically acquire and store the customer ID and matching purchase history. Table 1.2 shows an example of the purchase history acquired by a typical ID-POS system. The data mainly consists of date of purchase, store used, product bought, and quantity and cost. Some systems and retail store data management methods record the type of store such as convenience store or supermarket. Type of merchandise such as milk and chocolate may also be recorded along with individual product names.

Table 1.2: Example of purchasing history data

Attributes	Value Example
Consumer ID	0001
Purchase date	2018/01/01
Purchase time	15:00:00
Shop ID	001
Shop group	Supermarket
Product ID	4912345678900
Quantity	1
Price (Yen)	100

By using the purchase history, it is possible to estimate the preference and the consumption amount of the commodity without questioning or interviewing the consumer. It becomes possible to grasp the purchasing pattern of individual consumers and purchase frequency. For example, “Consumer A goes to the supermarket every week on Saturday evening and buys vegetables and meat once a week,” or “Consumer B goes to the supermarket around 19:00 on weekdays, Purchases alcoholic beverages and prepared dishes for lunch”. The contents that can be grasped are shown below. For example, “Consumer A shops every week on Saturday evening and purchases vegetables and meat once a week,” and “Consumer B visits the stores around 19:00 on weekdays, Buy alcoholic beverages and prepared lunch boxed lunches.”. Products preferred by consumers are considered to be products with high purchase frequency and large quantities.

Also, it can be estimated that products with high purchase frequency are the products that are consumed often. In addition, by using purchase history information and attribute data indicating consumer demographic information such as age and gender, it is possible to estimate the impact of age and gender on the products purchased, the visit time and frequency. As a result, it is possible to acquire purchasing trends as a form of demographics information, and to estimate the publics taste for products and visit time and frequency. This information is generated by the usage of membership cards as they allow the correspondence between the customer ID and the purchase history to be established.

Based on the information taken from the purchasing histories, it is possible to better understand consumers by estimating the timing of product purchases and information related to consumers. One to one marketing can be realized by understanding consumers more deeply. One to one marketing allows products and services to be specially tailored to the individual consumer. By realizing one to one marketing, we can fully utilize the merits of companies that manufacture products, retail stores, and consumers. It can be useful for companies to consider new product planning that fits the actual situation of consumers and take measures to promote sales of in-house products. Consumers can receive proposals for products and services that match their circumstances and conditions from enterprises and retailers, and enterprises and retailers can expect sales to improve as a result.

1.2 Research subjects

In order to realize one to one marketing, it is necessary to grasp the personal influences and environmental influences such as demographic factors and psychographic factors as they strong influence purchase decision making. These are generally acquired by using questionnaires and interviews, but it is difficult to obtain from all consumers. Therefore, a method of estimating the situation and state of consumers are proposed by using the purchase history automatically collected and accumulated using the ID-POS system. A method of estimating the situation and condition of consumers using Web browsing history and location information has also been proposed. However, there are three problems in actually implementing one to one marketing.

The first problem is that studies to date have not clarified exactly which factors are critical for ensuring marketing success.

The second problem is that no agreement has been reached on what types of data is needed to estimate the target factors; this is also true for the estimation accuracy that can be achieved.

The third problem is the weak estimation accuracy currently available. Although methods to estimate the situation and state of consumers have been studied, most studies use limited behavior data and so are handicapped by inadequate accuracy.

If we can solve these problems, we can understand consumers more fully and create proposals that will trigger the desired responses.

1.3 Motivation

This thesis focuses on multi-faceted and long-term single source behavior data. Specifically, it targets purchase histories, questionnaire information and position information. This study analyzes the impact of consumer information on purchasing behavior. A method which derives he factors determining purchase decision making from purchase histories are investigated and the feasibility of its application is confirmed. Blackwell [Roger et al., 1995] proposed the consumer decision process model shown in Figure 1.1. This model shows that personal influences and environmental influences are the key factors in purchase decision making. Table 1.3 lists the factors that affect purchase decision making.

Table 1.3: Factors affecting purchase decision making

Influence	Item
Personal influence	Age Gender Annual income Lifestyle Preference Personality Knowledge Values Purchasing intent
Environmental influence	Culture Marital status (Single / Married) Family structure Job Employment status

This study focuses on the personal influence factors and the environmental influence factors, all of which affect the purchasing decisions. Lifestyle and product preference and purchasing intention are covered by the personal influence factors. Family structure are covered by the environmental influences factors.

Lifestyle represents the consumer's lifestyle and appreciation of value. Lifestyle has been shown to interact with consumer behavior [Buckley et al., 2007] [Boer et al., 2004]. AIO (Activities, Interests and Opinions) approach [Wells and Tigert, 1971] [Ziff, 1971] and VALS (Values and Lifestyles) [Arnold, 1984] [Arnold et al., 1986] are widely known techniques for analyzing lifestyle. AIO tries to measure lifestyle by asking about activities, interests and opinions. Examples of questions are "What kind of activities are you doing?", "What kind of things are you interested in?" and "How do you feel about various events?". VALS classifies the consumer into one of nine types based on answers to about 800 questions of value and consumption behavior. Both AIO and VALS yield consumer lifestyles as measured by questionnaire and are widely used in marketing. If we can estimate the lifestyle from the purchase history without questionnaires, we can realize marketing techniques that can identify consumer lifestyle of more consumers.

Preference as regards products directly impacts purchasing behavior. It is considered to be a key factor in purchase decision making. Therefore, it is selected as a subject to be estimated by this thesis.

For companies, finding customers who are inclined to purchase their goods is an important issue because it can improve the efficiency of various measures for motivating consumers. Specifically, it can be used in optimizing the approaches made to customers and the selection of target customers such as rank up measures and activation measures of dormant customers. Therefore, this thesis focuses on the intention to purchase at shops targeted by consumers, not specific items.

Differences in family structure are reflected in the types and quantities of products purchased by consumers. This variable also has a big influence on purchase decision-making. However, research to date has not clarified the effectiveness of using family composition or indeed which data should be estimated from other data such as purchase history.

1.4 Purpose

This study aims to understand consumers while eliminating the physical and psychological burdens imposed by asking the consumer to explicitly disclose personal information. In order to realize these goals, this thesis targets the following three points.

1. To make it possible to identify the personal factors affecting purchase decision making from observable behavior information such as purchase and service use without directly disclosing personal information by questionnaire or interview to consumers.
2. To clarify which observable information should be collected by clarifying the factors affecting purchase decision making.
3. To clarify to what extent the estimated factors determine actual purchase behavior.

1.5 Contributions

This thesis proposes methods that makes it possible to identify the factors influencing purchase decision making and evaluate the accuracy possible. Although personal factors may changes over time, that they will not change for a certain period of time. The following three points are contributions.

1. Proposal of a machine learning-based method that can model the personal factors influencing purchase decision making and purchase history or service usage history.

2. Clarify the estimation accuracy of personal factors by using real world purchase history data and service usage history.

3. Confirmation of the personal factors influence on purchasing behavior by large-scale experiments in an environment with real customers.

The first point is to clarify which factors affect purchase decision making without imposing psychological or physical burdens on consumers. This is made possible by proposing a method that can identify factors such as lifestyle and preference from the purchase history and service usage history. By establishing the proposed method, it becomes possible to know more about consumers.

The second point is to evaluate the accuracy of the proposed method. This evaluation makes it possible to clarify to what extent the purchasing history and the service usage history can be used to identify the factors influencing purchase decision making. Moreover, it can clarify what kind of history yields the most accurate factors. By performing this evaluation, the explanatory variables needed to identify the factors become clear, and factor determination becomes possible with the minimum necessary input data. Moreover, an understanding of these relationships can be utilized for product planning and sales planning.

The third point is to clarify actual consumer trends using factors identified by the proposed method. This thesis actually conducts campaigns to promote store visits and promote the sales of products, while observing the behavior of consumers. This evaluation makes it possible to clarify the strength of influence of these factors on consumer behavior. Future tasks are identified and discussed.

1.6 Structure of thesis

The structure of this thesis is shown in Figure 1.2.

Chapter 2 describes work related to this thesis.

In Chapter 3, a method that can estimate lifestyle proposes and clarifies its performance. The method extracts from purchasing history data purchasing behavior concerning products that reflect the consumer's lifestyle. This thesis focuses on the lifestyle that is regarded as the standard for consumer segmentation. Lifestyle is known to interact with consumer behavior. It has already been clarified that lifestyle and purchasing behavior are related.

In Chapter 4, a method to estimate preference for products from location information, which is one bit of behavior information proposes, and clarifies the method's performance. The effectiveness and practicality of the preference estimation method are clarified by calculating the visit rate of a target shop while conducting a visit promotion campaign. This evaluation makes it possible to clarify how strongly the estimated preference influences the consumer's purchasing behavior.

In Chapter 5, a method to estimate the family composition from purchasing history of daily necessities proposes, and clarifies the method's performance. This thesis focuses on the family structure such as the number of family members and the age of family members. Families with preschool children and families with high school students tend to purchase different products and services.

In Chapter 6, a method to estimate customer level of a target store from purchase history target store as well as other stores proposes. The effectiveness and practicality of the estimation method is clarified by calculating the estimation accuracy of the good customer and the visit rate to the target shop during the store promotion. This thesis focuses on purchasing intent as customer level. The degree of the consumer's ability to pay and intentions towards products with strong attraction to each store can be used to target customers, such as rank up measures.

This thesis summarizes in Chapter 7.

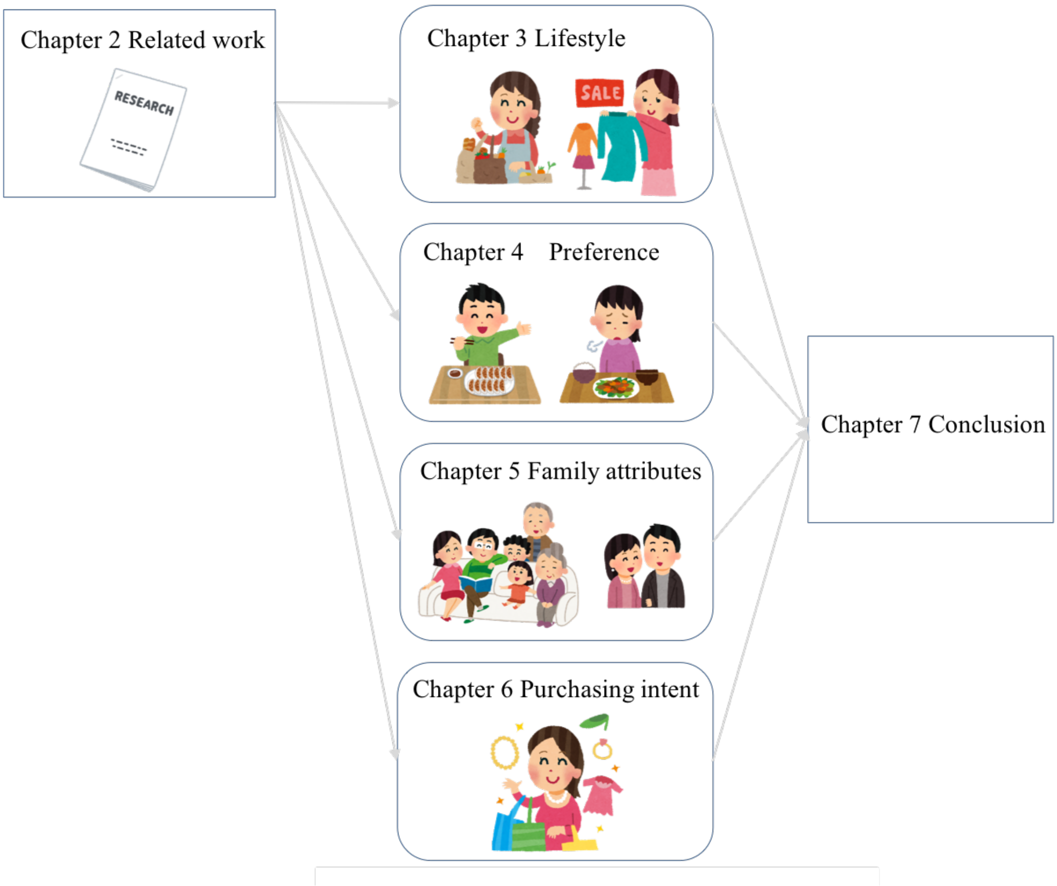


Figure 1.2: Structure of the thesis

Chapter 2

Related work

Methods of modeling factors influencing purchase decision making have already been proposed. An outline of the related work is shown in Figure 2.1. These studies use browsing history of Web service, purchase history of products, and usage history of services such as social network service. By modeling these, it is possible to estimate the condition and the situation of consumers without asking. The estimation method of each factor and the application of the estimated factor are described below.

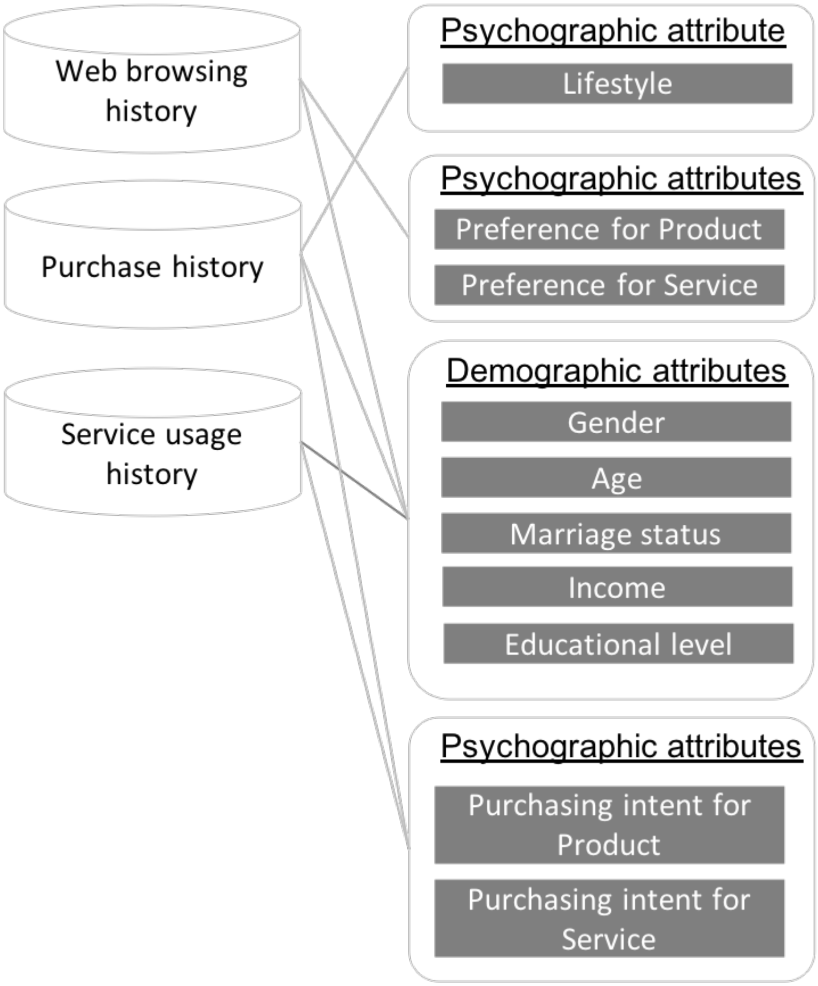


Figure 2.1: Related work

2.1 Lifestyle

Lifestyle is known to be correlated with consumption behavior. Research aimed at clarifying the difference in consumer behavior by lifestyle and utilizing it has already been conducted [Buckley et al., 2007] [Boer et al., 2004]. In this studies, consumer lifestyles are classified using questionnaire data.
Buckley et al. [Buckley et al., 2007] asked questions related to purchasing behavior and classified consumers as to their lifestyles. Instant food was the focus of this study. Furthermore, they revealed "Differences in consciousness concerning food and purchasing" and "Difference in motivation and behavior to purchase instant foods".

Research on modeling lifestyle and purchasing behavior has also been conducted [Ishigaki et al., 2011a] [Ishigaki et al., 2011b] [Ishigaki et al., 2010b] [Koshiba et al., 2013]. Ishigaki et al. [Ishigaki et al., 2011a] proposed a method of constructing a customer model in association with lifestyle and purchase history information. The lifestyle was defined by factor analysis of the questionnaire result. Ishigaki et al. [Ishigaki et al., 2011b] proposed a model that predicts the number of visitors to a store, focusing on the visit situation (day of the week, rainfall and temperature) for each lifestyle. Their analysis linked the lifestyle obtained from questionnaire results and the purchase history information as in previous research [Ishigaki et al., 2011a]. Koshiba et al. [Koshiba et al., 2013] proposed a method to estimate consumer lifestyle from just the purchasing history of consumers; its performance was evaluated. By modeling the relationship between lifestyle and purchase history, purchasing behavior for each lifestyle became clear. With this model, lifestyle can be estimated from purchasing history. However, since the number of products included in the purchase history is enormous, purchase information of the products distinguished by the JAN code is not included. They calculate several parameters that are assumed to characterize purchase behavior such as purchase quantity for each type of product such as “Vegetables” or “Fish” and the number of shop visits by day of the week. They modeled these parameters and the lifestyle. Their analysis was rather coarse-grained as products were grouped in terms of type of products such as “Milk” and “Natto”. However, it is reasonable to assume that the consumer will differentiate between manufacturers of the same product, say “Milk”. Furthermore, it is conceivable that the type of product, the product, and its strength, in which the features of each lifestyle appear is different. For these reasons, this thesis estimates lifestyle from purchase information at level of product type and product manufacturer.

2.2 Preference

Extensive research on customer preference has been carried out to enhance sales and advertising campaigns. There are several works on customer segmentation that attempt to gain a deep understanding of customer’s needs and wants [Liu et al., 2017] [Jingyan et al., 2016]. Liu et al. [Liu et al., 2017] modeled the interactions between consumption preferences, product attributes, and personality traits as purchase behavior and used the model to predict potential purchases. This thesis believes that this idea of combining data with different characteristics is very effective in determining customer preferences with regard to purchase behavior. However, there is a limit to extent to which the customer’s preferences can be understood if only a company data is used. Therefore, this thesis proposes a method to estimate consumer’s preference by combining the data of several companies in order to fully grasp the behavior of customers.

Point of Interest (POI) recommendation for services that use position information predicts the destination from the movement history of the customer and recommends the store that the customer will like. This helps to find new places suitable for the customer and also helps to exclude shops that do not fit the customer’s preferences. For example, customers who like meat dishes tend to prefer meat restaurants. For customers who frequently visit the same place, it is reasonable to recommend shops similar to the frequented places.

By using consumer preferences and visit frequency to stores, we can clarify consumer preferences. Several methods for POI recommendation have been investigated [Yonghong and Xingguo, 2015] [Zhao et al., 2016]. The influence of social effect on POI recommendation has been investigated. Since friends tend to share more interests than non-friends, it is well known that recommendation accuracy can be improved by considering the social effect [Zhang et al., 2016] [Li et al., 2016] [Ye et al., 2011]. On the other hand, some research found that most people share nothing in common as regards POI. This shows that there is a limit to the information that can be acquired from social networks [Yonghong and Xingguo, 2015] [Ye et al., 2010].

For a deeper understanding of the customer, this thesis focuses on the actions made at the places visited such as purchases made. In social network services, friending is the action of adding someone to a list of friends. This is an explicit action and so is seen as highly reliable information. Unfortunately, it is difficult to obtain relationship between people in the real world. The solution is to group customers via their purchase histories, which avoids the need for explicit customer opt-in.

2.3 Demographic attribute

It has been clarified that demographic attributes can be estimated from web browsing histories, service usage histories such as social network access, and purchase histories.

Lu et al. [Lu et al., 2015] proposed an architecture for gender prediction that uses the view logs of products that a customer browsed at an e-commerce web site.

The viewed product information reflects the user preferences. They leverage the user features, including preferences, to classify users as male or female. They showed the effectiveness of using viewed product logs for gender prediction.

Torres et al. [Duarte Torres and Weber, 2011] revealed that there are correlations between clicked web pages and demographic attributes, especially as regards age and educational level. Murray et al. [Murray and Durrell, 2000] used Latent Semantic Analysis to estimate the age, sex and income from web browsing histories. Zeng et al. [Hu et al., 2007] proposed a method to estimate gender and age from web browsing histories.

Research has also been conducted on estimating demographic attributes from service usage histories, such as those generated by social network services. Culotta et al. [Culotta et al., 2015] proposed a method to estimate the user's age, sex, educational background, existence of children, and nationality. They focused on the relationship between users of Twitter [Twitter, 2018] which is a key social network service and the demographic information of the followers of Twitter posters. Follower is person who subscribes to the posts of another user. By clinking on the "Follow" button, the posted contents of user A are displayed on the Twitter page belonging to user B. Mislove et al. [Mislove et al., 2010] revealed that social network service profile information indicates that friends have common or similar profile information. They proved that some information about educational background and interests can be estimated using this tendency. Dong et al. [Dong et al., 2014] revealed that the demographic attributes can be estimated by using a cell phone's call history and mail transmission history. Zhong et al. [Zhong et al., 2013] proposed a method to estimate gender, age, job, marital status, and the number of families from the use history of applications (programs) downloaded on the mobile phone.

In regard to demographic attribute prediction, Wang et al. [Wang et al., 2016] proposed a model for multi-task and multi-class prediction using purchase data. They focused on learning multiple tasks to improve prediction performance. They investigated the predictive accuracy of two prediction problems: partial label prediction and new-user prediction. Partial label prediction means predicting the remaining unknown demographic attributes. New-user prediction means predicting the demographic attributes for a new user. Wang et al. used a real world retail dataset to evaluate the prediction capability for demographic prediction and demonstrated the effectiveness of their proposed method. These studies have shown that partial demographic attributes can be predicted from purchase data. However, the estimation accuracy of the family structure from purchase data has not been investigated. For example, the quantity and type of products purchased is likely to depend on the members of the family and the number of people. It is also obvious that the presence or absence of diaper purchases is telling as are the meals selected as they depend on the age of the child. If we can estimate the family composition from the purchase history, we will be able to recommend more appropriate products and services to consumers. Therefore, this thesis proposes a method to estimate family structure from individual purchase history, and clarifies its estimation accuracy.

2.4 Purchasing intent

For many companies, to discover which of their own customers should be targeted is an important issue because they can improve the efficiency of various sales techniques. To optimize the approach taken to customers, selection of target customers such as rank up measures and measures to activate dormant customers [Hisamatsu et al., 2012] are also realized By discovering their customer to be targeted. Therefore, there is an RFM model based on purchase history that is widely used for customer analysis. This model makes it possible to estimate the customer's quality level by using Freshness (last purchase date), Frequency (purchase frequency), Monetary (purchase price) as indicators from the purchase history. If the purchase history is available, this method can to accurately determine customer level. However, in reality, there are many customers who have no purchase history, and a large percentage of them are highly desirable customers. These potential customers are considered to have the intention and ability to purchase products.

