Title	Applying Graph Mining to Discover Substructures of Room Layouts which Affect the Rent of Apartments
Author(s)	Takizawa, Atsushi; Yoshida, Kazuma; Katoh, Naoki
Citation	IEEE International Conference on on Systems, Man and Cybernetics, 2007 (2007): 3512-3518
Issue Date	2007-10
URL	http://hdl.handle.net/2433/84855
Right	Copyright(c) 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Type	Conference Paper
Textversion	publisher

Applying Graph Mining to Discover Substructures of Room Layouts which Affect the Rent of Apartments

Atsushi Takizawa, Kazuma Yoshida, *Nonmembers*, and Naoki Katoh, *Member*

***Abstract*—In this paper we will investigate the relationship between room layout and the rent of apartments by extracting meaningful substructures of a graph representing the room layout using a graph mining algorithm. We will then construct a prediction model for the rent of apartments with high accuracy. Through our analysis, we will reveal certain typical substructures in the room layout which strongly affect the rent.**

I. Introduction

Real-estate appraisal is important for trading properties, collateral evaluation, project evaluation, and so on. Recently in Japan, since securitization of real estate begins to intensify, the appraisal method for real estate with higher precision is needed. Hedonic approach is used as one of real-estate appraisal methods. In hedonic approach, the value of property or service is determined by its various attributes. Then the contribution of each attribute is statistically estimated by regression analysis.

In this paper we will present the price analysis of rental residences of apartment buildings. In a general hedonic approach for dwellings, the value (i.e. price or rent) of a residence is estimated based on attributes such as the occupied area, the distance from the nearest station, the building age, facilities and so on [1], [2]. However the information of the room layout has not been fully used in the conventional approach although it seems to be crucial to evaluate usability and fineness of space. Only the type of room layout or the area of the residence has been considered so far [3].

The room layout can be described by a graph, and it has often been used for spatial analysis in the field of architecture and urban planning [4]-[6]. Conventional statistical analysis usually deals with only numerical or categorical data, and has not directly dealt with the data of graph structure. Recent advance of data mining has been extending its applicable data structure from traditional tabular or transaction form data to structural data such as graph, semi-structured text and so on. Especially, data mining on graph structure is called graph mining, and several efficient algorithms have been developed [7]-[9]. Graph mining has been applied to the areas such as chemical, gene and web analysis. However, it has not been applied to the field of architecture. Moreover, the typical methods in previous studies on spatial analysis using graphs were the naive classification of space by graph topology and the characterization of space by several statistics of graphs. The enumeration and analysis of meaningful subgraphs has not been attempted yet.

With this background, using the data of family-oriented apartments in the suburb of Kyoto City, we will clarify the relationship between room layouts and rents by using graph mining algorithm and will construct a prediction model with high accuracy for the rent. We believe that this study contributes to both graph mining and spatial analysis in that it introduces a new spatial analysis method and adds a new application of graph mining. We believe that our study will help designers to design new apartments that increase the satisfaction level of customers.

The rest of this paper is organized as follows. In the next section the acquisition and details of data will be described. Section 3 will explain how we discover frequent subgraphs of room layout. In section 4 discriminative subgraphs strongly associated with rent or building age will be extracted by Emerging Patterns [10]. In section 5 three different regression models for rent prediction will be constructed; the first one uses only numerical and categorical data, the second uses only subgraph data, and the third combines both types of data. We will compare their accuracy and analyze the relationship between EPs and coefficients of the regression model. Section 6 concludes the paper.

II. Data

A. Target

Since in Japan the real estate database of residences which contains images of room layouts has not been available, previous studies [6] collected the real estate data from commercial information magazines which have only limited data because of the lack of space. Moreover, the information on the magazines is updated not more than twice per month. In contrast, web pages on real estate have detailed information which is updated almost every day. Furthermore, the data on the web can be downloaded and edited automatically. Consequently, in this study we use the data of CHINTAI web (http://www.chintai.net/) which is one of the most popular web sites on real estate especially rental apartments in Japan.

