CSIS Discussion Paper No. 47

An Approach to Extracting Unexpected Patterns from Massive Attributes: Understanding of Condominium Purchasers ' Behavior through Data Mining

May, 2002

Jungmin Choi*, Asami Yasushi**

* Department of Urban Engineering, University of Tokyo
** Center for Spatial Information Science, University of Tokyo

Center for Spatial Information Science
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

An Approach to Extracting Unexpected Patterns from Massive Attributes: Understanding of Condominium Purchasers ' Behavior through Data Mining[1]

Jungmin Choi[2], Asami Yasushi[3]

ABSTRACT

Marketing analysis is an indispensable step to establish strategies for the condominium supply for developers. In this context, it is crucial to grasp the potential buyers' behavioral characteristics as "rule" or "pattern" extracted from historical transaction data. The question that is often asked by developers is what kind of condominium customers are pursuing and who they are. So far, strategies in housing supply side rely mainly on the experience or intuition of the marketers/developers, and not from the comprehensive data analysis. In this paper, a new approach is demonstrated to extract informative and unexpected rules/patterns from a database with relatively small contract records but massive attributes. This approach can summarize information targeted to unexpected patterns in data structure using data mining methodology. For the validity of the proposed method, real data of about 800 condominium purchasers in Tokyo metropolitan area is analyzed. Our approach may contribute to develop the decision support systems that can provide managers with more innovative ideas from collected data in generating marketing strategies.

Keywords: Behavioral Characteristics of Purchasers, Data mining, Unexpected Patterns, Condominium, Marketing

---

[1] The authors would like to thank Recruit Co. Ltd., for kindly providing their invaluable data.

[2] Graduate Student M.Eng., Department of Urban Engineering, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 (haesong@ua.t.u-tokyo.ac.jp).

[3] Professor Ph.D., Center for Spatial Information Science, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan (asami@csis.u-tokyo.ac.jp).

1. INTRODUCTION

Recently there has been an apparent move of many industries towards customer oriented approach. In the midst of increasing competition, there is indeed a need to better understand of customers, and to quickly respond to their individual needs and wants. In the housing industry, especially the case of condominium market, this approach is still at its infancy stage. Nevertheless, due to many survey activities conducted for newly built condominium purchasers at contract stage, there is a wealth of data available to gain a better understanding of customer types and customer behavior. Most of major real estate companies in Japan have collected this sort of customer database in order to reflect customers' needs and tastes in designing of unit plan and ultimately to improve sales of their newly built condominiums. Recently there is new trend of supply like super high-rise condominiums located in city centers, large scaled buildings, open plan, etc in Japan. These may be the typical illustrations of what are consumers' new needs and requires which should be reflected in the new marketing strategies.

In general a detached house is a unique product which cannot be substituted, whereas in the case of condominium it has the characteristic of general goods because the process from designing to construction is typical and similar to the purchase of general goods. This implies that condominium business has similar aspects of selling branded goods and as such, market analysis is indispensable element in planning business strategies. This kind of recognition encourages real estate companies to start gathering significant number of large and heterogeneous databases related to their properties. On occasion, the collected databases need to be analyzed and applied to develop new business strategies and to identify opportunities, but the problem that most of them faced is that "We are rich in data and poor in information" (Bounsaythip and Rinta-Runsala, 2001).

Data mining has attracted marketers as a solution to this problem. In business, this emerging technology has been focused in market segmentation, customer profiling, risk analysis, and many other applications. It seems that the importance and the potential capacities of data mining have been well recognized to decision makers of housing market particularly in private sector, but in reality the collected data are not fully employed for this purpose. One of the reasons may be attributed to the lack of knowledge about data mining of the marketers and of proper techniques to apply. In fact, even though a huge data mining related literatures have been introduced and being applied in many fields, the actual samples applied in housing market are very scarce. A couple of reasons for this can be cited. For one, it is often very hard for a researcher to obtain detailed information about private housing sector, and even after successful gathering of this information, but owing to business confidentiality there are usually limitations on opening them to the public. The other is that the primary focus in data mining is oriented to discover meaningful rule patterns from a massive transaction database, which is somewhat different from our research in the context of relatively small contract records but massive attribute information. This causes a little

difficultly in implementation or application of data mining technology fully in real estate field. Indeed, the content of our dataset ranges from basic demographic attribute information to purchaser's behavioral information, which amounts to hundreds of attribute fields for each contractor.

As mentioned above, despite the increasing needs of market analysis of newly built condominium there are few empirical studies in the literature, which leads marketers or decision makers in this area to depend heavily on their experiences or intuitions. This motivated our study to develop a rather straightforward solution to extract and summarize unexpected or exception rule pattern from high (multiple) level of data structure space into low (simple) level of information space through the data mining method. To put it concretely, we propose a method to establish a good understanding of the current state of customer needs and behavioral characteristics on the basis of exception rule mining method. We expect that a clear and comprehensive market analysis of consumers' needs and behavioral characteristic in condominium can help to avoid the mismatches between supply and demand and cut down the unsold stock of condominium.

This article is organized as follows. In section 2 we introduce the background and a review of exception rule mining which explains the foundation of our method. In section 3, the algorithm that we proposed is outlined and discussed. Our proposed algorithm is applied to a large, real world condominium database in section 4. This case study is extensively described including the behavior characteristics of condominium purchasers in our research. Finally, the conclusions are given in section 5.

2. A REVIEW OF LITERATURE

2.1 Background

It should be noted that our dataset has a relatively small contract records with massive attribute information, and our focus is to grasp unexpected rule patterns projecting complicated data structure into simple information space. In this sense it could be referred to summarizing of information aimed at unexpected patterns in data structure. For this goal, we have tried to apply a variety of methodologies from conventional statistics to the latest neuro-computing methods. But we failed to find a meaningful result for this problem. Some of the representative methods that we have taken into consideration may be categorized into two domains from the perspective of linearity in relation to attribute variables. One is linear method, i.e., a conventional statistical methods represented by PCA (Principal Component Analysis), CA(Correspondence Analysis), and MDS(Multidimensional Scaling), etc. These traditional analytic methods are frequently termed as a data summarizing technique which is based on the assumption of linearity among attribute variables, and would perform excellently under ideal condition. However, when there is an increase of a vast number of attribute variables, and their relationship does not hold from the context of linearity, it would not operate properly. The other is as a non-linear method, including a sort of

neuro-computing methods, like ANN (Artificial Neural Networks), SOM (Self-Organized Maps), etc. It is reported that these comparatively recent techniques work well compared to conventional techniques with complicated data, but understanding of training processes of these techniques is quite difficult and in some cases the results varied on setting parameters and work environment (Worzala, Lenk and Silva, 1995). In either case if the number of attribute variables increases tremendously, then it is virtually impossible to extract summarized information.

Conventional data mining techniques are oriented to find mainly strong patterns. There is a good example why the exception rule mining is of importance instead of general data mining in business area (Liu et al., 1999; Padmanabhan and Tuzhilin, 1999). They illustrate the usefulness of unexpectedness as a measure of interestingness in KDD (Knowledge Discovery in Databases), and criticism about some of drawbacks in KDD by Padmanabhan and Tuzhilin (1999) can be summarized as follows:

1. It generates a very large number of rules, and most of them are obvious or irrelevant.
2. Because these obvious rules are mainly based on objective measures, such as confidence and support, they are not so informative to marketers, thus need additional introduction of subjective measure as unexpectedness.
3. Most of these existing algorithms are primarily data-driven and do not fully exploit domain knowledge and intuition that managers in a business environment have.