The effectiveness of analyzing and predicting the customer level from purchase history information has been clarified in many studies [Hisamatsu et al., 2012] [Matsumoto and Saigo, 2013] [Doi et al., 2018]. Research on estimating the customer level using data other than purchase history information has also been reported. Ohata et al. [Ohata et al., 2015] proposed a method of classifying customers who will purchase expensive products and customers who buy low-cost products from the purchase history information of supermarkets and the visit patterns and areas extracted from the movement history in the store. They used C4.5, which is one of the decision tree methods, to estimate the customer level from visit areas and patterns. The similar method is used in this thesis. However, rather than a single decision tree, estimation accuracy is improved by using a method of collective learning using a decision tree such as the Random Forest method. This thesis adopts a method that can estimate the level of customer desirability with higher accuracy than existing solutions. Note that the existing methods is effective only when purchases have already been actually made, they fail to support new customers. On the other hand, this thesis focuses on the behavior of customers outside the target store and proposes a method that can estimate the level of customer desirability even for new customers.

In order to grasp the behavior of customers new to the target shop, attention is paid to the fact that customer card and check-in services can provide past position information as well as current position information. The customer card holds information on the shops where purchases were made. Since it is possible to estimate the customer's living area and purchase intention from this accumulated position information, and to deliver information suitable for the customer.

Check-in information includes Shopkick [Shopkick, 2018], Rakuten check

[Rakuten-check, 2018], and Shoplier [Shoplier, 2018]. These services are used in combination with policies such as granting visiting points and providing coupons. Shopkick has obtained more than 15 million users and it is thought that Online to offline service with check-in will continue to expand, and users who will generate check-in history are expected to increase. The check-in history that can be acquired by these services is mainly recorded at the place where purchase behavior is performed. These check-in histories can be utilized to guide customers' purchasing behavior by recommending neighboring stores highly likely to be attractive to the customers, even if they are new to the store.

Several studies have already examined the processing of check-in history to create and distribute the information that meets the needs of customers and information providers.

Hayashi [Hayashi et al., 2014] proposed a method of delivering contents suitable for customers by quantitatively determining whether it is a habitual behavior or a non-habitual behavior by analyzing day and time from the check-in history. By using this method, information can be divided according to customer's behavior. If customers behave in accordance with their habit, they can provide information that is consistent with customer behavior trends. If a customer is acting in a novel way, information can be given considering the novelty. Also, by considering the history of other people whose habitual behavior tends to be similar, information to be delivered can be selected. However, this does not consider the actions taken at the places visited. There is a need on the store side to cultivate new customers. In this method, "customer who is likely to visit store A" can be estimated from the similarity of habitual behaviors with others. However, this method can not consider "good customers with high profit margin at store A"; information highly desirable to stores. Therefore, this thesis proposes a method that uses check-in histories to estimate desirable customers for particular stores. To the best of the author's knowledge, this research is the first to estimate the desirability of customers from the check-in histories of own and other stores. This thesis sets a hypothesis that the check-in history information is correlated with the purchase situation of a specific store. The validity of this hypothesis is validated, confirming that check-in histories may be useful discovering desirable customers and for developing information that can be put to practical use.

Chapter 3

Lifestyle estimation

3.1 Background

Due to the spread of the ID-POS (Point of Sales) service, detailed purchase history of consumers is now being collected and accumulated automatically. Many companies have adopted management strategy based on purchase histories acquired from ID-POS systems and analysis results include customer attribute data such as age and gender. Attempts have also been made to utilize the adoption of electronic money, credit cards, and customer cards to acquire purchase history information from companies in various fields. This makes it possible to comprehensively understand not only the local consumption behavior within each store or group store but also the consumption behavior at multiple stores visited by consumers. Rakuten Corporation [Rakuten, INC., 2018], a major e-commerce company, categorizes consumers into several groups, and presents banners for each group. They realized a method of personalizing banners for each consumer by paying attention to personal consumption behavior at multiple stores using customer attribute data and purchase histories.

Ponta [Ponta, 2018] and T point [T-Point, 2018] offer point cards and visualize potential customers and mutual customers across industries from point card usage histories at multiple stores. Such approaches that mine the consumer's purchase history are being carried out individually. Products and shop recommendations that use this purchase history information are performed in disparate business forms. However, there is a limit to the estimation of internal attributes such as consumer hobbies and values if only customer attribute data and purchase history are used.

This chapter focuses on lifestyle, which is a very important psychographic attribute. A psychographic attribute represents the one attribute of the psychological inner nature of the consumer. It is regarded as a standard for consumer segmentation. Furthermore, it is clear that it is more suitable than demographic attributes [Straughan and Roberts, 1999]. Lifestyle directly impacts consumption behavior. It has been proved that purchasing behavior depends on lifestyle [Buckley et al., 2007] [Boer et al., 2004]. Lifestyles are indicated by the purchasing behavior of daily necessities. Five features extracted from a questionnaire on the characteristics of consumers and purchasing consciousness are defined as lifestyle. Using this lifestyle information is expected to yield more advanced targeting.

Lifestyles are often acquired from questionnaires and interviews. However, it is impractical or impossible to subject all consumers to questionnaires and interviews. Therefore, a method for estimating lifestyle from products purchased by consumers by modeling the relationship between lifestyle and purchase history information has been proposed
[Koshiba et al., 2013]. This method makes it unnecessary to issue questionnaires or conduct interviews. Modern purchase history sets contain a huge number of items. For that reason, data is summarized by calculating several parameters that seem to characterize purchasing behavior such as the number of purchases by type of goods such as "Vegetables" or "Fish" and the number of visits by day of the week . They do not deal with product level purchasing information, i.e., JAN code information. However, there are differences in the types of products such as "Milk" and "Natto"

and exactly what is purchased will depend on the consumer's lifestyle. Even for the simple category of milk, differences in purchase frequency of goods such as "Full fat from natural cows" and "Low fat skim milk" may appear depending on the lifestyle. Furthermore, it is conceivable that the type of product or the level of the product also appears to depend on lifestyle. Therefore, this thesis estimates lifestyle by using purchase information at the product type level and the product level.

This thesis proposes a method to estimate lifestyle with high accuracy by extracting and using purchasing behavior associated with products in which consumer lifestyle strongly appears [Doi et al., 2017d]. Specifically, this thesis constructs individual models that realize lifestyle estimation by processing purchase information at the product unit level for each product type, selects individual models (product types) that have excellent estimation performance for each lifestyle, and combines them to yield comprehensive understanding of customer desirability. In order to better position the estimation performance of this method, this thesis examines the estimation performance of previous proposal, estimations that use only the product type information, and the proposed method. The evaluation measures are the F-measure and the correct answer rate as accuracy indicating the estimation ability of the lifestyle.

This chapter is organized as follows. Section 3.2 introduces the dataset used to construct the estimation model and evaluate the proposed method. Section 3.3 shows the detail of lifestyle. Section 3.4 explains a proposed method to estimate the lifestyle with high accuracy by extracting and using purchasing behavior on products that are thought to be strongly influenced by consumer lifestyle. Section 3.5 shows the evaluation result of the proposed method. Section 3.6 concludes the chapter with a summary and a view on future work.

3.2 Data summary

This study uses the data gathered by Intage Single Source Panel (i-SSP) dataset provided by INTAGE Inc., Tokyo, Japan for the commercial marketing. In addition to product purchase information, demographic information as attribute information of each monitor and questionnaire information on personality and purchase consciousness are included in the data. The number of monitor is 7,023. This dataset comprises the purchase history of daily necessities for a one-year period from January 1, 2012 to December 31, 2012. The data was gathered from all prefectures in Japan.

The demographic information contains 38 kinds of attributes, but only gender and age are used in this study. The attribute distribution (gender, age) of the target user is shown in Table 3.1.

Table 3.1: Attribute distribution of target user

Age	Male	Female	Total
10s	59	56	115
20s	450	497	947
30s	837	930	1767
40s	1007	997	2004
50s	733	660	1393
60s	509	288	797
Total	3595	3428	7023

Questionnaire were collected using approx. 1,600 questions on thinking and action behavior. Questions include those related to the perception of life and personality. Examples of questions about the perception of life are "I care about the calorie level of food." and "I choose meals considering nutritional balance". Examples of questions about personality are "I want to challenge new things more quickly" and "I dislike being in noisy places".

The item purchase information is a collection of records consisting of consumer ID, date and time of the purchase, product type, JAN code, the quantity, the price, and the store used. A record corresponds to one purchase of one product. Products to be handled in this study are Fast Moving Consumer Products excluding fresh foods. Each product is classified into one of 286 product types. The number of products, which is identified with JAN codes is 193,601. In this product purchase information, all purchases are recorded comprehensively for the target products.

The shop entry is just brand name, not the individual branch name. For example, convenience store A "Yokosuka branch" and convenience store A "Tameike Sanno branch" are both stored as a convenience store A. Store indicates a business handling daily consumable goods. The term store covers supermarkets, convenience stores, home centers, discount stores, pharmacies and drug stores, sake discount stores, department stores, vending machines, home delivery and mail order, 100 yen or 99 yen shops, home electronics mass merchandisers, bakery and confectionery stores, KIOSK , Liquor store and university co-op stores.

3.3 Lifestyle

Lifestyle is a feature derived using factor analysis from the questionnaire responses mentioned in Section 3.2. This method is the same as that proposed by Ishigaki et al. [Ishigaki et al., 2011a]. It is desirable to obtain a lifestyle that is valuable for the use intended. Since the lifestyle addressed in this thesis is mathematically obtained from the responses of the questionnaire, the lifestyle differs depending on the question and response. Also, it is important to note that it does not always match lifestyle that is effective for the use intended.

Specifically, 17 questions shown in Table 3.2 similar to the 20 question items used by Ishigaki et al.[Ishigaki et al., 2011a] were extracted from the questionnaire information (about 1,600 questions) included in the data described in Section 3.2.

Factor analysis was performed on the responses to the extracted questions. The responses followed the multiple choice formula of “Applicable”, “Well applicable”, “Neither”, “Not very applicable”, “Not applicable”. These options were scored 5, 4, 3, 2, 1 , respectively. Ishigaki et al. [Ishigaki et al., 2011a] used question items of “I have a household account book”, “I want to shop as soon as possible” and “There are items that I only buy at the main store”. Since there were no similar questionnaire items, this thesis does not consider these factors.

Factor analysis extracted common factors using the varimax method. The result of factor analysis on characteristic feature of lifestyle are shown in Table 3.3. “F” means factor in this table. Using scree standards, this thesis confirmed the Scree plot and determined the number of factors to 5. Consumers are thought to have multiple features with regard to lifestyle. For ease of understanding, the consumer’s lifestyle is plotted on one axis with the most characteristic feature in this thesis. This method is similar to that taken by [Ishigaki et al., 2011a].

Table 3.2: List of questionnaire items

Questionnaire number	Item
Q1	I am careful about the calorie intake in my diet.
Q2	I am cooking lunch.
Q3	Even if the price is high, I buy organic / pesticide-free vegetables.
Q4	I love cooking.
Q5	I dislike being in a noisy place.
Q6	I spend my money as much as I have.
Q7	I like food directly from the farm.
Q8	I try new products proactively.
Q9	I think it is important to have fun now.
Q10	I am eating a balanced diet of nutrition.
Q11	I want to challenge new things more and more.
Q12	I want to have a reasonable life without waste.
Q13	Products for bargain sale will be the opportunity to decide menu.
Q14	I am cautious about spending money.
Q15	I spend holidays actively.
Q16	I decide to buy product after comparing those.
Q17	I am looking for a shop selling cheaply and buying a product.

Table 3.3: Result of factor analysis on characteristic feature of lifestyle

Questionnaire					
number	F1	F2	F3	F4	F5
Q1		0.12	0.54	0.22	
Q2	-0.29		0.15		
Q3	0.89			0.15	
Q4	0.76				
Q5		0.14		-0.24	
Q6		-0.51	-0.11	0.15	0.18
Q7			0.22	0.66	
Q8			0.12	0.51	0.25
Q9					0.63
Q10			0.94	0.14	
Q11				0.16	0.46
Q12	0.80				
Q13	0.33				
Q14		0.81			
Q15				0.11	0.58
Q16		0.52			
Q17		0.29			

Table 3.4 shows the names of the features of each lifestyle and attribute distribution (age). The lifestyles are given subjective name in this study.

Table 3.4: Distribution of attributes of each lifestyle

Factor	lifestyle	10s	20s	30s	40s	50s	60s	Total
1	Strong preference	72	368	695	749	542	356	2782
2	Money saving	19	321	526	578	433	203	2080
3	Nutritional balance	18	154	318	427	294	195	1406
4	New items	1	47	119	147	79	25	418
5	Active	5	57	109	103	45	18	337
	Total	115	947	1767	2004	1393	797	7023

3.4 Lifestyle estimation by using purchasing history data

This section proposes a method to estimate the lifestyle with high accuracy by extracting and using purchasing behavior on products that are thought to be strongly influenced by consumer lifestyle.

3.4.1 Overview of the proposed method

In the approach that uses all product purchase information the lifestyle-dependent differences can be suppressed. This is because the extent to which differences in lifestyle appear in the purchase of products varies depending on the type of products and products. Therefore, when all the product purchase information is used, a full mixture of various product purchase information is obtained. In addition, lifestyles may also appear in purchased time periods, days of the week, shops, and prices which present in individual product purchase information. Therefore, this thesis takes the approach that extracting effective features from the product purchase information for each product type for each visit of the consumer may permit the lifestyle of the consumer to be predicted with high accuracy.

Product purchase information of each product type is treated as one record in the data set, and the extent to which each product type demonstrates each lifestyle is clarified. Then, this thesis proposes a method to estimate lifestyle by extracting only the product purchase information of the product type with high estimation performance for each lifestyle. Since lifestyle features are presented for each product, this thesis treats purchase history for each product as one record.

Figure 3.1 outlines the proposed method. The proposed method constructs a model for estimating lifestyle for each product type. It is constructed in the pre-analysis phase and the estimation phase. In the pre-analysis phase, models with high

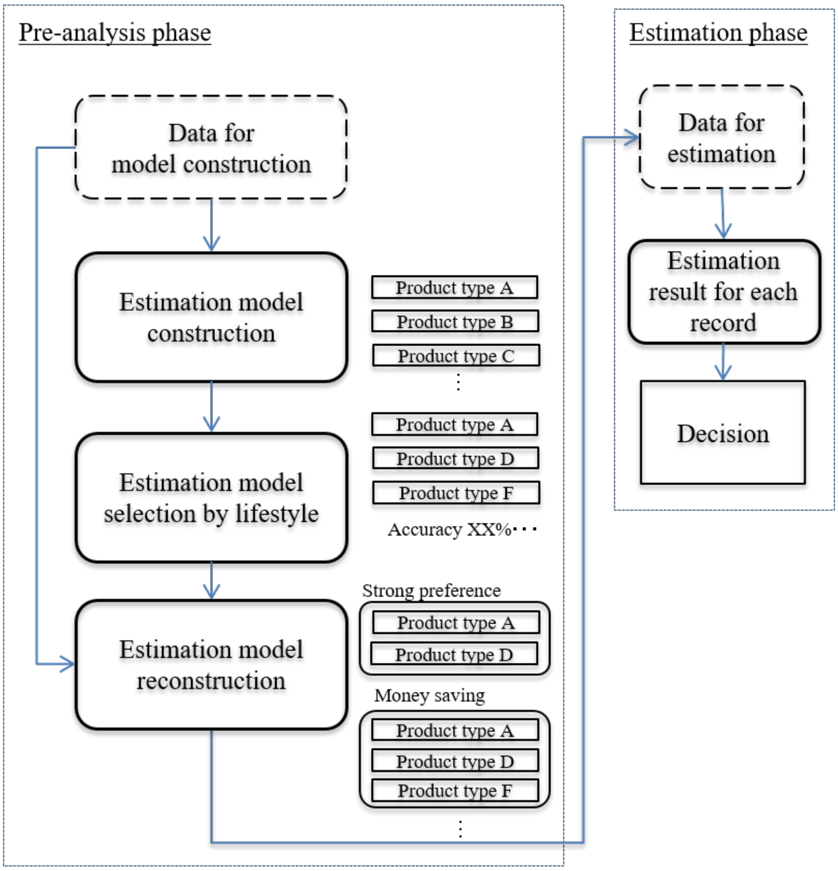


Figure 3.1: Lifestyle prediction method

estimation accuracy are automatically selected. In the estimation phase, estimation is performed using the model selected in the pre-analysis phase. Details of each phase are shown in subsections 3.4.4 and 3.4.5, respectively.

The lifestyles were derived using the responses to the questionnaire on consumer character and purchasing consciousness as described in Section 3.3. As such, the resulting lifestyles are not truly universal. In reality, it is necessary to design a questionnaire that can obtain the lifestyles suitable for the products and services to be recommended. The proposed method is applied to various lifestyles because it extracts product types effective for estimation according to the target lifestyle. It is expected to be applicable to various marketing applications.

3.4.2 Dataset

For model construction data and estimation data used for lifestyle estimation, purchase situation records are extracted from the purchase history data described in Section 3.2. The purchase situation record has information on the ID of the consumer, purchase date and time, JAN code of the product, purchase quantity, unit price and shop ID. Table 3.5 shows data items, dimensions, and values used in this thesis. The extracted purchase situation record is converted into the form of the following variable and used as data for model construction or estimation. A purchase situation record is information on a product in a purchase action. The number of JAN codes and the number of shop ID in Table 3.5 are indicated by the values appearing in the extracted purchase situation records.

Table 3.5: Data items for predicting the lifestyle

Data item	Dimensions	Value
JAN code	（Number of JAN code）	Indicator
Purchase quantity	1	Scalar
Unit price	1	Scalar
Three divisions of a month （Early・Middle・Late）	3	Indicator
Day of week	7	Indicator function
Time zone（by the hour）	24	Indicator
Shop ID	（Number of shop ID）	Indicator

3.4.3 Evaluation index

Precision, Recall, F-measure, and accuracy are used as measures for evaluating the performance of the estimation model in this thesis. Precision, Recall and F-measure are used for confirming the estimation ability of each estimation model. However, not all consumers are purchasing products of all product types. Even if the estimation accuracy of the model using the purchase history of the product type A is high, it indicates that some customer's lifestyles can not be estimated when the product type A is not purchased. Therefore, accuracy which considering the total number of consumers are also used to check the estimation ability.

The precision P(l, p) of lifestyle $l \in \{$ Strong preference, Money saving, Nutritional balance, New items, Active$\}$ using model p is calculated by equation 3.1 where $T(l,p)$ is the number of people estimated to have lifestyle l among the consumers who have purchased of product type p and $S(l,p)$ is the number of people who were estimated to have lifestyle l who actually had lifestyle l among the consumers who purchased product type p. Recall $R(l,p)$ is calculated by using equation 3.2. Depending on the product type, products are purchased only by some consumers, not all consumers. Recall $R(l,p)$ is calculated for all consumers, not just purchasers. $M(l)$ is the number of consumers having lifestyle l present in the estimation data.

F-measure is the harmonic mean of Precision $P(l, p)$ and Recall $R(l, p)$. F-measure is calculated by using equation 3.3. Accuracy $A(p)$ of the purchase of product type p among all consumers is calculated by equation 3.4 using the number of consumers N_p with correct lifestyle and the total number of consumers, O.

$$P(l,p) = \frac{S(l,p)}{T(l,p)} \tag{3.1}$$

$$R(l,p) = \frac{S(l,p)}{M(l)} \tag{3.2}$$

$$F(l,p) = \frac{2 \cdot P(l,p) \cdot R(l,p)}{P(l,p) + R(l,p)} \tag{3.3}$$

$$A(p) = \frac{N_p}{O} \tag{3.4}$$

$$N_p = \sum_l S(l,p) \tag{3.5}$$

$$O = \sum_l M(l) \tag{3.6}$$

3.4.4 Pre-analysis phase

The pre-analysis phase is shown in Figure 3.1. This phase consists of three steps: estimation model construction by product type, estimation model selection by lifestyle, and estimation model reconstruction.

In the first step for constructing estimation model, a hierarchical neural network (NN) [Riedmiller, 1994], which is one method of machine learning, is used to construct an estimation model that outputs a lifestyle upon the input of a purchase situation record for each product type. The reason for using a NN to construct the estimation model is that it yields higher estimation accuracy than the estimation result of the lifestyle yielded by previous study [Doi et al., 2015].

Result B of the NN for each purchase situation record indicated by equation 3.8 is a vector whose element is the assignment probability b_l for lifestyle l. The average of the assignment probability b_l of lifestyle l in product type p for consumer u is calculated using equation 3.7. rn_{up} is the number of records of consumer u that show product type p. L is the total number of lifestyles. B_{upi} is the output of the NN for the input of the i th purchase situation record for product type p of consumer u. Lifestyle that the highest affiliation probability was high among the average values of assignment probabilities of each lifestyle which A_{up} has as an element is taken as the lifestyle of the consumer.

$$A_{up} = \frac{1}{rn_{up}} \sum_{i=1}^{rn_{up}} B_{upi} \tag{3.7}$$

$$B = (b_1, \cdots, b_L). \tag{3.8}$$

The back propagation method is used as NN learning method. The hidden layer is one layer, the number of units of the hidden layer equals the number of units of the input layer. The number of dimensions of each input signal (the number of units of the input layer) is shown in Table 3.5. All weights of data are set to the same weight for learning. In order to suppress the increase in coupling load, a random number is set as the initial value of the weight of the weight decay term added to the objective function. The activation function of the hidden layer uses a sigmoid function. The activation function of the output layer uses a softmax function. To evaluated the performance of the estimation model, 10-fold cross validation is adopted. For model building data, 90% of consumer data is taken to be learning data and the remaining

10% was used as evaluation data. In order to evaluate each estimation model, the F-measure $F(l, p)$ is calculated.

In the second step, estimation model selection by lifestyle, several estimation models with high estimation accuracy for each lifestyle were selected from the estimation models for each product type constructed in the first step. Combining these selected models yielded an estimation model with high estimation performance. In this step, this thesis uses the F-measure $F(l, p)$ in each lifestyle of the estimation model and the average assignment probability A_{up} for each consumer processed by the estimation model. Combining the models of multiple product types means averaging the A_{up} values obtained from each product type.

The greedy algorithm is used in estimation model selection by lifestyle. Estimating lifestyle l by combining the product type models with the highest n in the F-measure $F(l, p)$ for each lifestyle yields a new F-measure $F(l, p)$. Start n from 1 and observe the F-measure. Next, increase n by 1. Continue to increment n until the improvement in F-measure saturates. At the saturation point, the estimation model that combines the top n th item types is adopted as the lifestyle model. The resulting set of product types is denoted as P_l. This thesis performs these procedures for all lifestyles.

In the first step, 90% of the data is used for constructing the estimation model. In the third step, the model construction data adopted in the second step is set as learning data. When a purchase situation record is input to the first step, an estimation model that outputs a lifestyle is constructed by using this data.