We limit the target of our research to the family-oriented residences of apartments. In Japan, dwelling-types are distinguished by the term such as 3LDK where “3” means the number of bed rooms, “L” a family room, “D” a dining room

Manuscript received April 15, 2007. This work was supported in part by the Grant-in-Aid for Scientific Research (C) (No.17500007) of Japan Society for the Promotion of Science.

Atsushi Takizawa and Naoki Katoh are with Department of Architecture and Architectural Engineering, Graduate School of Engineering, Kyoto University, Kyoto University Katsura Campus, Nishikyo-ku, Kyoto, 615-8540, Japan (e-mail: {kukure, naoki}@archi.kyoto-u.ac.jp). Kazuma Yoshida is with XYMAX AXIS Corporation, Umeda 1-11-4, Kita-ku, Osaka, 530-0001, Japan.

TABLE I
LIST OF NUMERICAL ATTRIBUTES

Attribute	Mean	Standard Deviation
Rent	81,770 yen	14,866
Building age	16.5 years	7.0
Occupied area	60.4 m^2	7.4
Area of the largest Japanese room	6.0 jou*	0.7
Area of the middle Japanese room	3.2 jou	2.7
Area of the smallest Japanese room	0.1 jou	0.7
Area of the largest Western room	5.7 jou	1.2
Area of the middle Western room	2.2 jou	2.7
Area of the smallest Western room	0.1 jou	0.5
Area of a dining room	9.6 jou	2.7
Floor of the residence	2.8	1.6
Number of stories of the building	4.8	2.2
Walking time to the nearest station	13.1 minutes	6.9
Distance to a super market	501 m	286
Distance to a convenience store	418 m	254
Distance to a hospital	258 m	216
Distance to police station	657 m	288
Distance to a kindergarten	417 m	205
Distance to an elementary school	504 m	255
Distance to a junior high school	753 m	377
Distance to a high school	1,242 m	608

* jou means a unit of area which corresponds to one sheet of tatami.

and "K" a kitchen. The 3LDK dwelling-type is most popular among Japanese condominiums. 2LDK dwellings are also popular, but the complexity of the room layout is not rich enough for our analysis considering space structure. On the other hand the number of 4LDK residences is much less than that of 3DLK residences. From this reason we will concentrate on 3LDK (including 3K and 3DK) residences in this paper. In addition residences which have lofts or two-story floors are excluded because these three-dimensional room layouts qualitatively much differ from single-story room layouts.

We choose the target area as that along the railways of Hankyu Kyoto Line and JR Tokaido Line located in the south west of Kyoto City. These areas have common features as dormitory suburb of Kyoto and Osaka. Among them 15 areas are picked up and classified according to the nearest station of these railways.

B. Data Acquisition and Cleaning

From the beginning of October to the end of November of 2006, we downloaded html source codes of the information page of residences on CHINTAI web page by using a free web downloader. However, image files of room layouts could not be downloaded by this software because of the cgi barrier, we created an image downloader to download image. Necessary data are extracted from html source codes by Perl script and we have created 996 data of residences as "tabular form data". Room layout images are abstracted into edge labeled graphs by hand, and then "room layout graph data" is constructed.

Below, we will explain the details of the tabular form data and the graph data of room layouts.

C. Tabular Form Data

Table I shows 21 numerical attributes, and Table II shows

TABLE II
LIST OF CATEGORICAL ATTRIBUTES

Attribute	Value (# of data)
Room layout type	3LDK(707), 3DK(285), 3K(4)
Residence type	Apartment(914), corporative(5), tenement(4), others(73),
Structure type	Reinforced concrete (821), steel framed reinforced concrete (28), steel-frame (68), light gauge steel(59), timber structure (17), others(3)
Orientation of main opening	East (361), west(84), south(450), north(9), south east (31),south west (47), north east (13), north west (1)
Nearest station	8 stations of Hankyu Kyoto line (8--220), 2 stations of Hankyu Arashiyama line (46 and 113), 4 stations of JR Tokaido line (9--98), 1 station of Keifuku Arashiyama line (16).
Facility	45 attributes; e.g. parking, air conditioner, etc.