Exception rules are those weak patterns outside the strong ones. Usually, such patterns (reliable exceptions) are unknown, unexpected, or contradictory to what the user believes. Thus, exception rule is often beneficial since it differs from a common sense. Naturally, the term surprise (unexpectedness) or usefulness should be argued from a “measure” point of view. There are a good number of literatures on measure of interestingness (Freitas, 1999; Silberschatz and Tuzhilin, 1996). This issue involves an evaluation of the patterns discovered, and generally there are two measures; an objective (data-driven) and a subjective (user-driven). Objective measure of interestingness is to capture the statistical strength of a pattern and ordinarily use “confidence” and “support” for the evaluation of the discovered patterns. On the other hand, subjective measures are related to a set of domain knowledge given by expert, assuming that the interestingness of a pattern depends on the user and does not only depend on the statistical strength of pattern.

In conclusion, it could be pointed out that for our problem an approach from statistics may be mainly targeted to summarizing of whole data structure with relatively small dataset, while an approach from data mining may be primarily oriented to discover a sort of strong patterns with a massive dataset. Hence, for a dataset like ours with relatively small size of records, compared to dataset of general data mining, and large size of attributes, there should be a new approach for both summarizing data structure and extracting informative rules for marketers

in accordance to users' demand in a given dataset. Note that in this article our concern is to extract only unexpected rule patterns and not all information from the perspective of data structure, leaving arguments on a comparison of statistics and data mining intact.

2.2 A review of literature related to exception rule mining

There are a few proposed exception rule mining methods in the literature. An approach based on a syntactic comparison between a rule and a belief, labeled as ZoomUAR (Padmanabhan and Tuzhilin, 1999), based on detecting occurrences of Simpson's paradox (Fabris and Freitas, 1999), and also there's an approach based on contingency table presented by Liu et al. (1999). In particular, the proposal by Liu et al. (1999) is quite similar to one proposed in this paper. But what makes it different is that they paid attention to "deviation" of actual frequency and expected frequency in a cell of given contingency table and identify outstanding negative deviations as reliable exceptions. However, in this algorithm it is not clarified on how to handle "low expected frequencies" in a contingency table and how to actually establish a point on the number of cells to be selected. We will revisit the issue of low expected frequencies and selection criteria for significant cells in section 3.2 and 3.4 respectively.

3. PROPOSED ALGORITHM

3.1 A contingency table

One may try to grasp the relationship among attribute variables using the association measure, and for this purpose there have been many association measures utilized. For instance, like Pearson's product moment correlation coefficient which is a measure of the linear association between two variables, there are a number of different correlation measures on the basis of the kinds of variables being studied. In particular, in social sciences it is very natural that category variables are frequently employed in many circumstances. Age, for instance, when we categorize it, classification is more general than actual age variants, like 20s, 30s, 40s, 50s, or above 60s, as the way presented in this article. Therefore, we focused on the category (discrete) variables and a combination of two category variables or a contingency table in order to grasp the relationship across variables. For this, it is necessary to classify attribute variables in accordance to their properties, and in the case that an attribute has continuous values, discretization is required. Figure 1(a) shows a simple example of a dataset, which has four attributes: two target variables and non-target variables. An attribute variable can be assigned to target variable or non-target variable in terms of users' concern. Here, a target variable indicates a variable on which a researcher focuses in terms of his/her main interest.

Contingency tables have been used in this article to extract exception cells or significant cells from the perspective of relationships across variables in the data. Figure 1(b) shows a combination of attribute variables. For convenience, we put target variables in row and

non-target variables in column in a combination table. Thus, the possible number of contingency tables in a combination table in Figure 1(b) is multiplication of number of target variables and number of non-target variables. A contingency table out of a combination table is exemplified in Figure 1(c), where r categories of a target variable are in row and c categories of a non-target variable are in column. A contingency table, also called a cross reference table, is a table showing the number of records for each value combination of two or more variables that constitute the table. In a two-dimensional $r \times c$ contingency table shown in Figure 1(c), a sample of N observations is classified with respect to two category variables; a target variable and a non-target variable. The entries in the cells in the contingency table are frequencies. These may be transformed into proportions or percentages but it should be noted that the data were originally frequencies or counts rather than continuous measurements.

Assume that a statistical model that a frequency n_{ij} in a given cell $R_i C_j$ is the observation of a probability variable N_{ij} which follows discrete distribution, then row total and column total are as follows.

(1) $$n_{i+} = \sum_{j=1}^{c} n_{ij}, \quad n_{+j} = \sum_{i=1}^{r} n_{ij}, \quad n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

Under the assumption that two category variables are independent, probability, P_{ij}, of an observation falling in the cell $R_i C_j$ of the table is simply the product of the marginal probabilities, P_{i+} and P_{+j}, in the equation (2).

(2) $$P_{ij} = P_{i+} \times P_{+j} \quad (i = 1, \cdots, r; j = 1, \cdots, c)$$

Suppose that a cell where its observation exceeds far from the expected frequency $N \cdot P_{ij}$, this implies that an event given by row and column condition is somewhat surprising or unexpected and vice versa. Namely, if an event occurs far less than expected, it also attracts concern or interest. Hence this sort of unexpectedness, or precisely speaking, events over- or under- the expectation could motivate marketers/decision makers to investigate the reason of those events. In this context, we propose an exception rule mining method using contingency tables.

The framework of the proposed method is outlined in Figure 2. In phase 1, as a pre-process in data mining like cleaning, selecting variables for mining is performed, and some sample cases for this process will be given in section 4. After data preparation, for all combinational contingency tables as shown in Figure 1(b) an extraction process of exception cells or

significant cells is carried out using the extraction measure. The detail process of extracting significant cells will be argued in section 3.4. Finally, for the extracted significant cells association rule mining method will be applied for the purpose of identifying relationships between significant cells described in section 3.5.

3.2 Issues related to contingency tables

There are some issues which should be considered when utilizing contingency tables. The most important question may be whether the category variables forming the contingency table are independent or not, because the probability P_{ij} in equation (2) is based on the assumption that two category variables are independent. For this purpose, χ^2 test is normally utilized for the test of hypothesis of independence. When this test applied, it is generally required to satisfy that the sample size is large and the expected frequencies are not too small, because the statistic χ^2 can be approximated to χ^2 distribution when sample size is large enough and identical for events (Dunning, 1993). A significant overall χ^2 test for an $r \times c$ contingency table indicates that the variables forming the table are not independent, but provides no information as to whether the lack of independence occurs throughout the table or only in a specific part. Whatever may be the case, for application of the proposed method, it should be desirable that sample size is large enough and the relationship between two category variables is not too strong.

Another issue related to contingency table is low expected frequencies. One of the assumptions made when deriving the χ^2 distribution is that the expected frequencies should not be too small, since otherwise the assumption of χ^2 distribution would not be acceptable. Namely, for small, sparse or skewed data the asymptotic theory may not be valid, although it is often difficult to predict a priori whether a given data set may cause problems. There are some arguments on how large of the expected frequencies in a cell should be (Everitt, 1992). One of the alternatives to deal with low expected frequencies is to analyze the collapsibility of the categories that originate in the cells with low expected frequencies, where concerning the possible loss of information due to the aggregation of categories (Ocerin et al., 1999). In relation to low expected frequencies, there is also a problematic issue; tables with a priori zeros. In the case of sampling zeros the solution may be either to increase the sample size or to add a small positive constant to each cell frequency. In many situations, however, tables arise in which it is theoretically impossible to have observations in a cell; in this case the empty cells are usually referred to as structural zeros, and the table as a whole is incomplete. In addition, two-dimensional contingency tables where the row and column variables have the same number of categories occur fairly frequently in practice and are known in general as square tables. For such tables, hypotheses relating simply to independence are not of major importance, instead interest centers on testing for symmetry and marginal homogeneity (Everitt, 1992). In summary, regarding contingency tables, it is recommended that several issues, as mentioned above, should be well recognized and carefully taken into consideration

in prior.