3.4.5 Estimation phase

The estimation phase of the proposed method shown in Figure 3.1 is described in this section. In the estimation phase, this thesis uses the estimation model constructed in the pre-analysis of Section 3.4.3. The estimation data is taken as the product purchase history of the consumer whose lifestyle is to be estimated. For consumer u, estimate the assignment probability of lifestyle l as follows. Extract records of product type $p \in P_l$ purchased by consumer u and obtain A by equation 3.7 for each p. This thesis uses equation 3.9 to obtain these averages. This is denoted by C_{ul}. l is taken as the lifestyle of the consumer when l matches the lifestyle with the highest assignment probability in C_{ul}. This thesis applies the above process to all consumer u and lifestyle l combinations. Because lifestyle estimation is done for each lifestyle, one consumer be assigned multiple lifestyles. In this case, the lifestyle with the highest assignment probability is selected using the corresponding C_{ul}. When the assignment probabilities are the same, the number obtained by dividing 1 by the estimated number of lifestyles is counted as the number of consumers.

$$C_{ul} = \frac{1}{|P_l|} \sum_{p \in P_l} A_{up} \tag{3.9}$$

3.5 Evaluation

This section evaluates the effectiveness of the proposed method by comparing the estimation performance of the previous study and the proposed method described.

3.5.1 Estimation accuracy of previous study

This study focuses on individual product levels as distinguished by JAN code to estimate the lifestyle of consumers. Koshiba et al. [Koshiba et al., 2013] estimate lifestyles by calculating and modeling several parameters characterizing purchasing behavior. This is because each purchase history has a huge number of items. As the method to estimate the lifestyle, the method proposed by Koshiba et al. [Koshiba et al., 2013] can be used. Therefore, Koshiba's proposed method [Koshiba et al., 2013] is compared to confirm the usefulness of the proposed method.

Since the estimation accuracy is considered to be depended on the data, the data used in this thesis is used for the evaluation of the previous study [Koshiba et al., 2013]. Product purchase information is counted for each person and individual purchase histories are not used in the previous study. Koshiba et al. [Koshiba et al., 2013] assumes the use of a purchase history for a single store.

The data examined in this thesis includes the purchase histories of multiple stores. For that reason, the number of visits to each store is calculated and used by using the store ID.

The data used in the evaluation of the previous research is shown in Table 3.6. The number of visits per time zone is a hour. This purchasing information also includes 24-hour stores. Therefore, this thesis adopted the time unit of one hour in this evaluation.

Table 3.6: Data used for evaluation of previous study

Data	Dimension	Value
Number of visits by store ID	485	Scalar
Number of visits (by day of the week)	7	Scalar
Number of visits (hourly)	24	Scalar
Product purchase price (total)	1	Scalar
Number of items purchased (total)	1	Scalar
Food purchase price (total)	1	Scalar
Food purchase (total)	1	Scalar
Number of products purchased by product type (total)	286	Scalar

Estimation of lifestyle is done using the random forest method [Breiman, 2001] as in previous study[Koshiba et al., 2013]. Evaluation is carried out by the 10 fold cross validation method, and the F-measure and accuracy are confirmed. In cross validation, 90% of consumer's data is used for model construction and 10% of consumer data was used for evaluation. The result of estimating the lifestyle by the method is shown in Figure 3.2 as "Prior". It is confirmed that lifestyle can be estimated with an accuracy of 37.6% by using a method in previous study.

3.5.2 Estimation accuracy of the proposed method

In order to confirm the operation of the proposed method, this section shows the estimation performance gained using only the product type. Moreover, the estimation performance of the proposed method is clarified and its effectiveness is shown.

Estimation accuracy of the proposed method

Figure 3.2 shows the results achieved when using the top five product types with high rates of accuracy. This thesis uses this result to clarify the estimation performance possible for each product type. The accuracy of lifestyle estimation estimated by

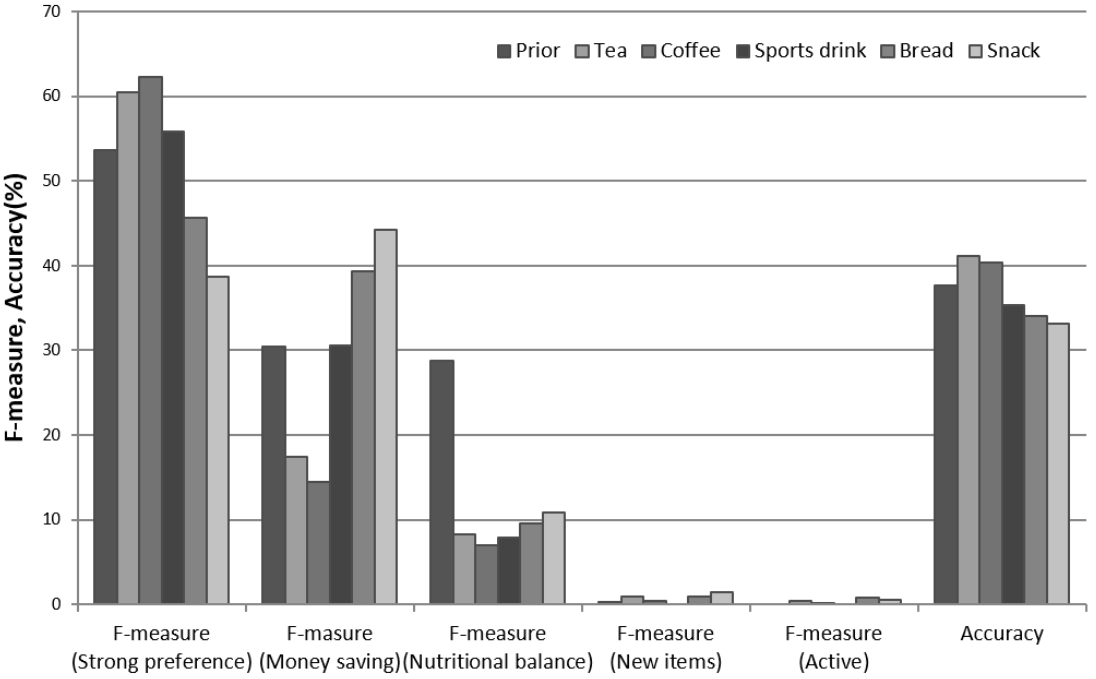


Figure 3.2: Comparison of lifestyle prediction accuracy (F-measure, Accuracy rate)

purchasing information of each product type (Using the average assignment probability $A_u p$ shown in equation 3.7) is shown. The accuracy for each type of product was 41.1% for “Tea”, and 40.3% for “Coffee”. This confirms that the proposed method estimates lifestyle more accurately than the previous study ”Prior”.

“Strong preference”, “Money saving”, “New items” and “Active” reveal the type of products that can be estimated with higher accuracy than possible in the previous study. However, it was confirmed that “Nutritional balance” had lower F-measure than in the previous study.

Table 3.7 to Table 3.11 show the types of products with high F-measure for each lifestyle.

Table 3.7: Product type and accuracy (F-measure) for strong preference estimation

Product type	F-measure（%）
Coffee	62.2
Tea	60.5
Sports drink	55.9
Fruit juice drink	52.0
Soda pop	51.1
Cola	49.6
Mineral water	49.2
Tea drink	48.9
Nutritious drink	47.2
Energy drink	45.8

Table 3.8: Product type and accuracy (F-measure) for money saving estimation

Product type	F-measure（%）
Tofu	48.9
Bread	48.3
Raw noodle	47.3
Yoghurt	45.6
Milk	45.4
Frozen meals and dinners	44.4
Snack	44.3
Chinese style food	43.9
Natto	43.8
Biscuit and Cracker	43.6

Table 3.9: Product type and accuracy (F-measure) for nutritional balance estimation

Product type	F-measure（%）
Cheese	27.9
Salad oil and Tempura oil	24.0
Pickle	24.0
Softening agent	23.9
Wrapping film	23.4
Other seasoning	23.2
Natto	23.0
Spice	23.0
Ham	23.0
Tofu	22.6

Table 3.10: Product type and accuracy (F-measure) for new items estimation

Product type	F-measure（%）
Lip cream	5.8
Other medicine for skin	5.0
Sponge	5.0
Lipstick	4.8
Skincare article	4.7
Other general merchandise	4.7
Eyebrow pencil	4.5
Syrup	4.4
House detergent	4.4
Foundation (Cosmetic)	4.3

Table 3.11: Product type and accuracy (F-measure) for active estimation

Product type	F-measure（%)
Other general merchandise	6.1
Miso-soup	4.5
Insecticide	4.1
Combined seasoning	3.9
Chinese tea	3.9
Test kit	3.7
Other canned food	3.5
Barely tea	3.2
Etiquette article	2.9
Candy	2.8

It is confirmed that the product types most clearly expressing lifestyle differs for each lifestyle. Strong preference was demonstrated by “Coffee drink” and “Liquid tea”. Money saving appeared in “Tofu” and “Bread”. Also, there was a difference in maximum value of F-measure depending on lifestyle. These results indicate that there are product types that are suitable for estimation and product types that are not suitable depending on the target lifestyle.

The difference in estimation accuracy due to lifestyle seems to be influenced by bias in the number of consumers having each lifestyle. Therefore, this thesis constructed and evaluated an estimation model with balanced learning data, each lifestyle had the same number of consumers. This increased the F-measure of active and new items are increased. This result suggests that imbalance in the number of consumers having each lifestyle may affect estimation accuracy. Compared with the case of not using balanced learning data, the types of products with high lifestyle estimation performance were generally the same. Based on this result, this thesis confirmed that it is possible to select the product type that can most accurately identify each lifestyle targeted even if bias in consumer number is present.

To clarify the how accurately the model of estimation by product type can be learned by NN, the learning accuracy and the evaluation accuracy are evaluated in Figure 3.2. The accuracy is the value obtained by dividing the number of records in which the correct lifestyle is estimated by the total number of records evaluated. The learning accuracy is the correct answer rate calculated by using the learning data used for the evaluation of NN for evaluation. The evaluation accuracy is the correct answer rate calculated using the evaluation data.

Table3.12 to Table3.16 show the learning accuracy and the evaluation accuracy of the model for the product type in each lifestyle. Strong preference, Money saving and Nutritional balance have learning accuracies of 40 to 60 % and evaluation accuracies of 30 to 50 %. New items and Active and have learning accuracies of 50 to 90 % and evaluation accuracies of 20 to 30 % . Compared to Strong preference, Money saving and Nutritional balance, the learning accuracies of New items and Active are high, while the evaluation accuracies are low. In order to consider the correspondence between learning accuracy and evaluation accuracy, the learning accuracy is divided by the evaluation accuracy to yield the indicator of correct answer ratio.

Table 3.12: Estimation performance of strong preference (learning accuracy, evaluation accuracy)

Product type	Learning accuracy（%）	Evaluation accuracy（%）
Coffee	59.6	52.2
Tea	46.5	46.0
Sports drink	44.8	46.2
Fruit juice drink	47.3	43.3
Soda pop	43.7	39.7
Cola	43.0	42.4
Mineral water	43.9	39.1
Tea drink	43.3	40.4
Nutritious drink	48.0	46.0
Energy drink	46.5	44.2

Table 3.13: Estimation performance of money saving (learning accuracy, evaluation accuracy)

Product type	Learning accuracy （%）	Evaluation accuracy （%）
Tofu	38.3	38.2
Bread	39.7	36.3
Raw noodle	40.8	35.9
Yoghurt	40.6	36.5
Milk	44.0	36.0
Frozen meals	40.7	35.7
Snack	45.0	37.2
Chinese style food	42.9	38.3
Natto	38.5	35.4
Biscuit and Cracker	42.9	33.7

Table 3.14: Estimation performance of nutritional balance (learning accuracy, evaluation accuracy)

Product type	Learning accuracy （%）	Evaluation accuracy （%）
Cheese	38.5	33.1
Salad or Tempura oil	45.1	31.7
Pickle	37.7	32.3
Softening agent	42.5	31.1
Wrapping film	43.2	32.2
Other seasoning	46.7	33.5
Natto	38.5	35.4
Spice	43.4	31.3
Ham	41.2	32.7
Tofu	38.3	38.2

Table 3.15: Estimation performance of new items (learning accuracy, evaluation accuracy)

Product type	Learning accuracy （%）	Evaluation accuracy （%）
Lip cream	56.5	29.5
Other medicine for skin	59.4	27.7
Sponge	62.9	31.4
Lipstick	99.0	32.0
Skincare article	61.3	29.1
Other merchandise	67.9	28.6
Eyebrow pencil	87.8	29.4
Syrup	60.8	27.8
House detergent	58.5	31.3
Foundation (Cosmetic)	90.6	26.5

Table 3.16: Estimation performance of active (learning accuracy, evaluation accuracy)

Product type	Learning accuracy （%）	Evaluation accuracy （%）
Other merchandise	67.9	28.6
Miso-soup	46.2	33.1
Insecticide	61.0	28.2
Combined seasoning	50.9	28.8
Chinese tea	64.8	33.3
Test kit	98.6	25.6
Other canned food	59.0	25.0
Barely tea	49.3	33.5
Etiquette article	81.0	28.6
Candy	47.7	34.0

Figure 3.3 plots each product type against the number of purchasers of it and the correct answer ratio. The nearer the correct answer ratio is to 1.0. If the correct answer ratio is low, over learning is suspected.

For product types with low F-measure, the correct answer ratio tended to be low. These result was confirmed that the estimation accuracy and the correct answer ratio were related. The ratio of correct answers tended to be closer to 1.0 for product types purchased more frequently.

For product types with low correct answer ratio values, it is conceivable that the data volume available for learning the features of lifestyle was too slight or that the product types were less likely to show lifestyle features. There is a possibility that the correct answer ratio will be improved by increasing the amount of data available for learning.

Estimation performance of the proposed method

Figure 3.4 shows the accuracy of the proposed method when estimating lifestyle using only product types with high lifestyle estimation accuracy. As a result of a preliminary analysis using the lifestyle estimation model, 5 product types for Strong preferences, 5 product types for Money saving, 4 products for Nutritional balance, 6 products for new items and 5 products for active models were adopted and combined as a final model. The accuracy rate of the proposed method reached 44.0%. This thesis confirmed that the correct answer rate is improved by selecting the product type appropriate for each lifestyle.

In addition, when using the method of the previous research and single product type, remarkably low F-measure were achieved for new items and active and estimation was problematic. However, the proposed method greatly improved F-measure to 10 pt or more in both cases. This reflects the benefits of using detailed product purchase histories appropriately.

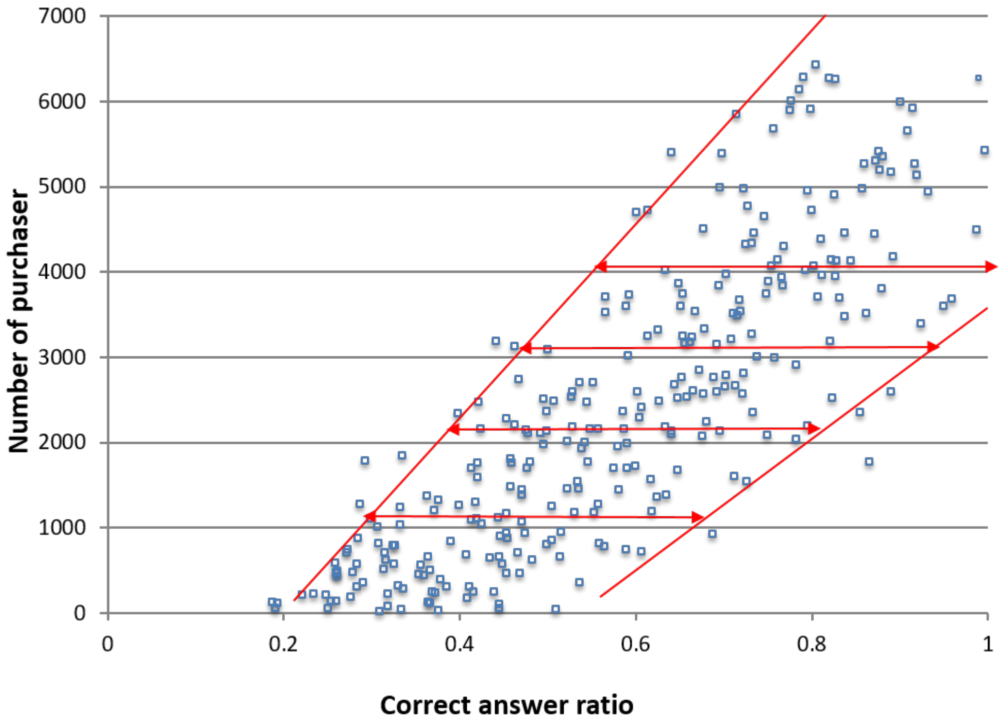


Figure 3.3: Number of purchaser and correct answer ratio

From these results, the effectiveness of the proposed method was confirmed. This thesis assumed that the proposed method is to be applied mass markets with millions of users. Therefore, it is considered that even small increases of a few points in the correct answer rate greatly increases the number of users for whom we can achieve correct targeting. The proposed method is useful because it offers 6.4 pt higher accuracy than the previous research.

By extracting purchasing behaviors concerning products that strong indicate consumer's lifestyle, this method makes it possible to estimate the lifestyle of a wide variety of consumers. However, there is a difference in the estimation performance of each lifestyle, and further improvement of estimation accuracy is needed. Depending

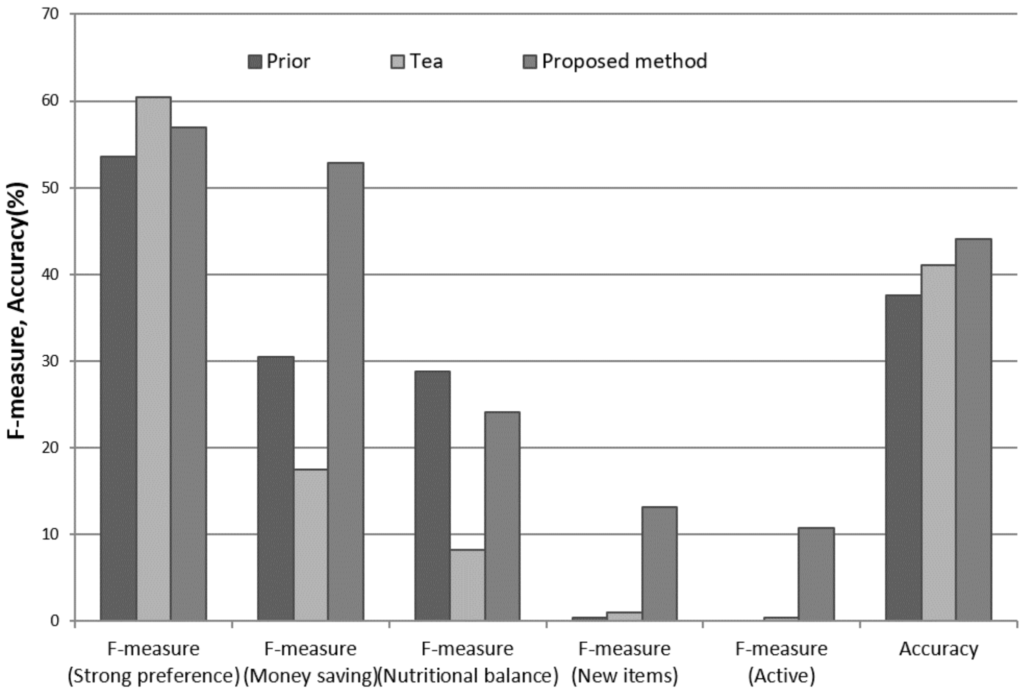


Figure 3.4: Lifestyle prediction accuracy of proposed method

on the target lifestyle, it may be difficult for extracting the effective characteristics in purchasing behavior. In such cases it will be necessary to add new information as clues.

3.6 Chapter summary

A method that can estimate lifestyle with high accuracy was proposed in this chapter. The method extracts and uses purchasing behavior concerning products that are strong indictors of lifestyle from the large numbers of products purchased by consumers. From evaluation which uses real world purchase history data, it was confirmed that lifestyle can be estimated with an accuracy rate of 44.0% by using the most appropriate target product types which are few in number. It was confirmed that the F-measure of some lifestyle estimations can be greatly improved by combining the purchase behavior of consumers across many different stores.

Accurately understanding the lifestyle of consumers makes it possible to realize more effective one to one marketing, such as recommending products and services that really suit each consumer. Note that the lifestyles examined in this chapter are comprehensive truly universal indicators, and it is necessary to design questionnaires so that the lifestyle that suits the targeted product can be obtained. However, this method has the advantage of being able to automatically extract product types effective for estimation according to lifestyles compared to those mentioned.

As future work, in order to further improve the estimation accuracy of each lifestyle, the author intends to construct and evaluate an estimation model based other methods such as the radial basis function network and the Bayesian network. Another challenge is to implement measures such as information distribution to customers by using the estimated lifestyles, and to clarify the effective features for improving the customer's unit price and purchasing frequency for each lifestyle. The purchase history used in this thesis contained goods that were rarely purchased, quite new products and discontinued products. This thesis will also consider measures to deal with changes in the status of these products and product extraction.

Chapter 4

Preference estimation for products

4.1 Background

Many companies conduct sales and advertising campaigns based on the results of analyzing service utilization histories and purchase histories acquired from credit card or electronic money companies. By analyzing own data along with other company's data, it becomes possible to discover the customer's preferences and actions to a level not possible from just own data. Credit card companies can obtain their cardholder attribute data and purchase history, which consists of just the store, date of purchase, and amount. The history does not include purchase description so exactly what the user is interested in is unknown. By using and analyzing "Like" and "Check-in" data from Facebook and Foursquare, American Express could realize the effective distribution of vouchers based on customer preferences, interests, and actions.

Several location-based services allow consumers to share physical location, the "Check-in" function, by using mobile applications on smart phones. The explicit behavior of check-in indicates the intention of the customer, and it is considered that using this data is effective for customer understanding. Facebook places

[Facebook places, 2018], Foursquare [Foursquare, 2018] and Yelp [Yelp, 2018] are examples of location-based services. Shopkick [Shopkick, 2018] is also one of the services that use the "Check-in" action. Customers are able to check into locations and redeem gift cards and vouchers as rewards. Shopkick is used by more than 15 million users. Online-to-offline services using "Check-in" are expected to expand in the future. Furthermore, it is also expected that the number of users accumulating large check-in histories will increase. The check-in data that can be acquired by these services includes where the purchase was made. By using check-in history, we can also recommend neighboring shops where customers are more likely to visit and induce purchases. If preferences of a customer who has not accumulated a purchase history could be estimated from check-in history, it would become possible to recommend products that match the consumer's preference.

On the other hand, Point of Interest (POI) recommendation in location-based services plays an important role in providing personalized recommendations of places to customers. It helps customers to find new places and filter out un-suitable places given their preference or interest. For example, a customer who often visits a delicatessen or restaurant that specializes in meat tends to prefer meat dishes. Also, customers who visit the same place may share the same or similar preferences. By sharing preferences and frequented places, we can recommend an unvisited place that meets the customer's preference.

Customer preference also plays an important role for retailers such as supermarkets or department stores as it should drive marketing strategies and provide personal recommendations. Although customer preference is generally obtained from questionnaires or purchase behavior, it is difficult to obtain personal information from all customers. Due to privacy concerns, most customers are reluctant or refuse to provide their personal information.

Previous research clarified the effect of social influence on POI recommendation [Yonghong and Xingguo, 2015] [Zhang et al., 2016] [Li et al., 2016] [Ye et al., 2011]

[Ye et al., 2010] [Huang and Dong, 2016] [Yao et al., 2016]. Since friends tend to share more interests than non-friends, it is well known that recommendation accuracy is improved by considering social effects. This thesis sets two hypotheses: 1. Customers could be grouped as friends if they have common preferences as indicated by similar purchase behaviors. 2. Customers who go to similar places may have similar preferences for foods or products . The goal is to estimate the customer's preference from their check-in history without using questionnaires.