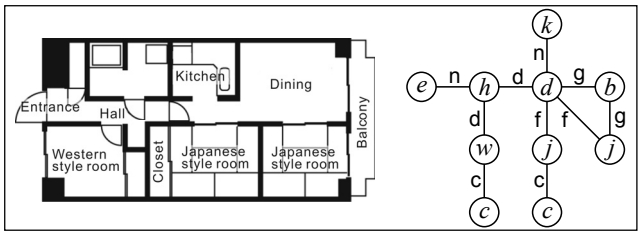

Fig. 1. Room layout and the corresponding graph.

categorical ones. Rent does not include utility fee. Facility attributes in Table II are limited to ones which correspond to more than five residences among all attributes on the website. Room area is set to be zero if there are no relevant ones in the property.

D. Room Layout Graph Data

As shown in fig.1, a room layout is abstracted to a graph structure which represents room types and adjacency relationships between rooms. A vertex indicates a room and an edge represents adjacency relationship. We call this abstract data as room layout graph data which is a simple undirected graph whose edge has a label which stand for denoting the type of partition between two rooms.

Next we will explain the details of vertex and edge types.

1) Vertex Types

Considering the general composition of Japanese residences of apartments, the following eight kinds of vertices are defined.

(*e*) : an entrance vertex. It can be found in all room layouts. For example, an entrance vertex is defined even if an entrance is directly connected to a dining room. The detail will be explained next.

(*h*): a hall vertex which is defined as the space which is between an entrance and a dining room. A hall vertex is defined if it is partitioned by a door or a fusuma (fusuma is a Japanese sliding door).

(*d*) : a dining room vertex. In modern Japanese residences, a lot of dining rooms contain a kitchen and a living room. They are usually distinguished by the term such as 3K, 3DK or 3LDK. However in this paper we do not distinguish them because this room might be a main space of the residence in daily use.

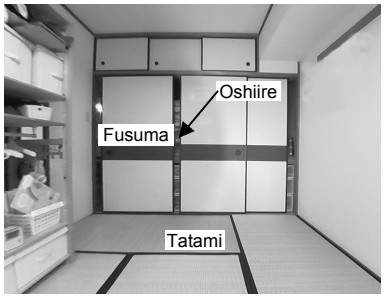

Fig. 2. An example of a Japanese style room in an apartment.

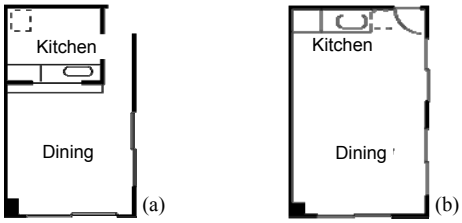

Fig. 3. Distinction of kitchen vertices; (a) has a kitchen vertex and (b) does not have it.

(j) : a Japanese room vertex which is a traditional room with tatami, fusuma, oshiire and other traditional Japanese fittings as shown in fig.2. Tatami is a Japanese floor mat which is made of rush. Fusuma is a sliding door. Oshiire is a Japanese closet whose opening is fusuma.

(w) : a western room vertex. The term "western" is contrasted with "Japanese". A western room has flooring or carpet, door and curtain.

(b) : a balcony vertex. Balcony is not usually large in Japan. The main function of the balcony is to hung out the laundry. If the room is on the first floor and has its own garden outside, we regard the garden as a balcony, too.

(c) : a closet vertex. Including oshiire, any kind of closet is defined by this vertex. If there are more than one closet in a room, they are combined into a single closet vertex.

(k) : a kitchen vertex which is defined only if the kitchen is separated from the dining room as shown in fig.3. This is an important distinction.