3.3 Alternatives to Extracting of significant cells

There are several measures that could be used to extract cells in accordance to a user's concern in a contingency table. Some alternatives for this measure are shown in the following. One possible measure could be the one that picks up the cells in terms of frequency ranking order. The cells extracted by this measure might be interpreted as a strong pattern, but it has a drawback from the viewpoint that since the pattern is very strong, therefore it is not of much interest. At the same time, distribution of extracted cells can vary according to the way of discretizing category variables. Another alternative could be the one that is based on the concept of "division" between observation and expected frequency in a cell, which is similar to one proposed by Liu et al. (1999). This alternative can be considered as a measure that is used to extract weak pattern cells or reliable exception cells. However, it is not clear on how to actually establish a criterion on the number of cells to be selected. The last alternative in this article is the revision of the second measure by establishing criteria of selection on the basis of statistical test whose detail information will be given in the next subsection. It could be cited that this alternative is the most well-grounded measure among the three proposals for our problem because it has clear statistical ground. In fact, our experiments with the dataset using the above three measurements suggest that the last alternative is superior to the others as exception rule mining measurement. Therefore, using the last measure we provide more detailed procedures for extraction of significant cells from a contingency table in the next subsection.

3.4 Detailed procedures for extraction of significant cells

As mentioned above, "difference" between actual observation and expected frequency in a cell being studied can give a good ground for extracting exception rules in contingency tables. Nevertheless because there should be more proper arguments given, consequently, the following four major steps are established in order to extract significant cells in a given contingency table. All significant cells are to be extracted throughout the following procedures that should be conducted in all combinations of target variables and non-target variables shown in Figure 1(b).

Step 1: Prepare an $r \times c$ contingency table

Initially, a contingency table should be made in advance, and in this time the number of contingency tables to be made is the number of combinations between target variables and non-target variables. A pair of category variables in a combination table consists of a target variable (in row) and a non-target variable (in column) as outlined in Figure 1(b). Then, calculation of ratio of frequencies and expected probabilities in each cell can be easily made.

Step 2: Regard distribution of a contingency table as a binomial distribution and then approximate it to a normal distribution

Under the assumption that two variables are independent, it is possible to regard distribution of frequencies for each cell in the table as binomial distribution. Binomial distributions arise commonly in statistical analysis when the data to be analyzed is derived by counting the number of positive outcomes of repeated identical and independent experiments. The task of counting observations in a contingency table can be cast into the form of a repeated sequence of binary trials comparing each observation in a cell with the case being counted. These comparisons can be viewed as a sequence of binary experiments. To the extent that these assumptions of independence and stationarity are valid, we can switch to a binary distribution of cells concerning Bernoulli trials. Defining a probability variable which conforms to a binary distribution as $X \sim B(n, p)$ with parameters n, p (number of trials and probability of events), then its expectation and variance are $E(X) = np,\ V(X) = np(1-p)$, respectively. If n is large enough, then cumulative probability $P(X \le x)$ of a binomial distribution can be approximated into cumulative probability of a normal distribution in terms of normal function Φ in the following equation (3). if n is not very large, the probability calculation can be improved by using the continuity correction, which considers that each whole number occupies the interval from 0.5 below to 0.5 above it. When an outcome x needs to be included in the probability calculation, the normal approximation uses the interval from $(x-0.5)$ to $(x+0.5)$.

(3) $$P(X \le x) \approx \Phi(u),\ u = \frac{x \pm 0.5 - np}{\sqrt{np(1-p)}},\ \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Step 3: Remove cells with minority frequencies in the table

As stated above, small frequencies in contingency table is problematic. There are some arguments on how large frequencies should be, however it is empirically known that normality assumptions are generally considered to hold well enough when $np(1-p) > 5$ (Dunning, 1993; Everitt, 1992). The agreement between the binomial and normal distributions is exactly what makes test statistics based on assumptions of normality so useful in the analysis of experiments based on counting. Moreover, from the practical aspect, given any cell with zero or near to zero frequencies, which frequently occur in real situation, all these cells could fall into significant cells on the basis of the proposed measure. Thus, if these small frequencies in a contingency table are not taken into consideration, then this could distort real information of data structure. Consequently there is a need to exclude those cells with scarce frequencies, as a result the proposed criterion is to calculate a theoretical number. That is, setting the frequency on binomial distribution to zero $(x = 0)$, to calculate the theoretical probability, p, which is less than 95 % significance level given by the form as:

(4) $$f(x) = {}_nC_x p^x (1-p)^{n-x} \le 0.05, \;\; E(x) \ge n(1-0.05^{n^{-1}})$$

In this paper, since the number of total observation is 786 as shown in section 4, the theoretical number of frequency calculated by the proposed concept above is near to 3. Thus, we set number 3 as a standpoint for our small frequencies' problem. Certainly because the 95% significance level above to some extent is arbitrary, one could set the desirable level to suit its purpose.

Step 4: Extract significant cells only at a given significance level

In the final step, for the cells processed from Step 1 through Step 3, it is required to determine whether a given cell is suitable to be selected as a significant cell. For this end, we compare the actual frequencies to theoretical frequencies processed in Step 2 and Step 3 over all cells in a contingency table using statistical significance level. Significant cells can be selectively extracted by picking up the cells in which frequencies are out of the theoretical frequencies calculated in Step 2 under a given statistical confidence level on normalized distribution N(0, 1), excluding the cells which are the case of scarce frequencies in Step 3. Specifically, we calculate the value of intensity of relationship for each significant cell by the following equation (5):

(5) $$I = \log(\mathit{freq} / \mathit{theor})$$

where *freq* is defined as actual frequency and *theor* is defined as theoretical frequency. For instance, comparing two significant cells A and B as I with 10.0 and 2.0 respectively, this can be interpreted as A is stronger than B in intensity of unexpectedness. Intuitively A may be more surprising than B, but we believe that the intensity here should be investigated more exhaustively. It should be noted that there are two types of significant cells; a cell surpasses theoretical expectation (cell-over-expectation,↑) and a cell fall below theoretical expectation(cell-under-expectation, ↓).

3.5 Significant cells according to target variables and application of association rule mining

By arranging the extracted significant cells according to target variables, to a certain extent latent information can be uncovered outside one's domain knowledge. It should be noted that the extracted significant cells in terms of the procedures discussed in section 3.4 is nothing but a collection of pairs or fragmented information that are regarded significant from the perspective of relationship between given target variable and non-target variable. As an illustrative example shown in section 4, there are two pairs of significant cells: <(Sex of household: FEMALE) - (Unit type: ONE ROOM)> and <(Sex of household: FEMALE) -

(Unit type: TWO ROOM)>. This means that if it is a female household then it is an exception pattern that her unit type of condominium is one or two rooms. This could be interesting, but the background or reason for this is not clear at this stage. Thus, if the relations between significant cells or fragmented information can be investigated, comprehensive information with consistency can be captured. In other words, the relationship among significant cells should be systemically synthesized for consistency.