This chapter proposes a method to estimate customer preference from check-in history [Doi et al., 2017c]. It defines customer preference clusters, which are extracted from the purchase history data, as customer preference. The thesis examines the determination of customer preference for foods and confectionaries that can be purchased in Tokyu Department Store, Shibuya station, Tokyo, Japan. To the best of the author's knowledge, this work is the first to confirm the effectiveness of estimating customer preference from check-in history.

This chapter is organized as follows. Section 4.2 introduces the dataset used to construct the estimation model and evaluate the proposed method. Section 4.3 explains how customer preference can be extracted from purchase history. The proposed method is detailed in Section 4.4. How to estimate customer preference from check-in history is explained. The performance of the proposed method is also shown. Section 4.5 describes visitor promotion trial and its results. Section 4.6 concludes the chapter with a summary and a view on future work.

4.2 Dataset

Datasets of check-in history and purchase history were used to construct a model to estimate customer preference. Permission to use these datasets was obtained from the customers. This section details the datasets used.

4.2.1 Targeted service

Shoplat is one of the services that allows customers to check into locations and earn points or digital stickers. Figure 4.1 shows an example of Shoplat application. This service is operated by NTT DOCOMO, INC., Tokyo, Japan. By downloading the free application onto their smartphone, customers can receive points that can be exchanged for gift cards or vouchers by checking in at specific places called check-in spots. Shop and event information can also be received.

The check-in spots exist in department stores, supermarkets, restaurants, drug store and karaoke places, mainly in Tokyo. We should note that a check-in spot is not placed the tenant areas (shops) of department stores.

Check-in is completed with the following two actions. The first action is to activate the application and move to the check-in spot. Then a tag with the point is displayed on the screen. Information on the place of check-in spot can be obtained by the application. The second action is to pull the tag down. The customer can acquire points after executing these actions.

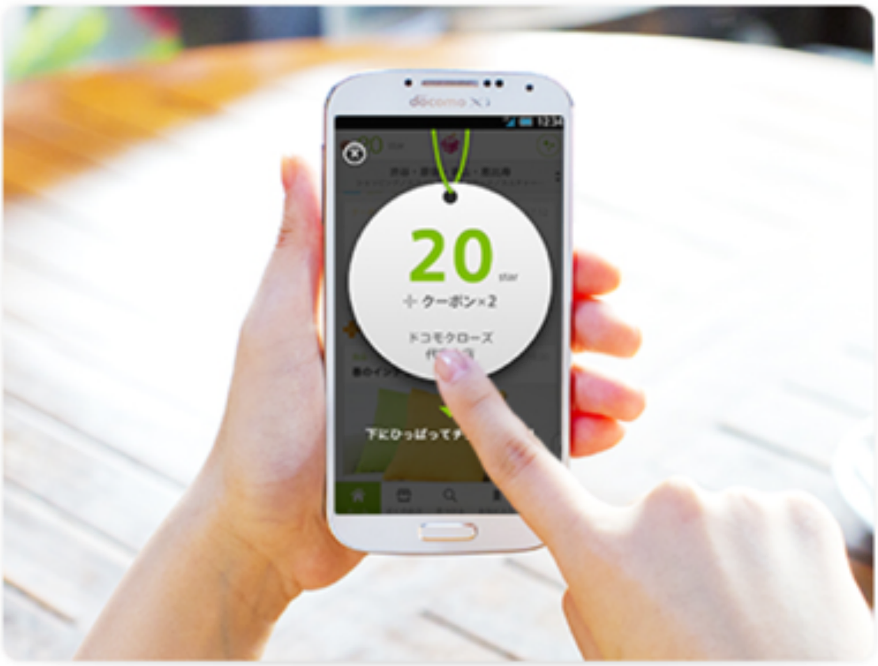


Figure 4.1: An example of Shoplat application

4.2.2 Check-in history data

The Shoplat dataset was adopted as the check-in history. A transaction comprises a Customer ID, Date, Time, Shop ID and Check-in Spot ID. The data items of the check-in history data and examples of values are shown in Table 4.1. Shop ID indicates ID numbers of a shop such as department stores A. For example, Shop ID “001” covers all branches of department store A and Shop ID “002” covers all branches of department store B. All check-in spots are in the shop, and Shop ID and Check-in Spot ID are given.

Table 4.1: List of check-in history data

Items	Value Example
Customer ID	0001
Date	2016/09/21
Time	15:00:00
Shop ID	001
Check-in Spot ID	100

4.2.3 Purchase history data

The credit card payment history of TOKYU CARD, INC. was adopted as purchase history data. Although cash payments are still common in the real world, only credit card payment is targeted in this study. This data is limited to purchases made in the Tokyu Department Store, Shibuya station. This dataset holds records for the 23 months period from 2013 to 2015. It contains 22,315 transactions made by 406 users who used the Shoplat application. The total number of specialty shops is 173. The purchases included daily necessities, delicatessen foods, and confectioneries. A transaction comprises a Customer ID, Date, Time, Tenant ID and Total price

(Price). The purchased products are not included in this transactions. Table 4.2 gives items and examples of values.

Table 4.2: List of purchase history data

Items	Value Example
Customer ID	0001
Date	2016/09/21
Time	15:00:00
Tenant ID	001
Price(Yen)	100

4.3 Customer preference

This section describes the extraction of customer preference clusters as customer preference from the purchase history data. This study considers only products from delicatessens and confectionery stores as purchased in Tokyu Department Store.

With the assistance of Tokyu Department Store staffs, the customer preference clusters of Cl.1, Cl.2, Cl.3 and Cl.4 were defined. In each cluster, this thesis confirmed that Western-style shops in Cl.1, Japanese-style shops with Cl.2, Cake shop in Cl.3 and Japanese sweets shops in Cl.4 are included mainly.

While each customer may exhibit multiple features (occupy multiple clusters), for simplicity and ease of understanding, this thesis assigns only one cluster to each customer, the cluster with the highest purchase frequency.

Table 4.3: Customer preference cluster

Cluster	Number of Persons
Cl.1	158
Cl.2	80
Cl.3	115
Cl.4	53

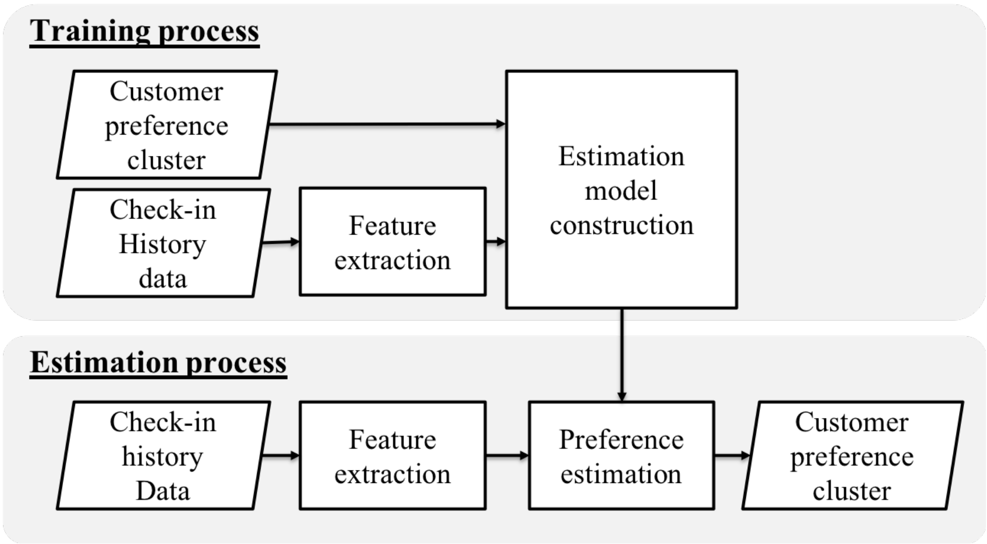


Figure 4.2: Overview of proposed method

4.4 Preference estimation

This section proposes a method to estimate the customer preference cluster from check-in history data. Figure 4.2 overviews the proposed method. The proposed method is divided into two processes: training and estimation.

In the training process, estimation models that estimate the customer preference cluster are constructed using multiple machine-learning algorithms. From these constructed models, the model that provides the highest estimation accuracy and F-measure is selected as the estimation model.

In the estimation process, customer preference clusters are estimated based on the estimation model acquired in the training process.

4.4.1 Feature extraction

The 5 features extracted from the check-in history data are listed in Table 4.1. The extracted features and number of dimensions are given in Table 4.4 . Check-in activities for each consumer are aggregated into a single transaction. These transactions are defined as input data. The items "Day", "Day of the week" and "Time" indicate when and how many times a consumer checked in.

Table 4.4: List of extracted features

Features	Number of Dimensions
Day of month	31
Time	24
Day of week	7
Check-in Spot ID	(Number of the Check-in spot)
Total number of check-in	1

4.4.2 Estimation model construction

The customer preference cluster is shown in Section 4.3. The extracted features shown in Table 4.4 are the input data used to construct the estimation model. The customer preference cluster and the check-in history data are linked via Consumer ID.

Estimating the customer preference cluster is a form of classification problem. This study adopts Random Forest [Breiman, 2001], Support vector machine [Boser et al., 1992] and Logistic regression [David, 1958] for model construction to solve this classification problem. This thesis uses the criteria proposed by Breiman [Breiman, 2001] to set parameters for Random Forest. For Support vector machine, the RBF kernel is used as the kernel function. After comparing the estimation ability, the constructed model that provides the highest F-measure is selected as the final estimation model and used in the estimation process to determine the customer's preference cluster.

4.4.3 Estimation model evaluation

In order to evaluate the accuracy of customer preference cluster estimation, this thesis uses 10-fold cross validation. The 10-fold cross validation works as follows: Nine subsamples are used for constructing the estimation model as learning data and the remaining subsample is used as test data to evaluate the estimation model. F-measure is used as the evaluation index. F-measure is a harmonic mean of precision and recall. Let i be the value of the attribute. The value i are $i \in \{Cl.1, Cl.2, Cl.3, Cl.4\}$. $U(i)$ is the number of correct estimations for attribute value i. $V(i)$ is the number of target value estimations for attribute value i. $W(i)$ be the total number of target values.

Precision $P(i)$ and Recall $R(i)$ are computed as

$$P(i) = \frac{U(i)}{V(i)} \quad , \tag{4.1}$$

and

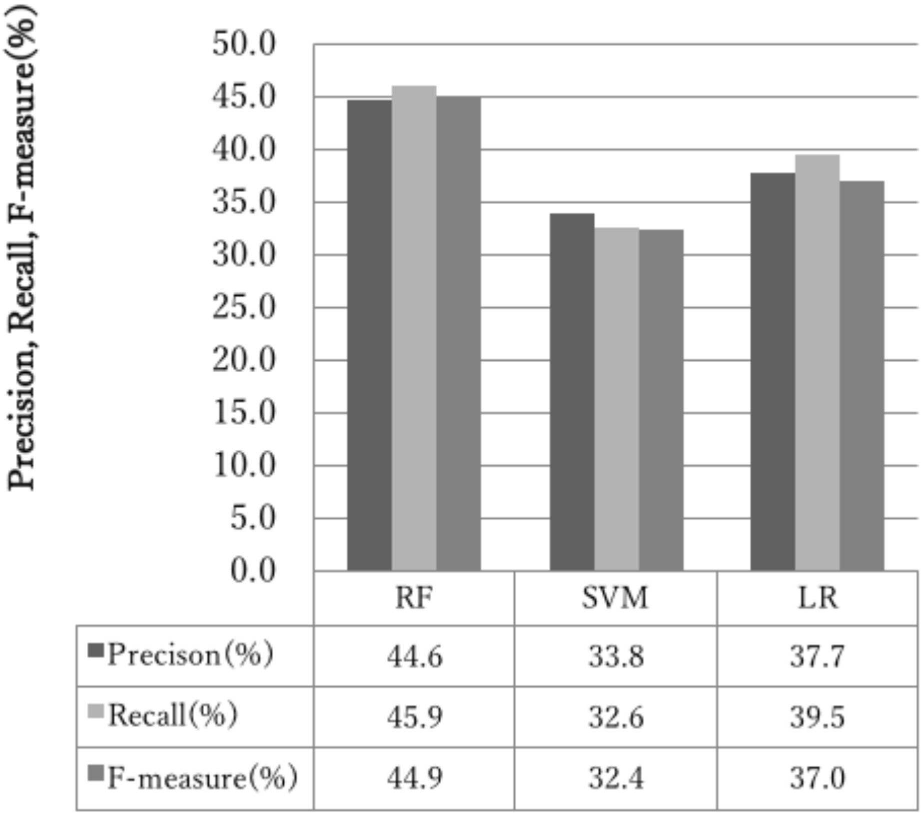
$$R(i) = \frac{U(i)}{W(i)} \quad . \tag{4.2}$$

F-measure $F(i)$ is given by

$$F(i) = \frac{2 \cdot P(i) \cdot R(i)}{P(i) + R(i)} \quad . \tag{4.3}$$

Figure 4.3 shows average estimation accuracy of each algorithm. RF represents Random Forest, SVM represents Support vector machine and LR represents Logistic regression.

A comparison of accuracy shows that RF estimated the customer preference

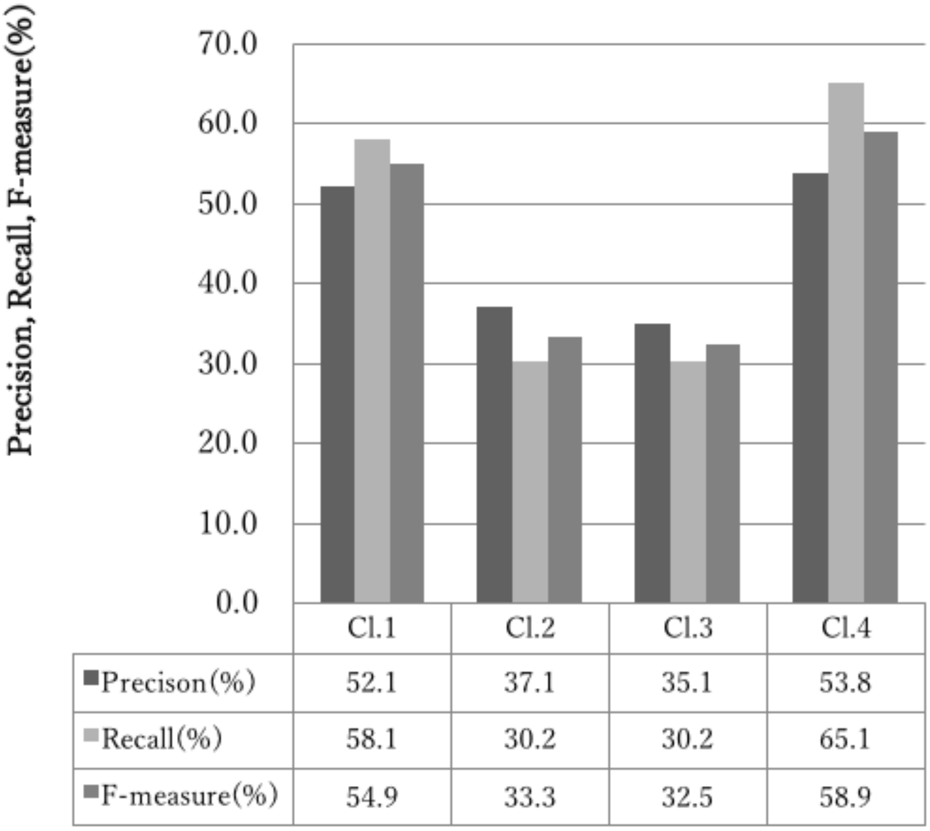


	RF	SVM	LR
■Precison(%)	44.6	33.8	37.7
■Recall(%)	45.9	32.6	39.5
■F-measure(%)	44.9	32.4	37.0

Figure 4.3: Average estimation accuracy of each cluster

cluster with higher accuracy (44.9%) than the other algorithms. Figure 4.4 shows the estimation accuracy of the RF algorithm for each cluster. The results show that Cl.4 was the easiest to estimate among all clusters. The second best was Cl.1. Cl.1 covered shops selling western style dishes and Cl.4 covered Japanese sweets shops. This confirms that there are clusters that are easy to estimate.

Next, the accuracy of the estimated customer preference clusters was confirmed. For the case of random cluster assignment, the estimation accuracy rate is considered to be 25%. It seems that using check-in history to estimate the customer preference cluster is more effective than random selection.



	Cl.1	Cl.2	Cl.3	Cl.4
■Precison(%)	52.1	37.1	35.1	53.8
■Recall(%)	58.1	30.2	30.2	65.1
■F-measure(%)	54.9	33.3	32.5	58.9

Figure 4.4: Accuracy rate of each customer preference cluster

4.5 Evaluation

In order to show the effectiveness of the proposed method, a visitor promotion event was held at Tokyu Department Store, Shibuya station. The purpose of this event was to increase the number of people visiting the store. Details of the visitor promotion and the results are described below.

4.5.1 Visitor promotion

The Shoplat application was used to distribute information, see Figure 4.5 for some translated examples. The information (written in Japanese) was selected by Tokyu Department Store.

The target customers are customers who used the Shoplat application and who had accumulated a check-in history. The number of target check-in spots was set to 873; customers who had both purchase history and check-in history could register at any spot. There were 8,863 target customers, people for whom only check-in history was available.

Cl.1

【Food Show in Tokyu department store】
Shibuya gastronomic festival

Fried rice spare ribs in a bowl
1,296 yen

This is the ultimate dish which mixed spareribs soaked in secret sauce over fried rice.

Cl.2

【Food Show in Tokyu department store】
Shibuya gastronomic festival

Grilled rice with wheat Fuji pork
670 yen

Fuji pork soaked in a sauce of grilled chicken overnight and baked softly. It is a rice bowl with well-balanced combination of meat and autumn vegetables.

Cl.3

【Food Show in Tokyu department store】
Shibuya gastronomic festival

New chestnut Mont Blanc
648 yen

This Mont Blanc uses new chestnut paste from Kumamoto prefecture. Chestnuts and fresh cream are the perfect combination.

Cl.4

【Food Show in Tokyu department store】
Shibuya gastronomic festival

Kuri Kinton
1,296 yen

Steaming light green chestnuts, sweet and comfortable scent. Chestnuts with rich taste carefully cooked. In combination with white beans, this is readily melts in the mouth.

Figure 4.5: Information distributed for visitor promotion

4.5.2 Effect of the visitor promotion

The effect of the visitor promotion was confirmed by comparing the view rate of delivered information and visit rate. The view rate is calculated using equation 4.4 where d is the number of customers who received the delivered information and e is the number of customers who viewed the delivered information on the Shoplat application. The visit rate was calculated by equation 4.5 where g is the number of people who visited Tokyu Department Store, *Shibuya station Toyoko store* after browsing the delivered information. The value of g does not include the following two cases. First, the customer visited a target shop without viewing delivered information. Second, the customer visited a target shop before viewing distributed information.

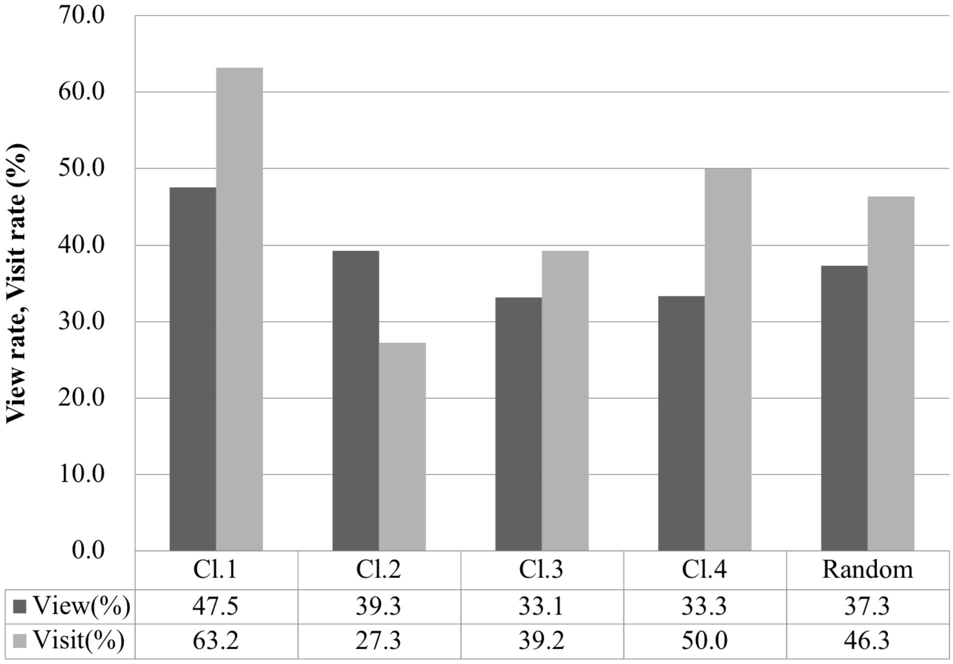
$$Viewrate = \frac{e}{d} \tag{4.4}$$

$$Visitrate = \frac{g}{e} \tag{4.5}$$

In order to confirm the effectiveness of the proposed method, this thesis considered the view rate and the visit rate for each cluster. By using purchase history, the correct customer preference cluster could be identified "Random" covers the users who received one of distributed information in Figure 4.5 without regard to customer preference clusters. "Random" consist of customers who were extracted and configured from each cluster.

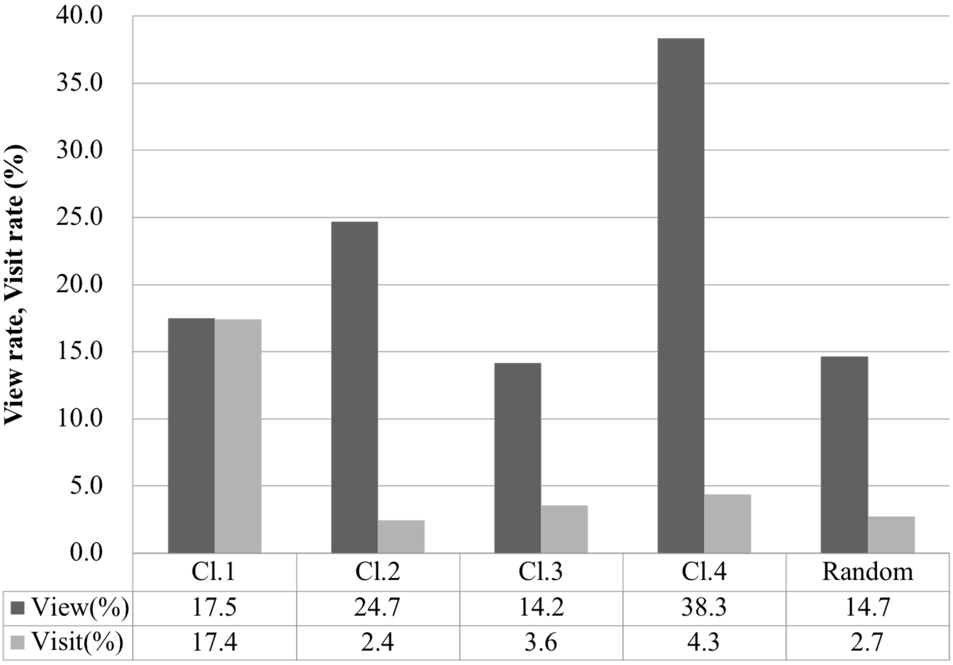
Figure 4.6 shows the verification results of customers who had purchase history and check-in history. Comparing the view rate and visit rate of each cluster can confirm the degree expected for the proposed method. Compared to the visit rate result yielded by a random distribution, Cl.1 attained 16.9 pt and Cl.4 attained 3.7 pt higher visit rates. However, the visit rates of Cl.2 and Cl.3 were lower than random.