For this purpose, we pay attention to formation of significant cells sorted by target variables. Namely, if we regard a set of significant cells, arranged by levels in a given target variable, as a set of frequent items, then it can be interpreted that in a cluster of frequent items (significant cells) there is close relation between the items. In fact, this concept is none other than association rule mining, which searches for interesting relationships among items in a given data set. This idea is motivated for market basket analysis, where the customer buying habits are analyzed by finding associations between the different items that customers place in their shopping baskets. In general, association rules are in the form of: "Head → Body [support, confidence]". In this paper, we exploit "a-priori" algorithm (Agrawal and Srikant, 1994), but as an alternative method to association rule for this purpose, "Rough set" method also could be taken into consideration because both of the algorithms can be handled from the perspective of a set theory. For investigation of the relationship between significant cells using association rule mining, the procedures and techniques for setting parameters in the mining could be the same as generally conducted.

4. APPLICATION OF THE PROPOSED METHOD AND DISCUSSIONS

4.1 Data

We tested our algorithm on condominium purchasers' contract database provided by Recruit Co. Ltd. The data consist of 798 household records in five wards of Tokyo metropolitan area in Japan (Shibuya, Setagaya, Ota, Shinagawa, Meguro) with over 200 attribute fields(tuples) from January 2000 to January 2001's transactions. The attribute fields are composed of a variety of information about demographics, behavior pattern, and the properties they actually purchased. As stated in section 2, with the objective of applying our approach to the dataset, some preprocess steps should be taken in advance. Firstly, selection of variables from the raw data, as well as designating them as target variables and non-target variables should be carried out properly so as to grasp a good understanding of the characteristics of condominium households. At the same time, because all variables used in the proposed approach should be discrete category variables, non-discrete variables should be discretized or categorized with proper size of levels in order to avoid some possible annoying problems discussed in section 3.2 in designing contingency tables. This issue may be referred as a "feature selection" matter in choosing proper variables and their levels (Ocerin et al., 1999).

As for the selection of variables to be used we paid attention to the well-known key

questions among real estate agencies; 4W+1H "Where, Who, Why, What, How" of customers' decision on final purchase. These questions are buyer oriented in respect to how a condominium can appeal to buyers, how it can offer a viable solution to buyers' housing needs and wants, and how best to convey the messages. According to these criteria, we selected and categorized the variables into three main categories as shown in Table 1. There are three main categories indicating "Where & Who", which represents household attributes (demographics variables), "How & Why" as behavior (behavioral variables), and "What" as property characteristics (condominium variables). As for discretization of the selected variables, since most of them are originally discrete, we integrated levels of the variables into larger levels rather than subdividing them. Because the variables are categorized too detailed in the raw data, otherwise there would be too many zero frequencies in contingency tables.

Moreover, for each behavior variables we classified households into some meaningful segments using cluster analysis. These segmented groups of households, or market segmentation, which normally defined as the dividing of groups of customers into sub-markets or segments, also could be a good result of categorization. With respect to market segmentation, each size of segments produced by segmentation should be an appropriate size for effectiveness and profitability of marketing strategies (Wedel and Kamakura, 1998). This implies that there would not be extremely small sized segments and as a result, it could be expected to reduce low values' frequencies in a contingency table. Consequently, the resulting variables selected are totally 27 variables shown in Table 1. Groups of target variables in this research are either of behavior and condominium variables, but most of the discussions in the remainder are focused on condominium variables as target variables. In this case, non-target variables are the rest demographics and behavior variables, whose levels are presented in Table 2. In this article, by limiting only on behavior variables in Table 1, we will provide more detailed information about the resulting categorization and characteristics of each segment in the next subsection.

4.2 Behavior characteristics

Motivation of purchase

The image of his or her own choice of housing varies in terms of individual's motivation. For instance, there is a gap in choice between household that purchase with the reason "Because of a child growing up, (s)he should change to a larger house", and the reason "because (s)he wants to live near the seaside leisurely". This questionnaire was delivered to households that have purchased, and asked on their motivation of purchase based on 23 criteria with 5 at maximum. We divided households into 4 categories on the basis of cluster of respondents; "investment (B11)", "acquisition of ownership (B12)", "acquisition of larger house (B13)", and "no particular reason (B14)". The most frequent answer is "investment" followed by "acquisition of ownership".

Concerned items as important factor

It is natural that the image of housing choice varies on what customers pursue. This question was asked about the important criteria in the selection. In this item, we categorized households into 5 segments on the basis of cluster of their responses; "price(B21)", "accessibility(B22)", "location(B23)", "building design(B24)", and "facilities and others(B25)". The most frequent answer is "price" which covers more than 70 %, followed by "accessibility".

Search pattern

For instance, there is difference in housing image between a household that "gathers information before visiting a model room" and a household that "visits a model room before gathering further information". It may be inferred that the former compares and makes decision after full investigation, whereas the latter can be categorized as impulse buyer. Such a difference in search patterns also seems to closely relate to the behavior of customer's housing choice.

Information sources for searching housing

The characteristics of customers also appeared to be closely related to where they obtain the information from. For example, comparing two customer layers, one customer layer that relied heavily on internet use and the other that favor of leaflet, there may appear quite different characteristics between them. From this point of view, we sorted seven representative information sources as: housing magazine, newspaper advertisement/leaflet, direct mail/bulletin publication, internet, signboard or poster at showroom, salesman, and introduced by others. We also have detailed information about what constitutes the main information sources in housing search for a household at five steps as: the stage for 1)consideration of purchase, 2)gathering information, 3)starting to visit showroom, 4)deciding a purchase, and 5)contracting. Note that the main information sources used for a household are different from step to step, and this requires to group households in terms of similar patterns or segments. For this, the partitioning clustering method "Daisy" (Kaufman and Rousseeuw, 1990) was employed. The result indicates that four segments are appropriate for the task as follow: internet/poster, housing magazine, leaflet, introduced by others as shown in Table 2.

Media environment

In addition, by investigating the media environment that customer accesses, we could figure out the characteristics of purchaser. Taking newspaper for example, there could be different customer layers generated depending on the type of newspaper that they read. For instance, compared the readers of economic newspapers and daily-sports, there could appear differences in interest and behavior between them. Thus, we divided purchasers into 3

categories of “high, medium, or low” based on the way they access the media environment of internet, fax, and housing magazine.

Customer satisfaction

Generally purchasers are quite sensitive to the level of services provided, it is known that a customer satisfied with good services may purchase a more expensive condominium. In this manner, services provided could affect the purchase activity of a customer. The services provided could be defined in many spectrums, but in this article, nine categories of satisfaction were provided in the questionnaire survey: the satisfaction of 1) explanation about target condominium, 2) explanation about surrounding environment, 3) explanation about financial planning, 4) explanation about purchasing procedures, 5) explanation about corporate management of the estate/after services, 6) response to questions, 7) politeness of reception, 8) speed of reception, and lastly 9) confidence of reception. Based on the degree of satisfaction of these criteria, we divided purchasers into three categories of “satisfied”, “in-between”, or “unsatisfied”.

Decisive factors or resigned factors in purchasing

These variables contain the items that were the decisive or resigned factors when a customer makes the ultimate decision in purchasing his/her condominium. The composition of questionnaire is the same as “Concerned items as important factor” described above. We believe that this kind of information is important, since it provides the insight into the change of psychology of the customer. That is what item(s) were actually more important or unimportant at the final stage compared to his/her concerned items at the beginning stage. The relationship between decisive and resigned factors of purchasing is to some extent trade-off. The most frequent decisive factors in purchasing are “location”(36%), “accessibility”(23%) followed by “price”(21%) as compared to “design”(45%), “price”(14%) and “accessibility”(12%) as resigned factors.

4.3 Extracted significant cells

In this subsection, the significant cells in a contingency table are extracted through all combination of variables using the proposed method in section 2. First, a combination table shown in Figure 1(b) should be designed, based upon how to designate attribute variables as target variables and the others as non-target variables. Again in this article, the term “target variables” indicates a group of variables on which a researcher focuses as his/her main interest. In this work, we set a scenario; condominium variables as target variables, and the remainder demographics variables and behavioral variables as non-target variables. In this case, the number of contingency table combinations by a set of target variables (8 attribute variables) and a set of non-target variables (the rest 19 attribute variables) consists of 152 pairs (8 x 19), where the total number of cells is 3,648. With this setting, we applied the

method presented in section 2. The resulting number of extracted significant cells is 240, at 99% confidence level set for suppressing emergence of too many pairs, which amounts to 6.6% of total cells. Figure 3(a) illustrates distribution of significant cells in this framework.

The prominent sets of combinations in the distribution are as follows. Firstly for the set of target variables, “price”(C1) and “ownership space”(C5) are high, and for the set of non-target variables, “number of tenant”(D3), “family type”(D4), “annual income of HH”(D11), and “decisive factor of purchasing”(B7) are high. In particular, as seen in Figure 3(a), there exist pairs like “price(C1) - annual income of HH(D11)”, “ownership space(C5) - annual income of HH(D11)” and “ownership space(C5) - number of tenant(D3)” that show prominent frequencies. This implies that in these pairs there are more exceptional patterns. Again note that the significant cells were extracted on the basis of two optional cases, namely either too much or too less frequency in a cell compared to its theoretical expectation.

After arranging the significant cells in terms of each level of target variables as a pivot, illustrated in Table 3(a), then we tried to summarize a series of exception rule patterns. The arranged set of significant cells is a collection of exception pairs, which are identified as characteristics of non-target variables related to target variables. Our resulting interpretation of them suggests that some of the combinations are quite contrary to our beliefs in the light of the domain knowledge. Some of these examples will be given in next subsection. Moreover, there seem to be repetitive pattern of apparently related pairs of significant cells.

4.4 Synthesis of significant cells

As mentioned in section 2, in order to grasp a comprehensive relation of the significant cells, we need to reorganize the significant cells or the fragmented information and apply association rule mining. In other words, relationship between the significant cells should be investigated systematically for consistent interpretation. In our framework, as in association rule mining, which is often labeled a basket analysis, transaction records of customers are analogous to each level of target variables, and items are correspondence to significant cells. Note that because the cells extracted from a contingency table are in one of the cases, either cell-over-expectation or cell-under-expectation, it is convenient to identify them in terms of unique code for each item. Any definition for the coding would be fine as long as the code is unique. For an intuitive example, the code “2C3↑” means that “2” as a variable ID, “3” as a level of the variable, and “↑” as the cell-over-expectation.

With the prepared codes exemplified in Table 3(b), we performed association rule mining and as a result, we obtained 1,819 association rules under the threshold support (5%) and confidence (100%), which are too many to grasp for meaningful information. The main reason that such a good number of rules emerge, despite 100% confidence, could be attributed to the fact that there are too many partially duplicated or nested rules. Thus, taking the nested rules away, we tried to summarize them by devising an intuitive but effective technique mainly aimed at our result from association rule.

The proposed technique is also based on a contingency table. An association rule is usually represented by X → Y with support (s) and confidence (c), where X and Y are a set of items. In our result of association rule, most of the items in both head part(X) and body part (Y) are single items, such as the form "(16C5H),(11C2H),(10.000% 100.00%)" which represents (head)(body)(support, confidence). Thus, assume that a contingency table composed of rows with a set of all unique items which occur in head part, columns with a set of all unique items in body part, then we can put a rule from association rule into the specified cell in accordance to a pair which can be decomposed into head (row) and body (column). Once built a contingency table, each frequency in the table can be interpreted as an integrated measure of support and confidence, where rows represent items of head part and columns represent items of body part. Considering the same example: "(16C5↑),(11C2↑)", by allocating the head (16C5↑) representing "Residential area:60s ㎡↑" in a row and the body (11C2↑) representing "Age of HH:「MALE」↓" in a column, we can easily count the frequency in the cell where two items crossed. Nevertheless, it should be noted that in this technique there can be three possible alternatives depending on the number of items in a head or a body; i.e. ①(r1) → (c1), ②(r1, r2,…,ri) → (c1), ③(r1, r2,…,ri) → (c1,c2,…,cj), where *ri*,*cj* items in row(head), items in column(body). Except alternative ① above, the other two alternatives could be unreasonable to be applied for the proposed technique because if there are more than two items in either head or body, it is unnatural to decompose items in the same manner.

Nevertheless, again for our task to outline result of association rule, the proposed technique performs successfully since most items are in the case of alternative ①, even though there are some cases that items fall into the alternative ② or ③. The resulted contingency table, as a summary of association rule, is illustrated in Figure 3(b). Due to the limitations of space, we omit the detailed result, but the result of association rule is sufficient to give clue to a good understanding of the relation between significant cells and it has closed to perfect consistency and easily understandable. Some of the association rule results are displayed in Table 4. The rule samples shown are in higher rank, and there are many pairs where the head of household is female. Note the case of line 2 in the table, "Sex of HH:「MALE」↓→ Sex of HH:「FEMALE」↑", which represents low probability of male household or high probability of a female household, namely a female household, a typical example of trade-off relation. Tracing these pairs, we can depict the characteristics of the female household as exception rules like this:

> *They have low income of average annual salaries in the range of 4~6 million Yen, and probably single families. Therefore, they are looking for a relatively small residential unit with one or two bedrooms. For them, "price" and "building facilities" are prone to be decisive factors of purchase.*

In the similar fashion, we present some other interesting rules found are followed:

- *If access time to the nearest railway station to be over 10 minutes, then "accessibility" as a resigned factors become increased, which implies that potential accessibility in mind would be around 10 minutes on walk or within about 350 meters from railway station.*
- *If occupancy space is around 60~70 m^2 or 2~3 bed rooms which is typical and popular, then the prominent items seem to be "investment" as a motivation of purchase and "price" as a decisive factor. Moreover, in this case, "introduction and others" as information sources for searching housing is extremely scarce.*
- *Households with average annual salary of 15~20 million Yen are targeted to high-rise building especially over 16 floors, and in this case "building facilities" is prone to be a decisive factor of purchase. In particular, the possibility that their occupations to be "financial" is very high.*

Again, note that these are exception rules, though it seems general or strong patterns, these rules are surpassing or falling short of theoretical expectation. Besides this analysis with condominium variables as target variables, we also carried out the same method to behavior variables as target variables, which also provides quite informative results relating to behavioral characteristics of purchasers explained by the demographics and condominium variables.

4.5 A validity of the proposed method

In this subsection we try to investigate the validity of the proposed method. Because it is not easy to find any methods oriented to our problem raised in section 1. Therefore, as a substitute, we compare some results produced by the proposed method to those from the histogram of the descriptive statistics. For simplicity, we take some extracted significant cells in Table 4 and some of their histograms in Figure 4, and in both cases, single family type is set as a criterion.