Figure 4.7 shows the verification results of customers who had only check-in history. From the results, the view rates of Cl.1, Cl.2 and Cl.4 are higher than random. The visit rates of Cl.1, Cl.3 and Cl.4 are higher than random. The clusters that the view rate and the visit rate are improved by the visitor promotion have been clarified.



	Cl.1	Cl.2	Cl.3	Cl.4	Random
■ View(%)	47.5	39.3	33.1	33.3	37.3
■ Visit(%)	63.2	27.3	39.2	50.0	46.3

Figure 4.6: View and visit rate for purchase and check-in history holders



	Cl.1	Cl.2	Cl.3	Cl.4	Random
■ View(%)	17.5	24.7	14.2	38.3	14.7
■ Visit(%)	17.4	2.4	3.6	4.3	2.7

Figure 4.7: View and visit rate for only check-in history holders

4.6 Chapter summary

This chapter proposed a method to estimate customer preference from check-in history. It introduced the idea of customer preference clusters and showed how they could be identified in purchase history data.

The accuracy of estimating customer preference clusters from check-in history was clarified. The average of F-measure was 44.9% when using RF. For the case of randomly cluster assignment (4 clusters), the estimation accuracy rate is considered to be 25%. It seems that using check-in history to estimate the customer preference cluster is more effective than random assignment. Based on the results, Cl.4 (Japanese sweets) was the easiest of the 4 clusters to estimate. The second best was Cl.1 (Western style dishes). Some clusters were confirmed to be easier to estimate than others.

Although the results of this experiment are interesting, it is considered necessary to clarify the usefulness of this method in depth. Specifically, it was assumed that the promotional materials delivered to the customers exhibited no difference in terms of effectiveness, although this is questionable. Thus, further experiments on controlled promotions and observing the results in order to clarify contribution factor on various contents are needed.

Chapter 5

Family structure estimation

5.1 Background

The amount and type of consumer products we consume daily vary based on user attributes such as demographics, lifestyle, and family structure. User attributes play an important role for retailers in deriving marketing strategies, and provide personalized recommendations. Although user attributes are obtained from questionnaires in most cases, it is difficult to obtain personal information from all customers. Due to privacy reasons, most customers are reluctant or refuse to provide individual information.

Some recent studies showed that user attributes such as demographics [Wang et al., 2016] [Siyu et al., 2015] and lifestyle [Ishigaki et al., 2010a] can be predicted from purchase data. Due to the spread of point of sale (POS) systems, it has become possible to obtain individual purchase histories from POS data with user ID tags. Thus, without the need to ask each customer, retailers are now able to obtain user attributes according to the actual results of purchased products.

Previous research investigated the effectiveness of using purchase data to predict five demographic attributes, i.e., gender, age, marital status, income, and education level [Wang et al., 2016]. Wang et al. assumed that family-structure attributes were correlated to purchasing behavior [Wang et al., 2016]. For example, the quantity of products to be purchased differs depending on the number of family members. Also, the product to be selected differs depending on the presence and age of children. With a deeper understanding of the influence of the family structure on purchase behavior, we can recommend more suitable products to a consumer. To the best of the author's knowledge, estimating the family structure from purchase data has yet to be attempted. Herein, this thesis makes the first attempt to reveal the effectiveness of utilizing individual purchase histories in estimating family structure.

To build the estimation module of the product recommendation system, this chapter proposes a method that estimates family structure attributes, especially in terms of the number of family members, family structure, and presence of children expressed as an educational level, which is used in the estimation module [Doi et al., 2017b]. A machine learning process is employed to construct the models used in the proposed method.

The chapter is organized as follows. Section 5.2 describes product recommendation system. Section 5.3 presents the dataset employed to evaluate the proposed method. Section 5.4 introduces the proposed method. It explains how to estimate the family-structure attributes from purchase data. Section 5.5 shows evaluation results for the proposed method. Section 5.6 concludes the chapter with a summary and a view on future work.

5.2 Product recommendation system

Figure 5.1 shows an overview of the product recommendation system. The system is divided into three parts: the estimation module, recommender module, and user interface module. The Estimation module outputs customer's family structure, lifestyle and demographic attribute by estimating from the purchasing history. The Recommender module determines the products to be recommended from the Product data DB based on the customer's information estimated by the Estimation module. The User interface module presents the products determined by the Recommender module to the customer's devices.

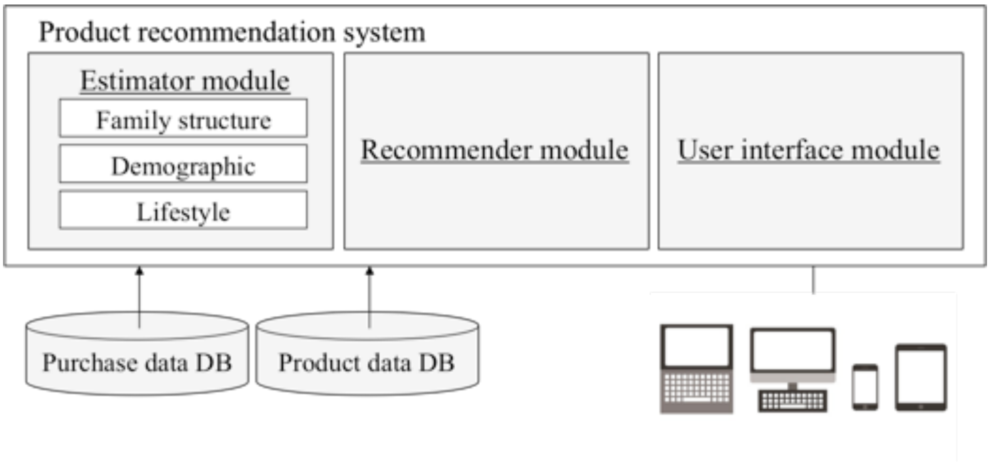


Figure 5.1: Overview of product recomendation system

5.3 Dataset

This section introduces the dataset of purchase data and questionnaire data used to construct an estimation model.

5.3.1 Purchase data

This study adopted the Intage Single Source Panel (i-SSP) dataset provided by INTAGE Inc., Tokyo, Japan. This dataset comprises the purchase history of daily necessities for a one-year period from 2014 to 2015. It contains 4,566,098 transactions belonging to 6,358 users. A transaction comprises a Consumer ID, Date, Time, Shop ID, Shop group, Product type, Product ID, Quantity, and Price. Table 5.1 gives attributes and examples of values. There are 518 kinds of shops included in the Shop IDs. Shops are grouped into 24 shop groups shown in Table 5.2. Products are grouped into 295 product types.

Table 5.1: List of purchase data

Attributes	Value Example
Consumer ID	0001
Date	2016/09/21
Time	15:00:00
Shop ID	001
Shop group	"Supermarket", "Convenience store"
Product type	"Milk", "Snacks"
Product ID	"4971198501140"
Quantity	"1", "2"
Price(Yen)	"100", "300"

Table 5.2: List of shop groups

Shop groups
Supermarkets
Convenience stores
Home improvement stores
Drug stores
Discount liquor stores
Department stores
Vending machines
Delivery services
Door-to-door sales
Electronics retail stores
100 yen shops (kind of dollar shop)
Coffee shops
Grocery stores
Specialty shops (butcher, fishmonger, green grocer)
Bakeries
Pet shops
General products stores
Baby products stores
Shops in train stations
Liquor shops
Consumer co-operative stores in universities
Cosmetic stores
Beauty salons
Animal hospitals

5.3.2 Questionnaire data

The family-structure attributes are acquired from questionnaires administered to 6,358 users identified in the i-SSP dataset. The family structure includes the number of family members, family structure, and presence of children expressed in terms of educational level. The values for each attribute are listed in Table 5.3. In the family structure, “Two generations” indicates that two generations are living in the same house. For example, parent and children where parents represent one generation and children represent the other generation, “Three generations or more” similarly indicates that three or more generations are living in the same house. For example, grandparents, parents and children. Parents who have multiple children are included in the dataset. In other words, not every parent has only one child.

Table 5.3: List of family structure attributes

Attributes	Values
Number of members	1,2,3,4 or more
Structure	Single person, Couple, Two generations, Three generations or more
Presence of infant(age 0-2)	YES/NO
Preschooler(age 3-5)	YES/NO
Elementary school student	YES/NO
Junior high school student	YES/NO

5.4 Family structure attribute estimation

This section presents the proposed method that estimates the family structure attributes from purchase data. Figure 5.2 shows an overview of the proposed method. The proposed method is divided into two processes: the training process and estimation process.

In the training process, estimation model that estimate the family structure attribute are constructed using machine-learning algorithm.

In the estimation process, family structure attributes are estimated based on the estimation model acquired in the training process.

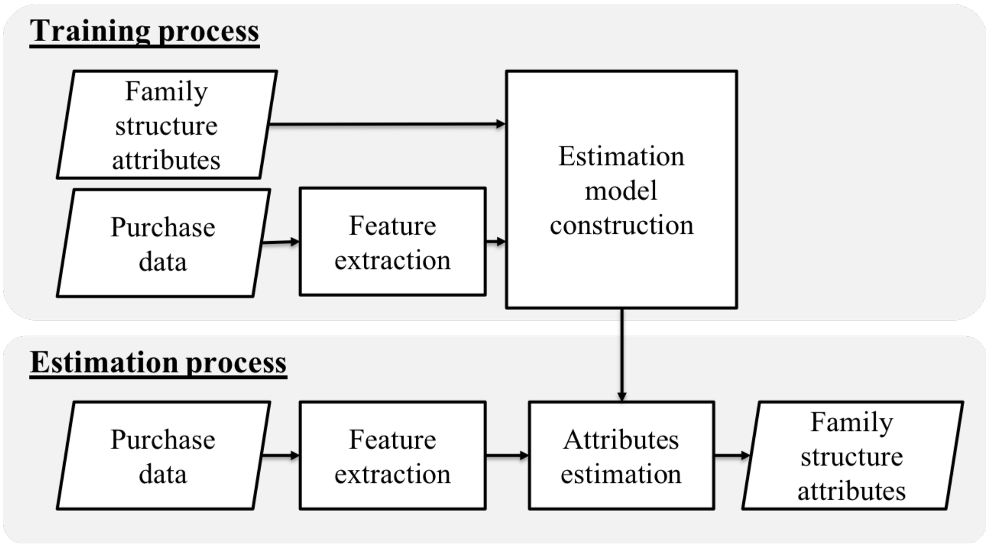


Figure 5.2: Overview of proposed method.

5.4.1 Feature extraction

For constructing the estimation model, nine features shown in Table 5.1 are extracted from the purchase data. Table 5.4 shows the extracted features and number of dimensions. Purchasing behaviors for a consumer are aggregated into a single transaction. This thesis defines these transactions as input data.

The items “Day” and “Day of the week” indicate when and how many times the consumer went shopping. The item “Time” however, does not represent the time the shop is visited. It represents when the user entered the purchase data.

Table 5.4: List of extracted features

Features	Number of Dimensions
Consumer ID	1
Day	31
Time	24
Day of week	7
Shop group	24
Total quantity for each product type	295
Total amount for each product type(Yen)	295
Total quantity of product	1
Total amount of payment(Yen)	1

5.4.2 Estimation model construction

Estimation models are constructed for each family structure attribute. Family structure attributes are shown in Section 5.3.2 and input data are shown in Section 5.4.1 . These are used to construct the estimation model. The family structure attributes and the purchase data are matched via the Consumer ID. The input dataset used for learning process is sampled randomly from the original dataset in order to set each value in attribute as same number. This thesis used a machine-learning algorithm to construct models. Estimating the family structure attributes can be classified as a classification problem because a consumer has one value for each family structure. To solve this classification problem, this thesis adopts Random Forest [Breiman, 2001] and Naive Bayes [John and Langley, 1995] for model construction.

The criteria proposed by Breiman [Breiman, 2001] are used to set parameters for Random Forest. For Naive Bayes, we define the classes so that they are distributed according to a normal distribution. After comparing the estimation ability, the constructed model that provides the highest estimation accuracy and F-measure is selected as the final estimation model. In attribute estimation, the family structure attributes are estimated by using the estimation model.

5.5 Evaluation

This section shows the evaluation results of the proposed method. The evaluation index is described in Subsection 5.5.1. Subsection 5.5.2 shows the estimation performance of each model.

5.5.1 Evaluation index

In order to evaluate the estimation accuracy of the family-structure attributes, this thesis uses 10-fold cross validation. The 10-fold cross validation works as follows: Nine subsamples represent learning data are used for constructing the estimation model and the remaining subsample represents the test data used for evaluating the estimation model. This thesis employs the following evaluation index. This thesis adopts F-measure and Accuracy. F-measure is the harmonic mean of precision and recall. i is the value in the attribute. For example, "Number of members" is attribute and "1" is value. All i values mean "1", "2", "3", "4" and "5 or more" for attribute "Number of member". $U(i)$ is the number of correct estimations for attribute value i. $V(i)$ is the number of target value estimations for attribute value i. $W(i)$ be the total number of target values. Precision $P(i)$ and Recall $R(i)$ are computed as

$$P(i) = \frac{U(i)}{V(i)} \quad , \tag{5.1}$$

and

$$R(i) = \frac{U(i)}{W(i)} \quad , \tag{5.2}$$

respectively. F-measure $F(i)$ equals

$$F(i) = \frac{2 \cdot P(i) \cdot R(i)}{P(i) + R(i)} \quad . \tag{5.3}$$

Let O be the number of all consumers in the test data. O is computed as

$$O = \sum_i W(i) \quad . \tag{5.4}$$

N be the total number of consumers who predicted the value correctly. N is computed as

$$N = \sum_i U(i) \quad . \tag{5.5}$$

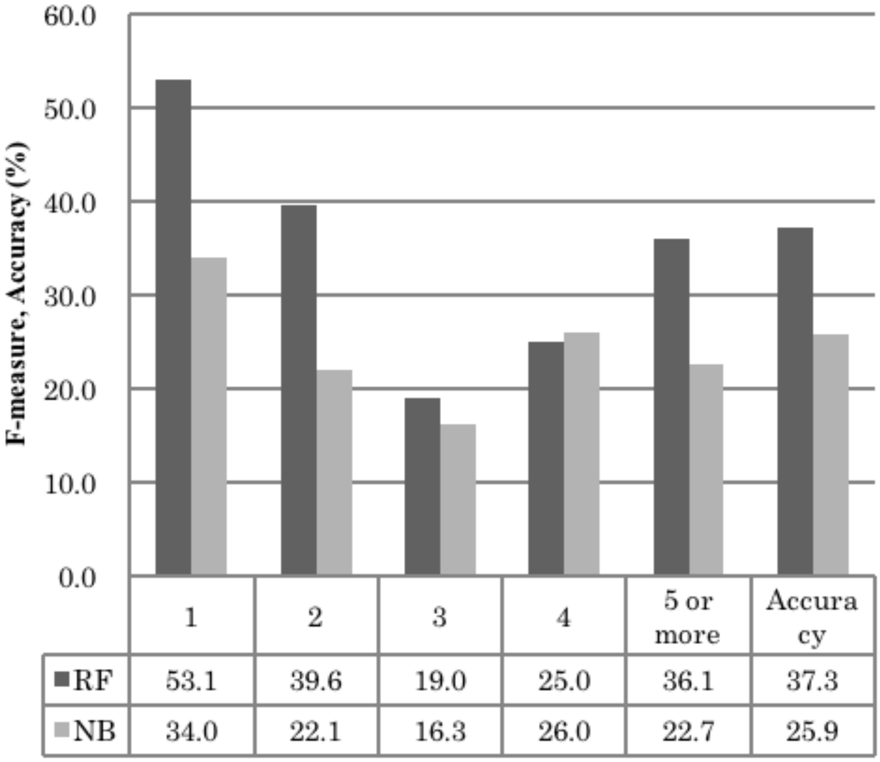
Accuracy A is computed as

$$A = \frac{N}{O} \quad . \tag{5.6}$$

5.5.2 Results

This subsection describes the accuracy of each estimation model. RF represents Random Forest and NB represent Naive Bayes. A comparison of the accuracy of RF to that of NB shows that RF estimates the family structure with high accuracy.

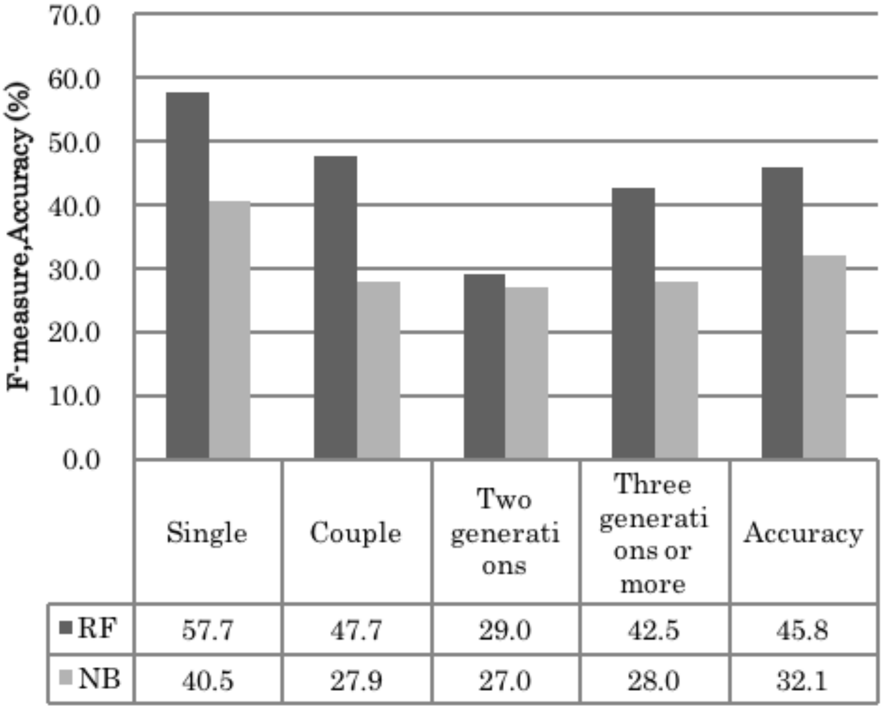
Figure 5.3 shows the estimation accuracy and F-measure for the number of family members. Based on the results, “1” which means the user lives alone is the easiest to estimate for all items among the considered number of family members. It is found that the total quantity and total amount of payment for daily necessities are effective features for determining the number of family members. In particular, the total quantity and total payment value for “Laundry detergent”, “Rice” (a staple food for Japanese) and “Toilet thesis” strongly indicate the number of family members. It is easy to understand that the quantity of bare necessities of life such as “Laundry detergent”, “Rice” and “Toilet thesis” are effective features for indicating the number of family members.



	1	2	3	4	5 or more	Accuracy
■RF	53.1	39.6	19.0	25.0	36.1	37.3
■NB	34.0	22.1	16.3	26.0	22.7	25.9

Figure 5.3: Estimation accuracy of number of family members.

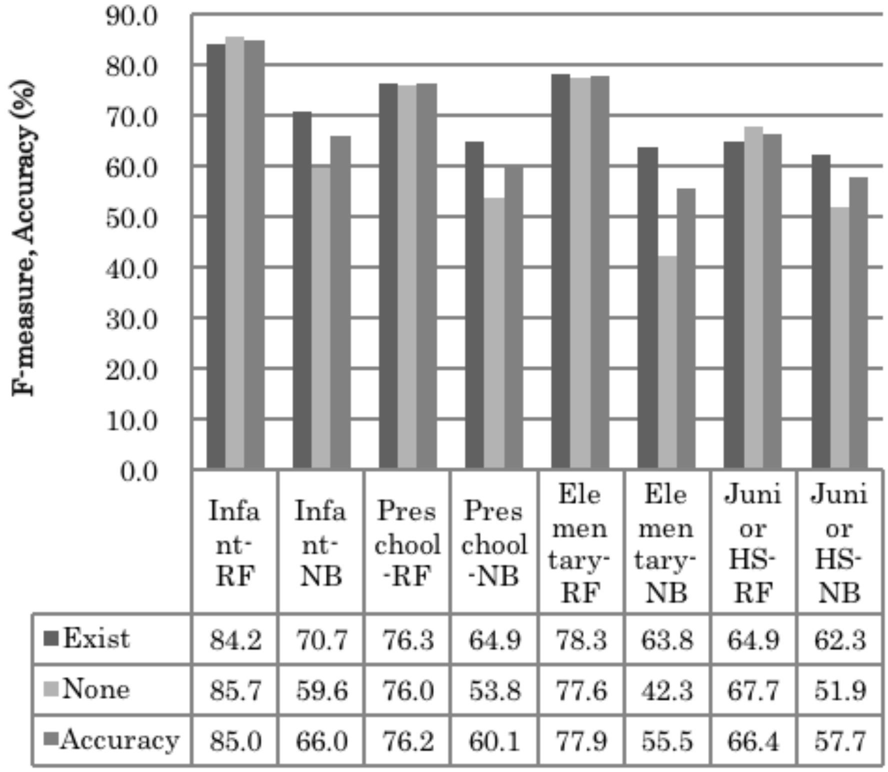
Figure 5.4 shows the estimation accuracy and F-measure for the family structure. "Single" is estimated with the highest F-measure. A similar tendency in effective features for estimation is observed between the number of family members and family structure.



	Single	Couple	Two generations	Three generations or more	Accuracy
■RF	57.7	47.7	29.0	42.5	45.8
■NB	40.5	27.9	27.0	28.0	32.1

Figure 5.4: Estimation accuracy of family structure.

Figure 5.5 shows the estimation accuracy and F-measure for the presence of children expressed in terms of educational level.



	Infant-RF	Infant-NB	Preschool-RF	Preschool-NB	Elementary-RF	Elementary-NB	Junior HS-RF	Junior HS-NB
■Exist	84.2	70.7	76.3	64.9	78.3	63.8	64.9	62.3
■None	85.7	59.6	76.0	53.8	77.6	42.3	67.7	51.9
■Accuracy	85.0	66.0	76.2	60.1	77.9	55.5	66.4	57.7

Figure 5.5: Estimation accuracy for presence of children.

Table 5.5 to Table 5.8 give the information gain. Information gain is an index used in feature selection of random forest. “Num” represents the total quantity and “Pri” represents the total payment value. “Shop” represents the number of times the target shop group was visited. “Time” represents the time that the user entered the purchase data. The features are arranged in decreasing order of information gain.