To begin with, distribution of variable "Annual income of HH" in Figure 4(a) is quite consistent with the first 5 lines in Table 4, where annual income of about 70 % of single families is fewer than 8 million Yen. The distribution of "Occupation of HH" in Figure 4(d) is also quite intelligible, because the occupation of "MANAGER" is quite rare in the case of single family. But, note that in the case of "Decisive factor of purchasing" in Figure 4(b) and "Media environment to access" in Figure 4(c), distribution of each variable disagrees to the rules in the Table 4, because "LOCATION" as decisive factor of purchasing is not so rare case and "LOW" as media environment to access is not frequent case, which contradicts the items in table 4. After all, as shown in this example, it could be pointed out that the structure of exception rules/patterns is not self-evident under statistical distributions, and in some cases it contradicts the distributions as shown in Figure 4 because these distributions are strong patterns.

5. CONCLUSION

In this work we have presented a rather straightforward method of extracting exception rules from a dataset with large attributes and applied the method in order to describe condominium purchasers' behavior. The method consists of two major steps: first, we focus on the contingency table showing the responses of subjects to one variable as a function of another variable as "association measure". We extract the cell(s) in the given statistical significance level, and summarize them on the basis of target variables. At this time the subjective evaluation of "exception" can be interpreted objectively by the significance level. Second, because these extracted cells or significant cells are fragment information, they should be synthesized systematically for consistency. For this, we utilize association rule mining in order to synthesize the fragment information as "pattern bundle in a basket". The rule patterns and their relationship by the proposed method are excellent in extracting and interpreting of exception rules because the rules are not only easily understandable but also consistent, and therefore it is expected to have high potential of application for a dataset with massive attributes.

For future work, different thresholds on association rule for significant cells can be experimented and a variety of results can be compared. For the reliability and validity of the proposed method, datasets with different characteristics and sizes can be evaluated and compared to the domain knowledge of marketers.

REFERENCES

Agrawal, Rakesh and Ramakrishnan Srikant (1994), "Fast algorithms for mining association rules," *In Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile.

Bounsaythip, Catherine and Esa Rinta-Runsala (2001), "Overview of Data Mining for Customer Behavior Modeling," *VTT Information Technology Research Report* TTE1-2001-18.

Dunning, Ted (1993), "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, 19(1), 61-74.

Everitt, BS (1992), *The analysis of contingency tables*. 2d ed. London:Chapman and Hall.

Fabris, Carem C. and Alex A. Freitas (1999), "Discovering surprising patterns by detecting occurrences of Simpson's paradox," *Research and Development in Intelligent Systems XVI*, 148-160.

Freitas, Alex A. (1999), "On Rule Interestingness Measures," *Knowledge-Based Systems*, Vol.12, 309-315

Kaufman, Leonard and Peter J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. New York:Wiley.

Liu, Huan, Hongjun Lu, Ling Feng and Farhad Hussain (1999), "Efficient search of reliable

exceptions," *PAKDD(Pacific-Asia Conference on Knowledge Discovery and Data Mining)*, 194-203.

Ocerin, J.M. Caridad, R. Espejo Mohedano, A. Gallego Segador, (1999), "Automatic aggregation of categories in multivariate contingency tables using information theory," *Computational Statistics & Data Analysis*, 29, 285-294.

Padmanabhan, Balaji and Alexander Tuzhilin (1999), "Unexpectedness as a measure of interestingness in knowledge discovery," *Decision Support Systems*, 27(3), 303-318.

Silberschatz, Abraham and Alexander, Tuzhilin (1996), "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Transactions of Knowledge and Data Engineering*, 8(6), 970-974.

Wedel, Michel and Wagner A. Kamakura (1998), *Market segmentation: Concepts and methodological foundations*. Boston: Kluwer Academic Publishers.

Worzala, Elaine, Lenk, Margarita and Silva, Ana (1995), "An Exploration of Neural Networks and Its *Application to Real Estate Valuation*," Journal of Real Estate Research, Vol.10, No.2, 185-201.

Table 1

SELECTED VARIABLES

Demographics Variables	Behavior Variables	Condominium Variables
D1. Type of address change(before/after) D2. Experience of residence purchase D3. Number of tenant D4. Family type of household D5. Means of transportation for work of HH D6. Age of HH D7. Sex of HH D8. Type of employment for HH D9. Occupation of HH D10. Category of business for HH D11. Annual income of HH	B1. Motivation of purchasing B2. Important factor of purchasing B3. Pattern of gathering information B4. Main mass media of purchasing B5. Media environment to access B6. Customer satisfaction B7. Decisive factor of purchasing B8. Resigned factor of purchasing	C1. Price C2. Time to walk from nearest railway station C3. Ward C4. Unit type C5. Ownership space C6. Number of room C7. Total number of floors C8. Total number of houses

※ HH : Head of a household

Table 2

LEVELS OF CATEGORY VARIABLES

Variables (Unit)	levels							
	1	2	3	4	5	6	7	8
D1	within a ward 60.5%(483)	outside a ward 39.5%(315)						
D2	for the first time 77.8%(620)	replace 15.1%(120)	increase 7.2%(57)					
D3 (Number)	1 0.5%(4)	2 21.8%(174)	3 35.5%(283)	4 23.4%(187)	5 16.2%(129)	6 2.4%(19)	7 0.3%(2)	
D4	single 21.6%(171)	couple 32.4%(257)	couple＋child 40.7%(322)	others 5.3%(42)				
D5	railway 85.7%(663)	car 4.8%(37)	bus 2.2%(17)	walk/cycle 7.4%(57)				
D6	20s 9.9%(79)	30s 54.4%(433)	40s 25.5%(203)	50s 7.5%(60)	over 60s 2.6%(21)			
D7	male 79.4%(632)	female 20.6%(164)						
D8	businessman 90.2%(715)	self-employed /specialist 6.6%(52)	student /others 3.3%(26)					
D9	business 24%(182)	service 22%(167)	engineer 16.4%(124)	specialist 15.6%(118)	manager 22%(167)			
D10	construction 7.5%(57)	manufacture 18%(137)	finance 13.7%(104)	broadcast 9.1%(69)	commerce 15.4%(117)	IT 11.6%(88)	others 24.8%(189)	
D11 (Million Yen)	～4 3.5%(28)	～6 23.2%(184)	～8 25.3%(201)	～10 18.1%(144)	～12 14.7%(117)	～15 9.8%(78)	～20 3.3%(26)	20～ 2%(16)
B1	investment 40.1%(319)	ownership 20.2%(161)	larger house 30.3%(241)	others 9.4%(75)				
B2	price independent 24.4%(195)	price dependent 75.6%(603)						
B3	standard 35.3%(269)	double (long-term) 30.7%(234)	double (short-term) 22%(168)	impulse 11.9%(91)				
B4	internet /poster 12%(96)	housing magazine 36.2%(289)	newspaper leaflet 43.6%(348)	introduction 8.1%(65)				
B5	high(internet・fax・JJ) 23.8%(190)	low (no internet) 12.7%(101)	medium 63.5%(507)					
B6	satisfied 38.8%(309)	in-between 53.1%(423)	unsatisfied 8.2%(65)					
B7	price 21.4%(169)	accessibility 23.2%(183)	area 37.1%(293)	design 11.5%(91)"	building facility 6.8%(54)			
B8	price 14.7%(108)	accessibility 13.1%(96)	area 13.5%(99)	design 49.2%(361)	building facility 9.5%(70)			
C1 (Million Yen)	～20 0.8%(6)	～30 4.9%(39)	～40 19.9%(157)	～50 29.9%(236)	～60 23.6%(186)	～70 11.3%(89)	～80 5.5%(43)	80～ 4.1%(32)
C2 (Minutes)	～5 28.1%(196)	～10 40%(279)	～15 26.5%(185)	～20 4.7%(33)	20～ 0.6%(4)			
C3	Shibuya 18.8%(150)	Setagaya 25.9%(207)	Ota 27.4%(219)	Shinagawa 18.2%(145)	Meguro 9.6%(77)			
C4	1R 13.7%(109)	2R 32.2%(256)	3R 48.2%(383)	4R+ 5.8%(46)				
C5 (㎡)	～30 1.8%(14)	～40 2.9%(23)	～50 4%(31)	～60 11.6%(91)	～70 23%(180)	～80 34.2%(268)	～90 14.9%(117)	90～ 7.7%(60)
C6	R1 10.7%(85)	R2 25.2%(200)	R3 55%(437)	R4+ 9.1%(72)				
C7	～3 F 8.8%(70)	～5 F 18.1%(144)	～10 F 53%(421)	～15 F 17.3%(137)	～20 F 1.4%(11)	～30 F 1.4%(11)		
C8	～30 23.6%(188)	～50 26.1%(208)	～100 29.8%(238)	～200 15%(120)	200～ 5.5%(44)			

※ JJ: monthly housing magazine for rent and purchase, D1~11 : Demographics, B1~7 : Behavior, C1~7 : Condominium

Table 3

SAMPLE OF ARRANGEMENT OF EXTRACTED SIGNIFICANT CELLS

(a) Sample of Actual Significant Cells

C1-1：Price:「Under 2 million Yen」 ----------

Experience of residence purchase:「INCREASE」 8.284

Family type of household:「SINGLE」 3.049

Annual income of HH:「Over 20 million Yen」 31.4

Decisive factor of purchasing:「BUILDING FACILITY」 8.83

Decisive factor of purchasing:「PRICE」 3.077

C1-2：Price:「2~3 million Yen」 ----------

…

※ The figure indicates the value of logarithm of ratio that actual frequency divided by expectation probability.

(b) Sample of Coding for Association Rule Mining

C1-1 : 2C3↑, 4C1↑, 11C8↑, 18C5↑, 18C1↑

C1-2 : 3C2↑, 4C1↑, 5C3↑, 7C2↑, 7C1↓, 10C1↑, 11C1↑, 11C2↑, 18C1↑

C1-3 : 3C2↑, 4C1↑, 5C4↑, 6C1↑, 7C2↑, 7C1↓, 9C5↓, 9C3↑, 10C3↓, 11C5↓, 11C6↓, …

C1-4 : 6C3↓, 8C2↓, 11C6↓, 11C3↑

C1-5 : 3C2↓, 4C1↓, 7C2↓, 11C2↓, 11C5↑, 11C4↑, 18C1↓, 18C5↑

…

※ *nCk* : *n* stands for a variable, *k* stands for the level of *k*.
↑: cell-over-expectation,
↓: cell-under-expectation.

Table 4

AN EXAMPLE OF RESULT FROM ASSOCIATION RULE MINING

(In the case of Family type:「SINGLE」↑)

Annual income of HH:「Under 4 Million Yen」 ↑
Annual income of HH:「4～6 Million Yen」 ↑
Annual income of HH:「8～10 Million Yen」 ↓
Annual income of HH:「10～12 Million Yen」 ↓
Annual income of HH:「12～15 Million Yen」 ↓
Sex of HH:「MALE」 ↓
Sex of HH:「FEMALE」 ↑
Number of tenant:「1」 ↑
Number of tenant:「3」 ↓
Number of tenant:「4」 ↓
Occupation of HH:「MANAGER」 ↓
Family type:「COUPLE + CHILD」 ↓
Decisive factor of purchasing:「PRICE」 ↑
Decisive factor of purchasing:「AREA」 ↓
Media environment to access:「LOW」 ↑

※ ↑ : cell-over-expectation, ↓ : cell-under-expectation

Figure 1

CONTINGENCY TABLES IN THE PROPOSED METHOD

(a) Example of a Dataset

Target variable		Non-target variable	
T1	T2	N1	N2
1	3	3	4
4	2	1	2
3	4	1	2
2	5	2	1
3	1	2	2
2	3	4	3

(b) Combination Table

		Non-target Vars.			
		N1	N2	…	Nn
Target Vars.	T1				
	T2				
	…				
	Tt				

(c) Contingency Table

	C_1	C_2	…	C_c	sum
R_1	n_{11}	n_{12}	…	n_{1c}	n_{1+}
R_2	n_{21}	n_{22}	…	n_{2c}	n_{2+}
…	…	…	…	…	…
R_r	n_{r1}	n_{r2}	…	n_{rc}	n_{r+}
sum	n_{+1}	n_{+2}	…	n_{+c}	n

Figure 2

EXCEPTION RULE MINING FRAMEWORK

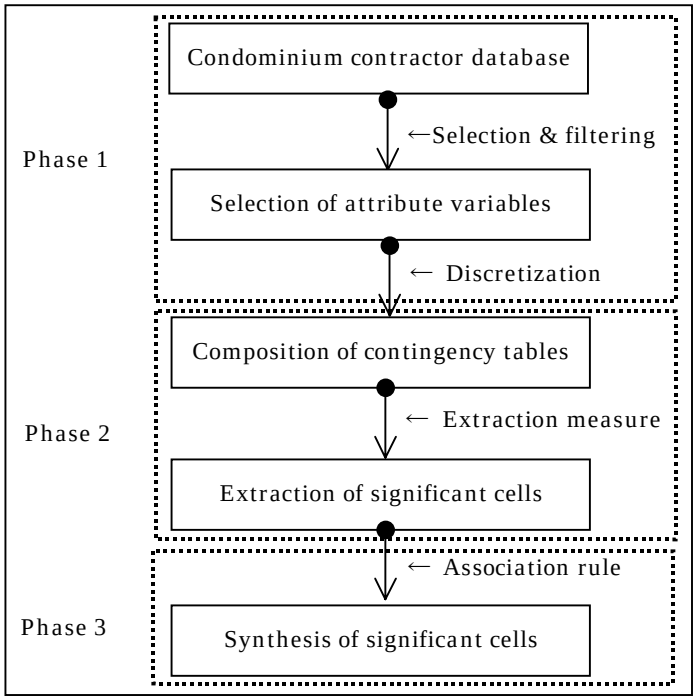
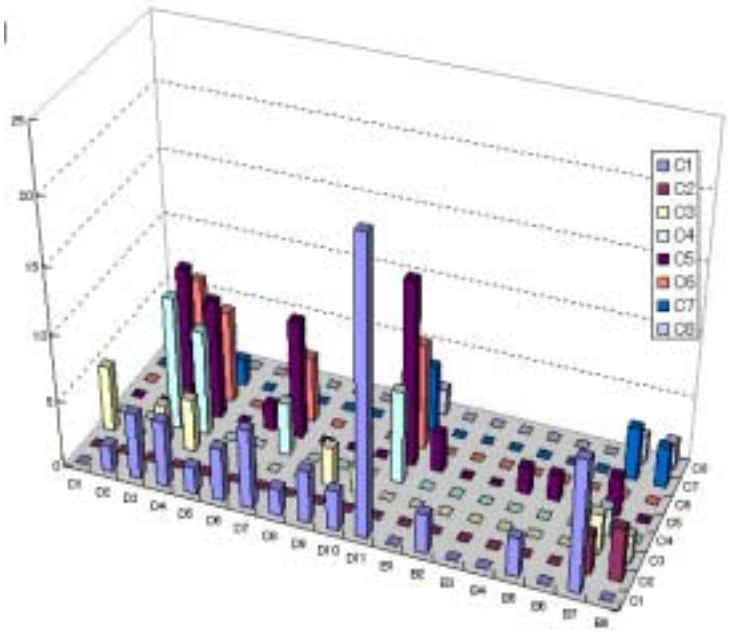