Table 5.5: Information gain of infant

Feature	Information Gain
Num Disposable diapers	0.37
Pri Disposable diapers	0.37
Pri Baby food	0.19
Num Baby food	0.19
Shop Baby products store	0.14
Pri Wet tissues	0.07
Pri Powdered milk	0.05
Num Powdered milk	0.05
Num Wet tissues	0.05
Pri Other confectionary	0.03
Pri Premix powder	0.03
Pri 100% juice	0.03
Num Side dishes	0.03
Pri Side dishes	0.03
Pri Cooked rice	0.03
Pri Confectionary from a toy manufacturer	0.03
Num Confectionary from a toy manufacturer	0.03

Table 5.6: Information gain of preschool

Features	Information Gain
Pri Confectionary from a toy manufacturer	0.13
Num Confectionary from a toy manufacturer	0.13
Num Disposable diapers	0.09
Pri Disposable diapers	0.09
Num Other confectionary	0.06
Pri Other confectionary	0.06
Num Premix powder	0.04
Pri Premix powder	0.04
Num Candy	0.03
Shop Baby products store	0.03
Num Extract	0.03
Time 20	0.03
Num Cooked rice	0.03
Pri Lactic drink	0.03
Num Lactic drink	0.03
Pri Whipped cream	0.02
Num Whipped cream	0.02

Table 5.7: Information gain of elementary

Features	Information Gain
Pri Confectionary from a toy manufacturer	0.08
Num Confectionary from a toy manufacturer	0.08
Num Premix powder	0.04
Num Candy	0.04
Num Snacks	0.04
Pri Premix powder	0.04
Pri Other confectionary	0.04
Pri Snacks	0.03
Num Other confectionary	0.03
Num Meat sausage	0.03
Pri Meat sausage	0.03
Num Whipped cream	0.03
Pri Candy	0.03
Num Barely tea	0.03
Pri Seaweed laver	0.03
Time 12	0.03
Pri Detergent for laundry	0.03

Table 5.8: Information gain of junior high school

Features	Information Gain
Num Furikake	0.04
Num Shampoo	0.04
Num Sanitary items	0.03
Pri Furikake	0.03
Pri Sanitary items	0.03
Pri Sports drinks	0.03
Num Tomato ketchup	0.03
Num Meat sausage	0.03
Shop Supermarket	0.03
Pri Premix powder	0.03
Pri Seaweed laver	0.03
Pri Shampoo	0.03
Pri Sauce for grilled meat	0.03
Num Frozen meals and dinners	0.03
Pri Meat sausage	0.03
Num Premix powder	0.03
Num Sauce for grilled meat	0.03

“Infant” is estimated with the highest accuracy of 85.0 %. There are several features that increase the accuracy. The total quantity and total payment value for “Disposable diapers”, “Baby food”, “Wet tissues” and “Powdered milk” are effective features. The number of times the “Baby products store” was visited is also a feature unique to “Infant”. It is natural that families with infants purchase items that only infants use. This is why the estimation accuracy of the “Infant” is very high.

The accuracy of “Elementary” is 77.9%, which is the second best. The total quantity and total payment value for confectionaries are effective features. Confectionaries include “Confectionary from a toy manufacturer”, “Candy”, “Snacks” and “Other confectionary” in this thesis. The effective features in “Preschool” appear similar to the features of “Infant” and “Elementary”. The data includes parents who have multiple children. There is a possibility that this can impact the estimation accuracy of “Preschool”.

From the results, this thesis clarified the accuracy with which the family structure can be estimated from purchase data. It seems that using purchase data to estimate the family structure is effective. The next step is to clarify how estimated family structure works in real retail scenarios.

5.6 Chapter summary

This chapter proposed a method that estimates family-structure attributes by focusing on user purchasing behavior. The method derives a relevance model for each product type between family structure attributes and purchase histories beforehand based on a consumer panel survey. Evaluations based on real datasets showed the effectiveness of the proposed method. The datasets contained 4,566,098 transactions from 6,358 users. The estimation accuracy of Presence of “Infant” is 85.0% and Presence of “Elementary” is 77.9%. The method is useful in deriving smart recommendations such as suggesting to the consumer products that suit the family structure.

As future work includes evaluating the estimation accuracy of multi-class estimation. Plans include extracting and using effective Product IDs to improve the accuracy of estimating the family-structure attributes. This model will be integrated into the proposed product recommendation system, as its performance will be enhanced by knowing the consumer’s attributes.

Chapter 6

Purchasing intent estimation

6.1 Background

Most companies use a customer relationships management system to improve their relationships with customers. Such systems record and manage customer attributes such as demography, preference, purchase history, usage history and contact history. Analyzing these data makes it is possible to respond to each customer's most effectively.

By analyzing own data and other company's data, it becomes possible to discover the customer's purchasing intent which would not be possible from just own data. Credit card company can obtain their cardholder attribute data and purchase history, which consists of just the store, date of purchase, and amount. The history does not include purchase description so exactly what the user is interested in is unknown. Gathering and analyzing "Like " and "Check-in " data from Facebook and Foursquare, American Express could realize the effective distribution of vouchers based on customer preferences, interests, and actions. The point card Ponta [Ponta, 2018] and T-point [T-Point, 2018] are visualizing latent customers from the point card usage histories of multiple stores.

The desirable customer is one that shops frequently and should be identified for improving sales and the efficiency of the various measures used for targeting customers. This thesis defines the level of customer desirability as customer level in this thesis. Specially, customer level can be used for customer selection such as optimization of the approach taken to customers, campaign to train best customer candidates and measures to activate dormant customers. By integrating customer level into Customer relationships management systems, this thesis can build better relationships with customers.

Several location-based services allow the consumer to share physical location, the “Check-in” function, by using mobile applications on smart phones. The explicit behavior of check-in indicates the intention of the customer, and it is considered that using this data is effective for customer understanding. Facebook places [Facebook places, 2018], Foursquare [Foursquare, 2018] and Yelp [Yelp, 2018] are examples of location-based services. Shopkick [Shopkick, 2018] is also one of the services that use the “Check-in” action. Customers are able to check into locations and redeem gift cards and vouchers as rewards. Shopkick is used by more than 15 million users. Online-to-offline services using “Check-in” are expected to expand in the future. Furthermore, it is also expected that the number of users accumulating large check-in histories will increase. The check-in data that can be acquired by these services includes where the purchase was made.

In order to calculate the customer level, RFM is often used. RFM is one of a method to analyze customer level from purchase history data. In this method, Freshness (last purchase date), Frequency (purchase frequency), Monetary (purchase price) are used as indicators from the purchase history. It can be used to indicate customer’s level.

By utilizing this check-in history information, if it is possible to estimate the customer level of those who do not hold purchase information, we can better maintain existing customers and acquire new desirable customers. To the best of the author's knowledge, no evaluation using actual service data has been conducted for estimating the degree of customer desirability from the check-in history information gathered inside and outside the target shop. Therefore, this thesis set the hypothesis that the check-in history information correlates with the purchase situation of a specific store and demonstrates the validity of the hypothesis.

This thesis proposes a method to estimate customer level from customers who only have check-in history data [Doi et al., 2016] [Doi et al., 2017a] . Specifically, customer level of the target customer in the target shop is estimated from combined store check-in history data. The purchase history information of the target shop and the store check-in history information are used as the data. An estimation model is constructed using the machine learning approach. An evaluation is performed using the accuracy of the estimation model and the results of a visitor promotion. This thesis shows the effectiveness and practicality of the customer level estimation method by using the visit rate to target store as discovered during the visitor promotion. This thesis believes that these methods are applicable to Online to Offline services that acquire store check-in history data.

The chapter is organized as follows. Section 6.2 introduces the dataset used to construct the estimation model and evaluate the proposed method. Section 6.3 defines customer level and introduces the customer level estimation method. Section 6.4 explains the evaluation conducted and shows the results. Section 6.5 discusses the study. Section 6.6 concludes the chapter with a summary and view on future work.

6.2 Dataset

Datasets of check-in history and purchase history shown in Section 4.2 were used to construct a model to estimate customer level. This section details the datasets used.

6.2.1 Check-in history data

The Shoplat dataset was adopted as check-in history shown in Section 4.2. All check-in spots are in the shop, and the Shop ID and Check-in Spot ID are given. There were 96 check-in spots. A total of 2,345,976 check-in histories were used.

6.2.2 Purchase history data

The credit card payment history of TOKYU CARD, INC. was adopted as purchase history data [Tokyu point Tokyu card, 2018] shown in Section 4.2. This data is limited to purchases made in the Tokyu Department Store, Shibuya Shinqs Hikarie [Shibuya Hikarie ShinQs, 2018]. This dataset holds records for the 23 month period from 2013 to 2015. It contains 36,522 transactions.

6.3 Customer level estimation method

This section describes customer level, customer estimation method, and estimation accuracy achieved.

6.3.1 Customer level

Customer level was calculated by using RFM. In this method, Freshness which means last purchase date), Frequency which means purchase frequency and Monetary which means purchase price are used as indicators from the purchase history. The duration of the data used is one year. Freshness is calculated using the day of purchase.

The customer is divided into equal parts for each indicator by using the purchase history information shown in Section 6.2.2. A score of 1 to 3 was given in descending order to each indicator. To set the customer cluster, the total value of the scores in each indicator were calculated and set Cl.1 to Cl.7 in decreasing order of total

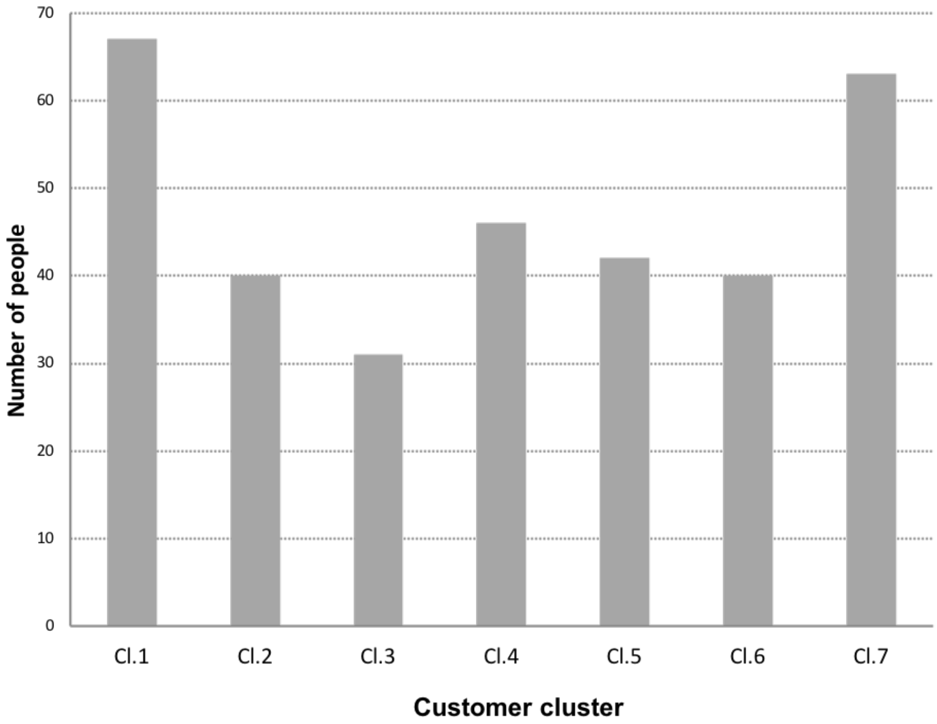


Figure 6.1: Distribution of customer cluster

value. For example, Cl.7 which has the highest customer cluster, each indicator is 3 and the total score is 9. Conversely, the total score of Cl.1 with the lowest customer cluster is 3. Figure 6.1 shows the number of distributions for each customer cluster.

The customers belonging to Cl.1 and Cl.2 were taken to be "Non-Active Customer", Cl.6 and Cl.7 as "Best Customer". These groups are used as customer level in this thesis. Customer level l_c of customer c is represented by $l_c \in \{$ Non-Active Customer, Best Customer$\}$.

This chapter aims to attract the "Best Customer" to the target shops. The assumption made is that the estimation accuracy of each customer level can be improved by using customer data in which features of each customer level appear. Therefore, learning data to be used for estimation of good customer level was selected considering two requirements. The first is the magnitude of the difference between the respective values of the customer's Freshness, Frequency, and Monetary. The second is ensuring that each customer level has many people. It is desirable that the number of people belonging to each customer level is the same. Three patterns were chosen in consideration of the number of people. In the first case (1), "Non-Active Customer" is Cl.1 and "Best Customer" is Cl.7. In the second case (2), "Non-Active Customer" is Cl.1 and Cl.2 and "Best Customer" is Cl.6 and Cl.7. In the third case (3), "Non-Active Customer" is Cl.1, Cl.2, Cl.3 and "Best Customer" is Cl.5, Cl.6 and Cl.7.

The distance Dis between customer levels is calculated using equation 6.1. Dis is expressed in the three dimensions of Freshness, Frequency and Monetary. These ranges have different dimensions. Therefore, normalization was performed with reference to the difference of each dimension from case (1). For the all customer level clusters, cl, the average value of Freshness is $Ravg_{cl}$, the average value of Frequency is $Favg_{cl}$, and the average value of Monetary is $Mavg_{cl}$.

When "Non-Active Customer" is Cl.1 and "Best Customer" is Cl.7 (Case(1)), this thesis defines the distance of Freshness as $RBAbs$ (equation 6.5), the distance of Frequency as $FBAbs$ (equation 6.6), the distance of Monetary as $MBAbs$ (equation 6.7). num is the number of people belonging to each target cluster.

Dis indicates that the characteristics of "Best Customer" and "Non-Active Customer" become more different as the value increases. Dis in Case (1) was 130.0, Case (2) was 143.7, Case (3) was 143.1. This result suggests that Case (2) is the best combination. Therefore, this chapter assumes Case (2) hereafter.

$$Dis = \frac{(RAbs + FAbs + MAbs)}{3} \cdot \sum_{cl} num_{cl} \tag{6.1}$$

$$RAbs = \frac{|Ravg_{cl(BestCustomer)} - Ravg_{cl(Non-ActiveCustomer)}|}{RBAbs} \tag{6.2}$$

$$FAbs = \frac{|Favg_{cl(BestCustomer)} - Favg_{cl(Non-ActiveCustomer)}|}{FBAbs} \tag{6.3}$$

$$MAbs = \frac{|Mavg_{cl(BestCustomer)} - Mavg_{cl(Non-ActiveCustomer)}|}{MBAbs} \tag{6.4}$$

$$RBAbs = |Ravg_{cl.7} - Ravg_{cl.1}| \tag{6.5}$$

$$FBAbs = |Favg_{cl.7} - Favg_{cl.1}| \tag{6.6}$$

$$MBAbs = |Mavg_{cl.7} - Mavg_{cl.1}| \tag{6.7}$$

6.3.2 Customer level estimation

This section presents the proposed method; it estimates the customer level from check-in history data. It is assumed that explanatory variable i_c of customer c indicates the store check-in situation. Explanatory variable i_c uses the value obtained by normalizing the check-in number of customer c at each check-in spot using equation 6.8. The approach of estimating good customers by setting thresholding the number of check-in times is conceivable. There is a difference between the minimum value and the maximum value of the check-in frequency for each customer or shop. It is thought that the difference between these values may affect the estimation accuracy of good customers. Therefore, three kinds of estimation accuracy were calculated beforehand. First case uses the number of check-ins. Second case normalizes the number of check-ins for each customer. Third case normalizes the number of check-ins for each store. The estimation accuracy was highest for the second case and lowest for the first case. The difference between the estimation accuracy of these two cases (F-measure, adopting the Random Forest method) was 12.6%. Therefore, the number of check-ins normalized for each customer is used in this method.

h indicates a check-in spot. H indicates the total number of target check-in spots. $a_{c,h}$ indicates the number of check-ins at check-in spot h made by customer c. The normalized check-in number $n_{c,h}$ is calculated using equation 6.9. The total check-in number s_c of customer c is calculated using equation 6.10.

$$i_c = (n_{c,1}, \cdots, n_{c,H}) \tag{6.8}$$

$$n_{c,h} = \frac{a_{c,h}}{s_c} \tag{6.9}$$

$$S_c = \sum_{h=1}^{H} a_{c,h} \tag{6.10}$$

The chapter categorizes customers as "Non-Active Customer" or "Best Customer". In building an estimation model, good customer level l_c mentioned in Section 6.3.1 is used. At the time of estimation, a model is built that outputs either "Non-Active Customer" or "Best Customer".

As the machine learning technique used in estimating good customer level the approaches of Random Forest method, Logistic regression method (Logistic) and Support vector machine (SVM) are considered as candidates. Random Forest method [Breiman, 2001] is a method of estimating a target variable by using multiple decision trees created by random sampling of explanatory variables.

Criterion of Breiman [Breiman, 2001] are used as the number of explanatory variables to be sampled and the depth of the decision tree. The logistic regression method [David, 1958] is a method used for binary discrimination and prediction of occurrence probability. SVM [Boser et al., 1992] is a method that performs binary discrimination; its kernel function uses the RBF kernel.

Evaluation is performed by 10-fold cross validation, learning 90% of data and with the remaining 10% used for evaluation. For cross validation, the data of "Best Customer" (Cl.6, Cl.7) or the "Non-Active Customer" (Cl.1, Cl.2) was targeted at good customer level l_c. Customer cluster data labelled Cl.3, Cl.4 and Cl.5 were not used in learning.

Next, the estimated performance of the proposed method will be described. Precision rate, Recall rate and F-measure are used as the evaluation metrics.

In cross validation, data of Cl.3, Cl.4 and Cl.5 are not used. However, customers actually belonging to these clusters actually exist. Consumers belonging to Cl.3, Cl.4 and Cl.5 were classified as either "Best Customer" or "Non-Active Customer" as the result of the estimation. Therefore, this thesis evaluated the performance taking account of consumers belonging to Cl.3, Cl.4 and Cl.5. The precision rate obtained by cross validation is represented as p. The precision rate, p', is a precision rate corrected by considering the consumers labelled Cl.3, Cl.4 and Cl.5. Cl.3, Cl.4, Cl.5 consumers do not affect the recall rate r. The precision rate, p_l, of customer level l is calculated using equation 6.11. The matching rate p'_l is calculated using the equation 6.12.

t_l indicates the number of customers estimated to be customer level l. v_l indicates the number of customers who actually are customer level l among the estimated level l customers.

u_l represents the number of customers belonging to the customer level l. W indicates the number of customers belonging to Cl.1, Cl.2, Cl.6 and Cl.7 used for cross validation. O is the total number of customers belonging to customer clusters Cl.3, Cl.4 and Cl.5 mentioned in Section 6.3.1.

Recall ratio r_l is calculated using equation 6.14. F-measure f_l by F-measure 1 is calculated from the harmonic mean of the precision rate p_l and the recall rate r_l using equation 6.15. F-measure f_l' by F-measure 2 is calculated from the precision rate p_l' and the recall rate r_l using equation 6.16.

$$p_l = \frac{v_l}{t_l} \tag{6.11}$$

$$p_l' = \frac{v_l}{t_l + O \cdot \frac{t_l}{W}} \tag{6.12}$$

$$W = \sum_{l \in (BestCustomer, Non-ActiveCustomer)} u_l \tag{6.13}$$

$$r_l = \frac{v_l}{u_l} \tag{6.14}$$

$$f_l = \frac{p_l \cdot r_l \cdot 2}{p_l + r_l} \tag{6.15}$$

$$f_l' = \frac{p_l' \cdot r_l \cdot 2}{p_l' + r_l} \tag{6.16}$$

Figure 6.2 shows the result of “Best Customer” estimation. Precision ratio p is indicated by Precision 1, precision rate p' by Precision 2, and recall rate r by recall. In the case of using the Random Forest method, F-measure 1 f_l was 78.9%. F-measure 2 f_l' which indicates the estimation accuracy when considering the distribution of all customers, was 60.8%.

Figure 6.3 is the result of estimating “Non-Active Customer”. F-measure 1 f_l was 77.6% for the Random Forest method and 75.5% for SVM. F-measure 2 f'_l was 60.8% in the case of using the Random Forest method and 56.6% when using SVM. These results confirm that “Best Customer” can be estimated only from the check-in history although the machine learning method used has some impact.

The purpose of this research is to select customers with high customer desirability such as “Best Customer” and direct them to the target shop. For this reason, this thesis adopts the Random Forest method as its F-measure 2 f'_1 was the highest for “Best Customer” estimation. The target customers of visit promotion explained in Chapter 6.4 were selected using this model.

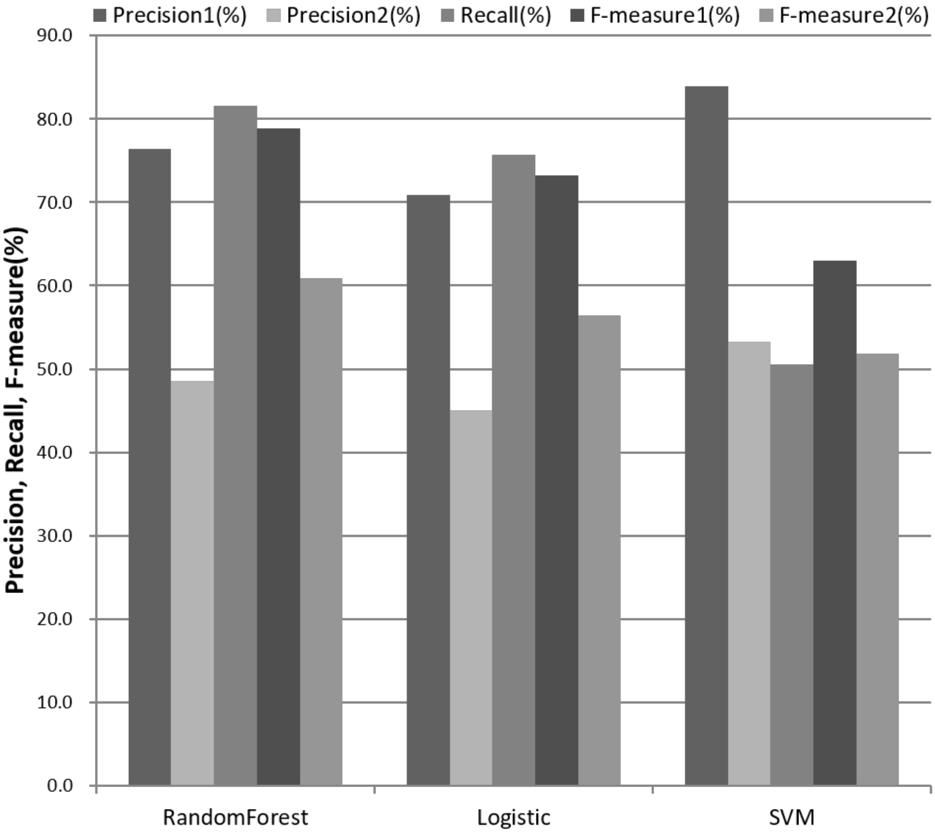


Figure 6.2: Accuracy rate of each method on Best customers

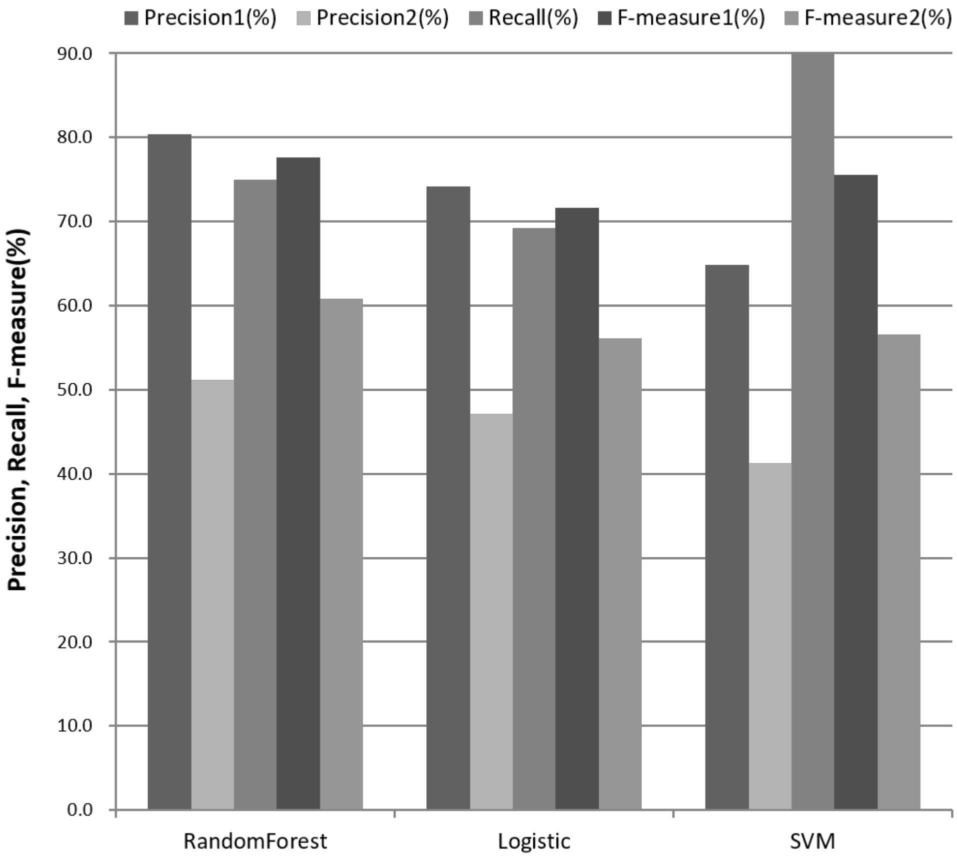


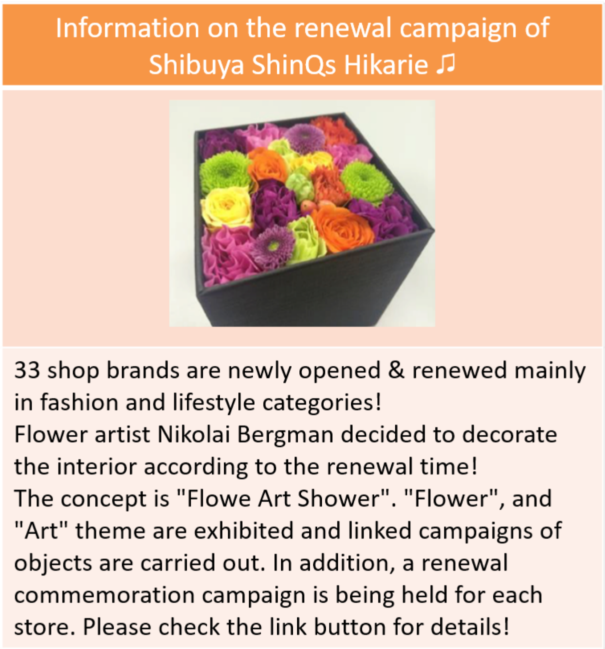
Figure 6.3: Accuracy rate of each method on Non-Active customers

6.4 Evaluation for visitor promotion

In order to confirm the usefulness of the proposed method introduced in Section 6.3.1, this thesis implemented visitor promotion for “Best Customer” and “Non-Active Customer” estimated using the proposed method. This section details the visitor promotion conducted and the results gained.

6.4.1 Visitor promotion

The target customers for visitor promotion were selected by using the proposed method. The Shoplat application was used to distribute information, see Figure 6.4 for example. The contents to be delivered consisted of information on the opening of a new tenant in Shibuya ShinQs Hikarie. The evaluation involved observing the effect of information distribution on visit rate.



Information on the renewal campaign of Shibuya ShinQs Hikarie ♫

33 shop brands are newly opened & renewed mainly in fashion and lifestyle categories!
Flower artist Nikolai Bergman decided to decorate the interior according to the renewal time!
The concept is "Flowe Art Shower". "Flower", and "Art" theme are exhibited and linked campaigns of objects are carried out. In addition, a renewal commemoration campaign is being held for each store. Please check the link button for details!

Figure 6.4: Delivered content for Shibuya ShinQs Hikarie trial

The target customers were customer who used the Shoplat application and had generated store check-in information but not purchase history information. This thesis selected “Best Customer” and “Non-Active Customer” from among these customers by applying the customer level estimation method.

The target check-in sites consisted of 96 spots where a customer holding both purchase history information and store check-in history information checked in at least once. Customers who did not check in at any of the check-in spots could not be predicted by the proposed method and are labelled as “Unforeseeable Customer”.

Furthermore, this thesis defines existing customer and new customer using the store check-in history for the past year in Shoplat service. Existing customers are defined as customers whose check-in history information contained a reference to Shibuya ShinQs Hikarie. New customers are defined as customers whose check-in history information showed no reference to Shibuya ShinQs Hikarie. By comparing existing customers with new customers, the effectiveness of the customer level estimation method was confirmed. Table 6.1 shows the number distribution of target customers according to estimated customer level. All “Unforeseeable Customer” are new customers.

Table 6.1: Distribution of target customers

Customer level	New・Existing	Number of customer
Best Customer	New	111
	Existing	516
Non-Active Customer	New	11,397
	Existing	2,602
Unforeseeable Customer	New	2,665

6.4.2 Effect of the visitor promotion

This section describes the results of the visitor promotion held at Shibuya ShinQs Hikarie. Effectivness is confirmed by using the view rate of the distribution information and the visit rate.

The visitor promotion involved 1,011 people. The effect is checked by the view rate of distributed information, visit rate and purchase rate limited to the purchase behavior of Shibuya ShinQs Hikarie. The view rate was calculated using equation 6.17. d is the number of customers to whom the information was distributed. e is the number of customers who viewed the distributed information. The visit rate was calculated using equation 6.18. g is the number of customers who visited the target shop after viewing the delivered information. The customers who visited the store without viewing the delivered information or browsed the information after visiting the target shop, are not included in the number of customers of g. The purchase rate was calculated using equation 6.19. b is the number of customers who visited the target shop after viewing the delivered information and purchased an item by credit card at the target shop.

$$View\ rate = \frac{e}{d} \tag{6.17}$$

$$Visit\ rate = \frac{g}{e} \tag{6.18}$$

$$Purchase\ rate = \frac{b}{g} \tag{6.19}$$

Customer with a purchase history can be given an accurate customer level. The view rate, the visit rate, the purchase rate are checked for the evaluation. By this observation, it is possible to confirm the effectiveness and degree of estimation for classification at the customer level. As a result of calculating a customer level using purchase history, 305 “Best Customer” and 334 “Non-Active Customer” were existed. This thesis conducts a verification to confirm that the use promotion measure result is significantly different depending on the customer level.

The chi-squared test was adopted as a verification method. The chi-squared test is carried out assuming that 1 can be observed when each customer’s viewing, visiting, purchasing can be observed, and 0 when they cannot be observed.

The effectiveness verification result of the customer with purchase history is shown in Figure 6.5. Compared to “Non-Active Customer”, “Best Customer” had superior view rate of 7 pt, visit rate of 15 pt, and purchase rate of 26 pt. Moreover, the chi-squared test indicated the significance level to be 5%, confirming that there was a statistically significant difference between the customer types. The effectiveness of RFM analysis was also confirmed.

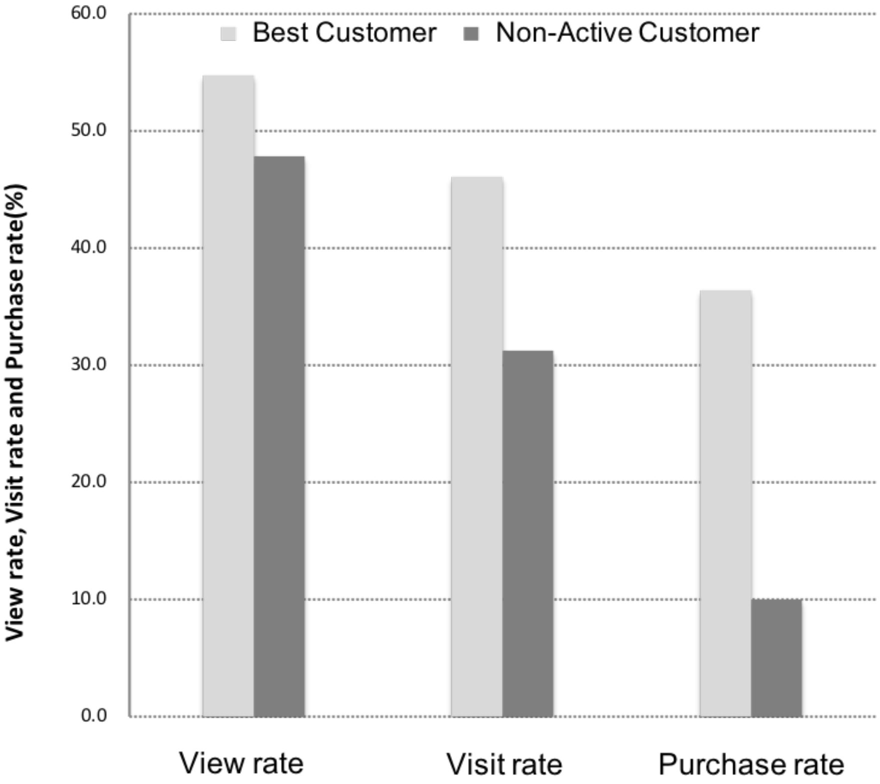


Figure 6.5: View rate, visit rate and purchase rate on customers who have purchasing history data

Next, the visit promotion results for 17,291 customers who had no purchase history information are shown in Figure 6.6 and Figure 6.7 . For customers with no purchase history information, this thesis estimates the best customer level using the customer level estimation method described in Section 6.3.1. Figure 6.6 shows the results for new customers. The view rate of “Best Customer” was 12 pt higher than that of “Non-Active Customer”. The view rate of “Best Customer” was 19 pt higher than that of “Unforeseeable Customer”. This confirms that the view rate of “Non-Active Customer” is 7pt higher than that of “Unforeseeable Customer”. It also confirms that the browsing rate decreases in the order of “Best Customer”, “Non-Active Customer”, “Unforeseeable Customer”. There was a statistically significant difference for each customer type ($p < 0.05$, Chi-squared test). In addition, the visit rate of “Best Customer” was 5 pt higher than that of “Non-Active Customer”, and 6 pt higher than “Unforeseeable Customer”.

However, no significant difference in the visit rate was confirmed for “Best Customer” and “Non-Active Customer” ($p < 0.05$, chi-squared test). The weak performance for “Best Customer” that had no purchase history, is considered to be due to the small number of samples. Figure 6.7 shows the results for existing customers. Existing “Best Customer” had a view rate 3 pt and a visit rate 18 pt higher than those of “Non-Active Customer”. Similarly, for new “Best Customer”, this thesis confirmed the tendency that the view rate and the visit rate were relatively high.

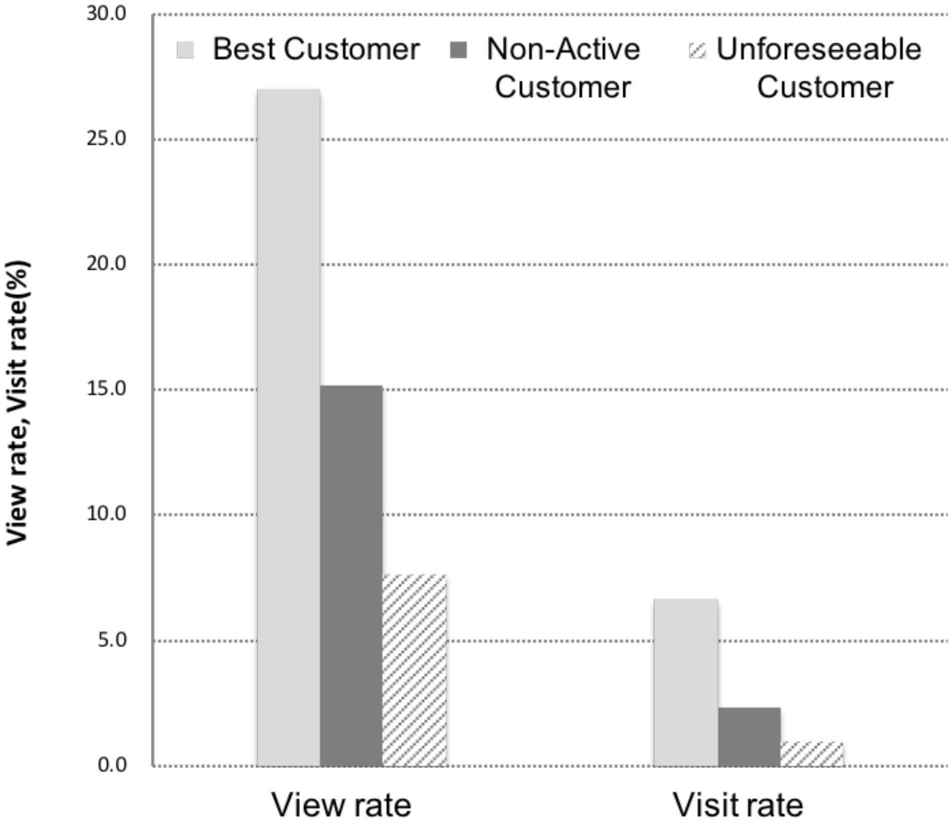


Figure 6.6: View rate and visit rate on new customers

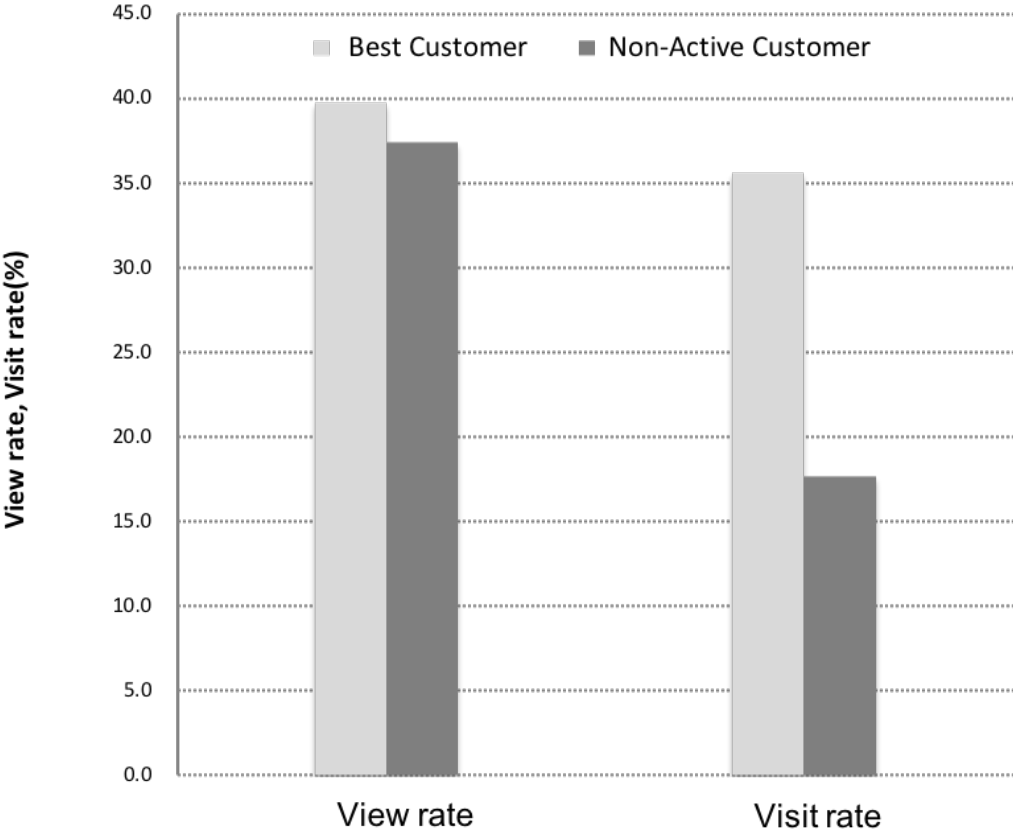


Figure 6.7: View rate and visit rate on existing customers

The effectiveness of the customer level estimation method using check-in history has been proved by the higher view rate and the visit rate of "Best Customer" compared to the other customer types. Furthermore, "Best Customer", both new and existing, was confirmed to browse the information delivered by the promotion. Compared with existing customers, new customers exhibited statistically significant differences in terms of lower view rate, 12.7 pt down, and visit rate, 28.9 pt, down ($p < 0.05$, binomial test) shown in Figure 6.8. This also indicated that new customers are more difficult to motivate than existing customers into visiting target shops. Similar results were confirmed for Non-Active customer shown in Figure 6.9. Compared with existing customers, new customers exhibited statistically significant differences in terms of lower view rate, 22.2 pt down, and visit rate, 15.4 pt, down ($p < 0.05$, binomial test) .

"Unforeseeable Customer" clearly exhibited lower view rates and lower visit rates are lower than the other customer types.

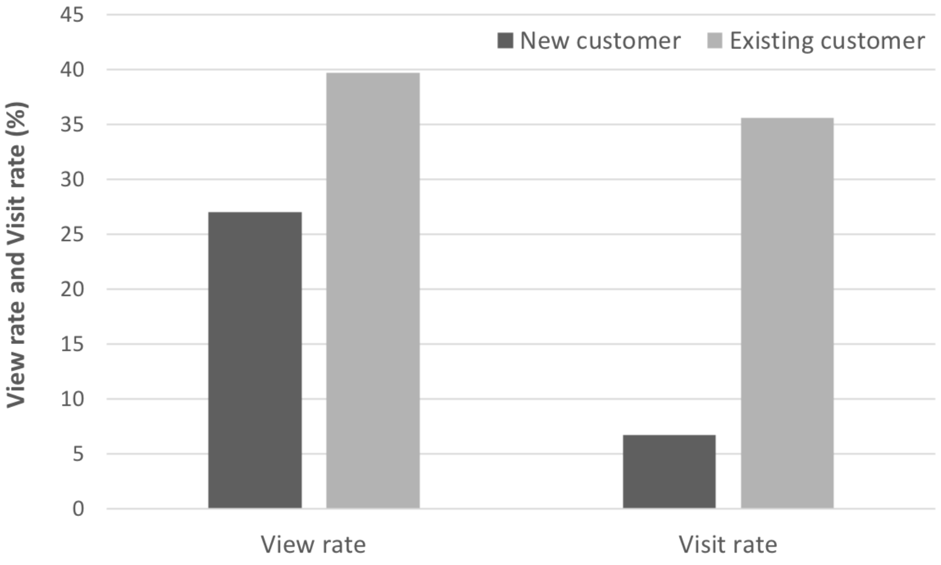


Figure 6.8: View rate and visit rate on estimated Best customers

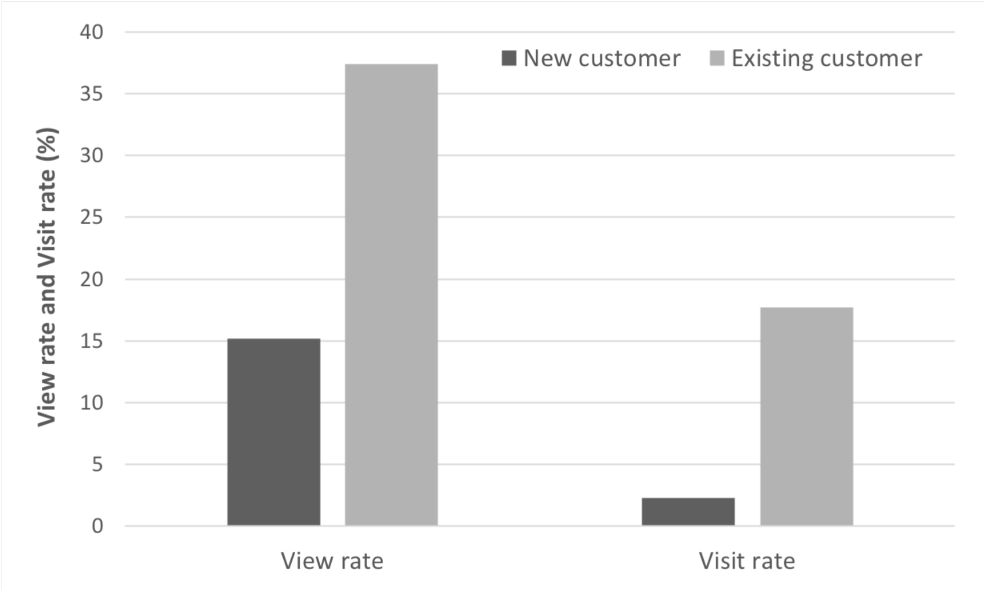


Figure 6.9: View rate and visit rate on estimated Non-Active customers

6.5 Discussion

When judging customer desirability by RFM analysis, this thesis used total purchases in the target shop without considering the tenant's business style. However, in reality, the "Best Customer" of a tenant dealing mainly with confectionery is not always the "Best Customer" of a tenant selling clothing. It is assumed that when the goods are different, the perceived degree of value will differ greatly for Freshness, Frequency and Monetary. Taking the tenant's business into account will make it possible to estimate "Best Customer" for individual tenants. A future plan includes conducting RFM analyses that consider the tenant's business when estimating the customer level. This thesis is a valuable first step as it suggest how to clarify the explanatory variables that improve the estimation accuracy.

In the proposed method, the number of check-ins was used for customer level estimation. As one approach to estimating customer level, the method of simply providing a threshold value for the total number of check-ins can be considered. For example, if the number exceeds the threshold value, the customer is tagged "Best Customer", otherwise "Non-Active Customer". However, the store check-in history is a record captured in various areas and check-in spots. This makes thresholding less effective in estimating the customer level of a target shop. As an example, it is considered that a customer who has a large number of check-ins in area A lives in or next to area A, and it may be difficult for the customer to visit a store in area B. The proposed method allows the difference in areas accessible to the customer to be considered, and so makes it is possible to estimate the customer level of the target shop more accurately. In addition, there are variables that affect estimation accuracy at the customer level. These include the time of day and the day of the week at which the check-in was created, and the spatial order of check-ins on the same day. If enough data volume can be accumulated to express the customer's behavior, we can improve the estimation accuracy and distribute attractive promotional material that suits each customer.