Figure 3

EXTRACTED SIGNIFICANT CELLS

(a) Distribution of Significant Cells (Target Variables : Condominium Variables)



(b) Relation Table Among Significant Cells By Association Rule

	1YC2H	1YC3L	1YD8H	3C3H	3C2L	3C3L	3D5H	3D8H	4C3H	4C1L	4D3H	7C1L	7C2H	7C2L
1YC1H	16	0	0	35	0	0	0	0	35	0	0	16	35	0
1YC2H	0	0	0	61	0	0	0	0	61	0	0	60	61	0
1YC3L	0	0	0	0	10	0	0	0	0	10	1	0	0	13
1YD4L	40	0	0	40	0	0	0	0	40	0	0	40	40	0
1YD5H	0	20	0	0	20	0	0	0	0	20	0	0	0	20
1YD8H	0	5	0	0	3	0	14	0	0	3	13	0	0	0
1YD9H	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1BC2H	40	0	0	40	0	0	0	0	40	0	0	40	40	0
1BC3H	34	0	0	33	0	0	0	0	34	0	0	33	30	0
1BD3H	0	0	0	0	1	0	0	0	0	1	0	0	0	0
3C2H	64	0	0	0	0	0	0	0	70	0	0	53	74	0
3C3L	0	19	0	0	0	0	0	0	0	27	0	0	0	19
3D3L	0	0	16	0	0	0	16	16	0	0	16	0	0	0
3D5H	0	0	14	0	0	0	0	0	0	0	14	0	0	0
3D8H	0	0	29	0	0	29	29	0	0	0	29	0	0	0
4C3H	60	0	0	74	0	0	0	0	0	0	0	49	70	0
4C1L	0	19	0	0	27	0	0	0	0	0	0	0	0	19
4D3L	0	0	0	0	0	1	0	0	0	0	0	0	0	0
4D3H	0	16	13	0	13	0	14	0	0	13	0	0	0	13
4D3L	0	0	0	14	0	0	0	0	14	0	0	0	0	0
7C1H	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7C1L	84	0	0	84	0	0	0	0	84	0	0	0	84	0
7C2H	69	0	0	82	0	0	0	0	82	0	0	53	0	0
7C2L	0	16	0	0	25	0	0	0	0	25	0	0	0	0

※ Row : Head, Column : Body

Figure 4

HISTOGRAM OF ATTRIBUTE VARIABLES IN THE CASE OF FAMILY TYPE : SINGLE

Appendix

Table A-1

A SAMPLE OF ASSOCIATION RULES ACROSS EXTRACTED SIGNIFICANT CELLS

```
Number of frequency : 84
      Sex of HH: 「MALE」 ↓      ⇒   Annual income of HH: 「400～600 million Yen」 ↑
      Sex of HH: 「MALE」 ↓      ⇒   Sex of HH: 「FEMALE」 ↑
      Sex of HH: 「MALE」 ↓      ⇒   Number of tenant: 「1」 ↑
      Sex of HH: 「MALE」 ↓      ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 82
      Sex of HH: 「FEMALE」 ↑    ⇒   Number of tenant: 「1」 ↑
      Sex of HH: 「FEMALE」 ↑    ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 81
      Annual income of HH: 「400～600 million Yen」 ↑⇒   Sex of HH: 「FEMALE」 ↑
      Annual income of HH: 「400～600 million Yen」 ↑⇒   Number of tenant: 「1」 ↑
      Annual income of HH: 「400～600 million Yen」 ↑⇒   Family type: 「SINGLE」 ↑
Number of frequency : 79
      Number of tenant: 「1」 ↑  ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 74
      Number of tenant: 「1」 ↑  ⇒   Sex of HH: 「FEMALE」 ↑
      Family type: 「SINGLE」 ↑  ⇒   Number of tenant: 「1」 ↑
Number of frequency : 70
      Family type: 「SINGLE」 ↑  ⇒   Sex of HH: 「FEMALE」 ↑
Number of frequency : 69
      Sex of HH: 「FEMALE」 ↑    ⇒   Annual income of HH: 「400～600 million Yen」 ↑
Number of frequency : 64
      Number of tenant: 「1」 ↑  ⇒   Annual income of HH: 「400～600 million Yen」 ↑
Number of frequency : 60
      Family type: 「SINGLE」 ↑  ⇒   Annual income of HH: 「400～600 million Yen」 ↑
      Annual income of HH: 「400～600 million Yen」 ↑⇒   Sex of HH: 「MALE」 ↓
Number of frequency : 53
      Number of tenant: 「1」 ↑  ⇒   Sex of HH: 「MALE」 ↓
      Sex of HH: 「FEMALE」 ↑    ⇒   Sex of HH: 「MALE」 ↓
Number of frequency : 49
      Family type: 「SINGLE」 ↑  ⇒   Sex of HH: 「MALE」 ↓
Number of frequency : 43
      Annual income of HH: 「800～1,000 million Yen」 ↓    ⇒Annual income of HH: 「400～600 million Yen」 ↑
      Annual income of HH: 「800～1,000 million Yen」 ↓    ⇒   Sex of HH: 「MALE」 ↓
      Annual income of HH: 「800～1,000 million Yen」 ↓    ⇒   Sex of HH: 「FEMALE」 ↑
      Annual income of HH: 「800～1,000 million Yen」 ↓    ⇒   Number of tenant: 「1」 ↑
      Annual income of HH: 「800～1,000 million Yen」 ↓    ⇒   Family type: 「SINGLE」 ↑
      Ownership space: 「30㎡ ZONE」 ↑     ⇒   Annual income of HH: 「400～600 million Yen」 ↑
      Ownership space: 「30㎡ ZONE」 ↑     ⇒   Sex of HH: 「MALE」 ↓
      Ownership space: 「30㎡ ZONE」 ↑     ⇒   Sex of HH: 「FEMALE」 ↑
      Ownership space: 「30㎡ ZONE」 ↑     ⇒   Number of tenant: 「1」 ↑
      Ownership space: 「30㎡ ZONE」 ↑     ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 35
      Annual income of HH: 「400 under million Yen」 ↑     ⇒   Sex of HH: 「FEMALE」 ↑
      Annual income of HH: 「400 under million Yen」 ↑     ⇒   Number of tenant: 「1」 ↑
      Annual income of HH: 「400 under million Yen」 ↑     ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 34
      Total number of floor: 「3F Under」 ↑     ⇒   Annual income of HH: 「400～600 million Yen」 ↑
      Total number of floor: 「3F Under」 ↑     ⇒   Family type: 「SINGLE」 ↑
Number of frequency : 30
      Total number of floor: 「3F Under」 ↑     ⇒   Sex of HH: 「MALE」 ↓
      Total number of floor: 「3F Under」 ↑     ⇒   Sex of HH: 「FEMALE」 ↑
      Total number of floor: 「3F Under」 ↑     ⇒   Number of tenant: 「1」 ↑
Number of frequency : 27
      Number of tenant: 「1」 ↓  ⇒   Family type: 「SINGLE」 ↓
      Family type: 「SINGLE」 ↓  ⇒   Number of tenant: 「1」 ↓
…
```