Furthermore, this thesis proposes a good customer estimation method with the aim of estimating “Best Customer” (customer with high score of any RFM). However, by interpreting and using each score value of Freshness, Frequency, Monetary as a three-dimensional space, RFM analysis can better identify customer desirability. For example, it is becomes possible to identify customers who were good customers in the past but have not visited for some time. This thesis also showed how to categorize customers who are lie in middle ground between “Best Customer” and “Non-Active Customer”. By refining the proposed method, it will be possible to implement measures aimed at maintaining existing customers by those who are more likely to lose interest.

This thesis performed evaluations using purchase history information generated by credit card payment, so cash payments were not considered. Future work is to identify how cash payments can be used to generate useful purchasing information.

6.6 Chapter summary

This thesis proposed a method to estimate customer level in the target store from just the store check-in history information of the customer. Specifically, features of the store check-in history were learned for each customer level by using the purchase history of some customers and the store check-in history information in advance. The evaluation results confirmed that good customers can be estimated from shop check-in history information with accuracy of about 60%.

In order to demonstrate the usefulness of the customer level estimation, this thesis conducted a visitor promotion to attract "Best Customer" for Shibuya ShinQs Hikarie from among the users of the Shoplat service. The effectiveness was revealed by the relatively high view rate and visit rate of the "Best Customer", the most desirable customer level. In addition, store check-in history information was shown to be correlated with the goods sold by specific stores. This thesis clarified that customer level can be estimated by using these correlations.

The future plan is to focus on the using the time of check-in information and the spatial order of check-in for improving the accuracy of estimating the customer level. especially for "Best Customer". This research assumed that the customer has both purchase history and store check-in history. Of course, estimation accuracy is doubtful if the customer's store check-in history is small. One solution, grouping by shop type rather than individual stores, might maintain the estimation accuracy of desirable customers. The cold start problem should also be addressed, along with the prediction of customer churn.

Chapter 7

Conclusion

This thesis showed how to estimate the personal factors that influence purchase behavior from the consumer's purchase history and service usage history in order to make proposals for services and products suitable for each consumer. The effectiveness of the proposals made were clarified by a demonstration experiment. By understanding the estimated personal factors of consumers, one to one marketing can be realized that recommends products and services suitable for the consumer.

When consumers purchase products, they recognize the products from advertisements and word-of-mouth information. In deciding whether to purchase the products, they consider their condition and situation.

For consumers who are committed to the origin and materials of products, information distribution that matches or is similar to the conditions preferred by consumers is considered to be effective. In order to understand the situation and condition of consumers, it is common to ask consumers directly through interviews and questionnaires. However, this technique is burdensome to both sides, to the interviewer and the interviewee. To ensure adequate coverage, the interviewer often has to pay the interviewee, which raises costs. The interviewee takes time to respond to the queries. Furthermore, the disclosure of personal information becomes a psychological burden. For these reasons, it is difficult to obtain answers from most consumers.

Understanding consumers while eliminating the physical and psychological burdens imposed by asking the consumer to explicitly disclose personal information is the aim in this study. In order to realize this goal, the following three points are targeted. First point is to make it possible to identify the personal factors affecting purchase decision making from observable behavior information such as purchase and service use without directly disclosing personal information by questionnaire or interview to consumers. Second point is to clarify which observable information should be collected by clarifying the factors affecting purchase decision making. Third point is to clarify to what extent the estimated factors determine actual purchase behavior.

Personal influences and environmental influences are clarified as the key factors in purchase decision making. This study focused on lifestyle, preferences for products, family structure and purchasing intent which are thought to directly affect purchasing of daily necessities.

Chapter 3 focused on lifestyle, a standard tool used for consumer segmentation. This thesis proposed a method to estimate lifestyle and clarified its estimation performance. The proposal estimates lifestyle by extracting the purchasing behavior of products that strongly indicate consumer lifestyle from purchase history.

Chapter 4 focused on preferences for products. A method was proposed that can estimate the customer's preference for products from location information, which is one piece of behavior information, and its estimation performance was verified. The effectiveness and practicality of the preference estimation method were clarified by calculating the visit rate of target shops during a visit promotion campaign. This evaluation clarified how the estimated preference influences the consumer's purchasing behavior.

Chapter 5 focused on the family structure, i.e., the number of family members and the age of family members. Families with preschool children and families with high school students purchase different products and services. Therefore, this thesis proposed a method to estimate the family composition from the purchase history of daily necessaries and clarified its performance.

Chapter 6 focused on purchasing intent. By obtaining the degree of the consumer's ability to pay and the level of intentions to buy products for each store, target customers can be identified and suitably handled. Therefore, this thesis proposed a method to estimate customers desirable for the target store from store check-in histories including those of other stores. The effectiveness and practicality of the estimation method was confirmed by calculating the estimation accuracy of the good customer and the visit rate to the target shop during a store promotion.

As conclusions, this thesis proposed a machine learning-based method that can model the personal factors influencing purchase decision making and purchase history or service usage history. The estimation accuracy of personal factors are clarified by using real world purchase history data and service usage history. The personal factors influence on purchasing behavior are confirmed by large-scale experiments in an environment with real customers.

Purchasing history data for one year was used to estimate personal factors, lifestyle and family structure in this study. To estimate changes in personal factors, it is necessary to analyze data accumulated in the long term, which is the future plan.

Appendix

Product List

Rice

Material for pot dish

Butter

Bread

Caramel

Cheese

Cereal

Chewing gum

Other spread

Instant noodle cup

Snack

Furikake

Raw noodle

Side dish

Mix for boiled rice seasoned

Macaroni

Other oil

Cooked rice

Chocolate

Margarine

Sweet buns and stuffed bread

Candy

Jam and Marmalade

Instant noodle package

Biscuit and Cracker

Seaweed laver

Dried noodle

Rice cookie

Material for Ochazuke

Spaghetti

Confectionary from a toy manufacture

Other mixed seasoning

Nutritional supplement

Curry

Flour

Ice cream

Pasta sauce

Flour for fried chicken

Powdered milk for a baby

Premixed sauce

Premix powder

Instant cream

Miso-soup

Miso

Whipped cream

Frozen seafood

Cooking wine spirits

Condensed milk

Frozen meals and dinners

Low calorie sweetener

Yoghurt

Canned fish

Honey

Soy milk

Canned fruit

Tomato ketchup

Other noodles

Other confectionary

Stew

Tempura flour

Dessert

Stew mix

Bread crumb

Baby food

Soup

Soy sauce

Fresh cream

Soup stock for cooking

Salt

Skim milk

Frozen fruits and vegetables

Sugar

Milk

Other frozen food

Syrup

lactic drink

Canned vegetable

Sauce

Instant coffee

Canned meat

Coffee	Mayonnaise
Other Canned food	Tea
Dressing	Food in pouch
Cocoa	Spice
Japanese style food	Malt beverage
Extract	Western style food
Japanese tea	Other seasoning
Chinese style food	Barely tea
Sauce for grilled meat and shabu-shabu	Other cooked food
Chinese tea	Vinegar
Ham	100% juice
Ponzu vinegar	Meat sausage
Fruit juice drink	Hon mirin
Fish meat ham	Tomato juice
Mirin like seasoning	Fish meat sausage
Vegetable juice	Liquid soup stock
Roast pork	Cola
Flavored seasoning	Bacon
Soda pop	Shavings of dried bonito
Boiled fish paste	Other Carbonated drink
Dried sardine	Tube shaped fish paste
Coffee drink	Sea mustard and Sea tangle
Hanpen	Tea drink
Sauce for boiled cuisine	Fritter

Japanese tea drink

Other paste product

Combined seasoning

Energy drink

Kuzukiri

Nutritious drink

Cooked bean

Koya-dofu

Pasteurized lactic drink

Mekabu

Sesame oil

Beer

Wrapping film

Wine

Nail polish

Food packaging ware

Shochu

Cosmetic cotton

Food storage container

Cigarette

Health drink

Toilet paper

Electric tooth brush

Nourishing tonic

Chemical seasoning

Sports drink

Pickle

Gelatin noodle

Natto

Seaweed salad

Mineral water

Tsukudani

Salad oil and Tempura oil

Other beverage

Tofu

Whisky

Eyebrow pencil

Aluminum foil

Japanese sake

Perfume

Gas range cover

Other Japanese sake

Cosmetic accessory

Tissue

Tooth brush

Mini health drink

Other paper

Tooth paste

Paper towel

Mouth wash

Vitamin B1

Paper cleaner

Other oral hygiene material product

Vitamin E

Paper diaper for adults

Bathing agent

Calcium supplements

Sanitary shorts

Hair rinse

Combination cold remedy

Wrinkle removing agent

Other hair care product

Medicine for allergic

Disposable warmer

Tonic for hair

Antitussive agent

Agent for contact lens

Other cosmetic for men

Gargle

Other general merchandise

Neutral detergent

Sedative

Multivitamin

Wet tissue

Cleaning agent for denture

Vitamin C

Disposable diaper

Soap

Medicine for woman

Sanitary item

Shampoo

Herbal medicine

Water repellent agent

Hair treatment

Medicine for rhinitis

Adhesive bandage

Hair color

Analgesic antipyretic

Cotton swab

Anhidrotic deodorizer

Oral preparation

Therapeutic agent for pyorrhea alveolaris

Detergent for laundry

Ant dizziness drug

Dog food

Bleaching agent

Cat food

Softening agent

Medicine for respiratory disease

Pet products

Other detergent

Intestinal drug

Paper pack for vacuum cleaner

Cleanser detergent

Medicine for hemorrhoid

Face washing cream

Floor wax

Other medicine for organs

Skin lotion

Bath cleaner

Dermatologic preparation

Nutritive cream

Pipe cleaner

Eye drop

Skin beauty liquid

Adhesion cleaner

Contraceptive drug

Other Cosmetic product

Dust cloth

Cardio tonic drug

Sleepiness inhibitor

Other pet food

Laundry starch

Gastrointestinal agent

Razor

Kitchen detergent

Laxative

Cleansing agent

House detergent

Enema

Cold cream

Toilet cleaner

Antiphlogistic sedative drug

Milky lotion

Glass cleaner

Medicine for athlete’s foot

Face pack

Other house detergent

Other medicine for skin

Cosmetic paper product

Household gloves

Test kit

Skincare article

Sponge

Body care article

Draining bag

Other medicine

Etiquette article

Insecticide

Other health food

Makeup base

Antifungal agent

Perfume agent

Aromatic agent for toilet

Deodorizing agent

Dehumidifying agent

Herb medicine for kids

Sunscreen cosmetic

Waste cooking oil treating agent

Health food

Lip cream

Mothproof agent

Diet food

Foundation (Cosmetic)

Face powder

Cheek rouge

Lipstick

Other product for lip

Bibliography

[Arnold, 1984] Arnold, M. (1984). *The Nine American Lifestyles: Who We are and where We're Going*. Warner Books edition. Warner Books.

[Arnold et al., 1986] Arnold, M., Ogilvy, J., and Schwartz, P. (1986). *The VALS Typology: A New. Perspective on America*. SRI International.

[Boer et al., 2004] Boer, M., McCarthy, M., Cowan, C., and Ryan, I. (2004). The influence of lifestyle characteristics and beliefs about convenience food on the demand for convenience foods in the irish market. *Food Quality and Preference*, 15(2):155 – 165.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Buckley et al., 2007] Buckley, M., Cowan, C., and McCarthy, M. (2007). The convenience food market in great britain: Convenience food lifestyle (cfl) segments. In *Appetite*, volume 49, pages 600–617.

[Culotta et al., 2015] Culotta, A., Ravi, N. K., and Cutler, J. (2015). Predicting the demographics of twitter users from website traffic data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 72–78. AAAI Press.

[David, 1958] David, C. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society,Series B (Methodological)*, 20(2):215–242.

[Doi et al., 2017a] Doi, C., Ishii, A., Araki, T., Inamura, H., Ohta, K., Shigeno, H., and Katagiri, M. (2017a). Estimating value of customer through store check-in histories and its application for visitor promotion (in japanese). *IPSJ Transactions on Consumer Devices and Systems*, 7(2):115–124.

[Doi et al., 2017b] Doi, C., Katagiri, M., Araki, T., Ikeda, D., and Shigeno, H. (2017b). Family structure attribute estimation method for product recommendation system. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pages 167–173.

[Doi et al., 2018] Doi, C., Katagiri, M., Araki, T., Ikeda, D., and Shigeno, H. (2018). Is he becoming an excellent customer for us? a customer level prediction method for a customer relationship management system. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 320–326.

[Doi et al., 2016] Doi, C., Katagiri, M., Ishii, A., Araki, T., Inamura, H., and Ohta, K. (2016). Estimating value of customer through store check-in histories and its application for visitor promotion (in japanese). In *Multimedia, Distributed, Cooperative, and Mobile Symposium*, pages 735–741.

[Doi et al., 2017c] Doi, C., Katagiri, M., Ishii, A., Konishi, T., Araki, T., Ohta, K., Ikeda, D., Inamura, H., and Shigeno, H. (2017c). Estimating customer preference through store check-in histories and its use in visitor promotion. In *2017 Tenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, pages 1–6.

[Doi et al., 2017d] Doi, C., Katagiri, M., Ohta, K., and Shigeno, H. (2017d). Lifestyle prediction by capturing micro purchasing behavior (in japanese). *Journal of Information Processing*, 58(2):298–307.

[Doi et al., 2015] Doi, C., Konishi, T., Nakagawa, T., Katagiri, M., Inamura, H., and Ohta, K. (2015). Estimation of lifestyles focusing on purchasing behavior at purchase product level in multiple stores. (in japanese). In *Multimedia, Distributed, Cooperative, and Mobile Symposium*, pages 672–683. Information Processing Society of Japan.

[Dong et al., 2014] Dong, Y., Yang, Y., Tang, J., Yang, Y., and Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 15–24, New York, NY, USA. ACM.

[Duarte Torres and Weber, 2011] Duarte Torres, S. and Weber, I. (2011). What and how children search on the web. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 393–402. ACM.

[Facebook places, 2018] Facebook places (2018). `https://www.facebook.com/places/(Online)(2018.1.1).`

[Foursquare, 2018] Foursquare (2018). `https://foursquare.com/(Online)(2018.1.1).`

[Hayashi et al., 2014] Hayashi, A., Matsubayashi, T., and Hiroshi, S. (2014). Regular behavior measure for location based services (in japanese). *DBSJ Japanese Journal*, 13-j(1):64–71.

[Hisamatsu et al., 2012] Hisamatsu, T., Asahi, Y., and Yamaguchi, T. (2012). Comparative analysis of daily goods purchasing patterns using id-pos data in drugstore (in japanese). *Operations research as a management science research*, 57(2):63–69.

[Hu et al., 2007] Hu, J., Zeng, H.-J., Li, H., Niu, C., and Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 151–160, New York, NY, USA. ACM.

[Huang and Dong, 2016] Huang, C. and Dong, W. (2016). Unsupervised interesting places discovery in location-based social sensing. In *2016 International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 67–74.

[Ishigaki et al., 2010a] Ishigaki, T., Takenaka, T., and Motomura, Y. (2010a). Computational customer behavior modeling for knowledge management with an automatic categorization using retail service's datasets. In *IEEE International Conference on E-Business Engineering (ICEBE)*, volume 25, pages 528–533.

[Ishigaki et al., 2010b] Ishigaki, T., Takenaka, T., and Motomura, Y. (2010b). Customer-item category based knowledge discovery support system and its application to department store service. In *2010 IEEE Asia-Pacific Services Computing Conference*, pages 371–377.

[Ishigaki et al., 2011a] Ishigaki, T., Takenaka, T., and Motomura, Y. (2011a). Customer behavior prediction system by large scale data fusion in a retail service (in japanese). *The Japanese Society for Artificial Intelligence*, 26(6):670–681.

[Ishigaki et al., 2011b] Ishigaki, T., Takenaka, T., and Motomura, Y. (2011b). Improvement of prediction accuracy of the number of customers by latent class model (in japanese). In *The 25th Annual Conference of the Japanese Society for Artificial Intelligence*, volume 25, pages 1–4. The Japanese Society for Artificial Intelligence.

[Jingyan et al., 2016] Jingyan, Y., Chuanren, L., Mingfei, T., March, L., and Hui, X. (2016). Buyer targeting optimization: A unified customer segmentation perspective. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1262–1271.

[John and Langley, 1995] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Koshiba et al., 2013] Koshiba, H., Ishigaki, T., Takenaka, T., Sakurai, E., and Motomura, Y. (2013). Customer modeling method constructed from behavioral data and lifestyle survey (in japanese). *The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and Systems Society*, 133(9):1787–1795.

[Li et al., 2016] Li, H., Ge, Y., Hong, R., and Zhu, H. (2016). Point-of-interest recommendations: Learning potential check-ins from friends. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 975–984, New York, NY, USA. ACM.

[Lindquist and Sirgy, 2009] Lindquist, J. D. and Sirgy, M. J. (2009). *Shopper, buyer, and consumer behavior: Theory, marketing applications and public policy implications.* Atomic Dog/Cengage Learning.

[Liu et al., 2017] Liu, X., Xu, A., Akkiraju, R., and Sinha, V. (2017). Understanding purchase behaviors through personality-driven traces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 1837–1843, New York, NY, USA. ACM.

[Lu et al., 2015] Lu, S., Zhao, M., Zhang, H., Zhang, C., Wang, W., and Wang, H. (2015). Genderpredictor: A method to predict gender of customers from e-commerce website. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 13–16.

[Matsumoto and Saigo, 2013] Matsumoto, K. and Saigo, A. (2013). Data analysis competition assignment setting department:customer segment forecast for ec site users (in japanese). *Operations research as a management science research*, 58(2):68–73.

[Mislove et al., 2010] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 251–260, New York, NY, USA. ACM.

[Murray and Durrell, 2000] Murray, D. and Durrell, K. (2000). Inferring demographic attributes of anonymus internet users. In *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, WEBKDD '99, pages 7–20, Berlin, Heidelberg. Springer-Verlag.

[Ohata et al., 2015] Ohata, Y., Ohno, A., Yamasaki, T., and Tokiwa, K.-i. (2015). An analysis of shopping behavior pattern and purchase amount in the inner areas of the sales floor in a retail store (in japanese). *Forum on Information Technology*, 14(2):297–302.

[Ponta, 2018] Ponta (2018). `http://www.ponta.jp/(Online)(2018.1.1).`

[Rakuten-check, 2018] Rakuten-check (2018). `https://check.rakuten.co.jp/(Online)(2018.1.1)`.

[Rakuten, INC., 2018] Rakuten, INC. (2018). `https://global.rakuten.com/corp/(Online)(2018.1.1)`.

[Riedmiller, 1994] Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons — from backpropagation to adaptive learning algorithms. *Computer Standards and Interfaces*, 16(3):265–278.

[Roger et al., 1995] Roger, D. B., Engel, F. J., and Paul, W. M. (1995). *Consumer behavior*. Dryden.Press.

[Shibuya Hikarie ShinQs, 2018] Shibuya Hikarie ShinQs (2018). `https://www.tokyu-dept.co.jp/shinqs/(Online)(2018.1.1)`.

[Shopkick, 2018] Shopkick (2018). `http://www.shopkick.com/(Online)(2018.1.1)`.

[Shoplier, 2018] Shoplier (2018). `https://shoplier.jp/(Online)(2018.1.1)`.

[Siyu et al., 2015] Siyu, L., Meng, Z., Hui, Z., Chen, Z., Wei, W., and Hao, W. (2015). Genderpredictor: A method to predict gender of customers from e-commerce website. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 13–16.

[Straughan and Roberts, 1999] Straughan, R. D. and Roberts, J. A. (1999). Environmental segmentation alternatives: a look at green consumer behavior in the new millennium. *Journal of Consumer Marketing*, 16(6):558–575.

[T-Point, 2018] T-Point (2018). `http://tsite.jp/(Online)(2018.1.1)`.

[Tokyu point Tokyu card, 2018] Tokyu point Tokyu card (2018). http://www.topcard.co.jp/(Online)(2018.1.1).

[Twitter, 2018] Twitter (2018). https://twitter.com/(Online)(2018.2.14).

[Wang et al., 2016] Wang, P., Guo, J., Lan, Y., Xu, J., and Cheng, X. (2016). Your cart tells you: Inferring demographic attributes from purchase data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 173–182, New York, NY, USA. ACM.

[Wells and Tigert, 1971] Wells, W. D. and Tigert, D. J. (1971). Activities, interests, and opinions. *Journal of Advertising Research*, 11(4):27–35.

[Yao et al., 2016] Yao, Z., Yanjie, F., Bin, L., Yanchi, L., and Hui, X. (2016). Poi recommendation: A temporal matching between poi popularity and user regularity. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 549–558.

[Ye et al., 2010] Ye, M., Yin, P., and Lee, W.-C. (2010). Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 458–461, New York, NY, USA. ACM.

[Ye et al., 2011] Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 325–334, New York, NY, USA. ACM.

[Yelp, 2018] Yelp (2018). https://www.yelp.com/(Online)(2018.1.1).

[Yonghong and Xingguo, 2015] Yonghong, Y. and Xingguo, C. (2015). A survey of point-of-interest recommendation in location-based social networks. In *Proc. of 29th AAAI conference on Artifical Intellifence*, pages 53–60.

[Zhang et al., 2016] Zhang, D. C., Mei, L., and Chang Dong, W. (2016). Point of interest recommendation with social and geographical influence. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1070–1075.

[Zhao et al., 2016] Zhao, S., Irwin, K., and Michael, L. (2016). A survey of point-of-interest recommendation in location-based social networks. *CoRR*.

[Zhong et al., 2013] Zhong, E., Tan, B., Mo, K., and Yang, Q. (2013). User demographics prediction based on mobile data. *Pervasive Mob. Comput.*, 9(6):823–837.

[Ziff, 1971] Ziff, R. (1971). Psychographics of marketing segmentation. *Journal of Advertising Research*, 11(2):3–9.

Acknowledgement

I would like to express my heartfelt gratitude toward Professor Hiroshi Shigeno for his great supports and advices to obtain this doctoral degree.

I would also like to thank Professor Masafumi Hagiwara, Professor Hiroaki Saito and Professor Yoshihisa Shinozawa for reviewing this thesis and giving me valuable suggestions.

I have been supported and helped by many respectable fellows in NTT DOCOMO, INC..

Especially, Doctor Masaji Katagiri gave me various gifts such as the ability of thinking logically and skills in writing technical papers.

Doctor Daizo Ikeda gave me insightful comments and a favorable environment for researching .

Mr. Takashi Araki, Doctor Hiroyuki Sano and Ms. Sawa Korogi supported me to keep a good balance of job and school work.

Doctor Ken Ohta gave me constructive comments and valuable suggestions.

I would like to show my greatest appreciation to Mr. Wataru Takita and Doctor Narumi Umeda.

Professor Hiroshi Inamura from future university Hakodate gave me constructive comments and warm encouragement.

I have been supported and helped by many respectable fellows in Shigeno Laboratory.I have enjoyed spending time with all laboratory members. I am very glad I shared many rewarding experiences with them.

I would like to offer my special thanks to researcher fellow. Doctor Miki Enoki and Mr. Mori Kurokawa always gave me warm encouragement.

I am also very grateful to the NTT DOCOMO, INC. for making my Ph.D. study possible by the financial support.

Finally, I would like to sincerely thank my family for supporting me always. I could not accomplish this work without their support and understanding.

Chiaki Doi
August 2018