2) Edge Types

As mentioned above, we distinguish the connection between rooms by the following five kinds of edges. d is a door edge. A hinged door corresponds to it. f is a fusuma edge. A fusuma edge is defined if the border of the rooms is partitioned by a sliding door. c is a closet edge which corresponds to a door of a closet. The difference of the door type such as hinged or sliding door is not distinguished. g is a glass edge. It is used for the border of an internal room and an outside balcony. n is a no-partition edge. This edge has a special meaning that it is defined for distinguishing two rooms where there is no physical partition such as an entrance and a hall.

The average number of vertices and edges is 10, the maximum number of vertices is 13, and the maximum number of edges is 14 in our data.

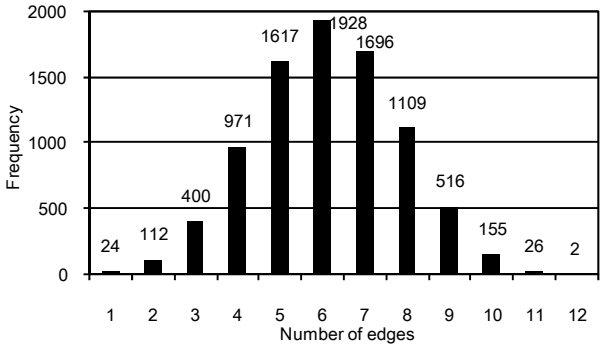

Fig. 4. Histogram of subgraphs in terms of the number of edges.

III. Extracting Frequent Subgraphs

Frequent subgraphs are extracted by using a graph mining tool, and those "subgraph data" is constructed and analyzed. Subgraph data is then transformed to a 0-1 categorical one which indicates the non-existence or the existence of the subgraph in the room layout.

We adopt FSG [8] as graph mining algorithm. FSG is one of apriori based graph mining algorithms such as AGM [7] and gSPAN [9]. The reason why we adopt FSG among them is simply that FSG can output not only frequent subgraphs but additional information such as the parent-children list. This information is useful for our analysis.

The minimum support of a subgraph is set to be 0.5%. This corresponds to 5 residences in our data set consisting of 996 residences. This seems to be very small. However, considering the diversity of room layouts, only general patterns with large frequency do not seem to be enough for evaluating room layouts. Rarer patterns with small frequency should be considered. As a result of graph mining, 8,556 subgraphs were extracted. Fig.4 shows the histogram of subgraphs in terms of the number of edges. It looks like a standard distribution whose peak is six.

IV. Discovering Primary Subgraphs by Emerging Patterns

A. Emerging Pattern Analysis

In this section, the influence of subgraphs on rent is examined. In addition, the difference of room layouts according to their building age is also examined. For this purpose, we will use Emerging Pattern (EP) [10], which is a kind of association rule for finding significant subgraphs from among a lot of subgraphs extracted by FSG.

EP is defined as an itemset whose support increases significantly from one dataset to another, and can find the itemset which are typical to the class concerned but not to the other class. The degree of significance of EP is measured by the value called the growth-rate (gr). The detail of EP is given in the appendix. Many subgraphs for which in parent-child relationship holds tend to have the same gr.

B. EPs for the Rent

At first we will examine the EP for the rent. According to the rent, we divide the whole data into two classes H and L

such that class H (resp. L) are the subsets of data whose rent fall into the upper (resp. lower) half among those whose nearest railway station is the same.

As the first step, EPs of simple subgraphs whose number of edges is one or two are extracted in order to examine the basic influence of subgraphs on rent. The sizes of classes H and L are 490 and 506, respectively.

1) EPs for the Classes H and L

Top three EPs on classes H and L are shown in fig.5. We select EPs having at least 10 residences which occupy about 1% of the whole data in order to keep the minimum amount of generality. Let us look at three subgraphs having one edge which reads that the graphs 1-22, 1-16 and 1-12 respectively indicate a separate kitchen, a dining room facing to a balcony, and the existence of a closet in the western style room. The subgraphs 2-71, 2-66 and 2-57 having two edges in the figure respectively indicate two western style rooms connected to a balcony, a dining room with a separate kitchen that faces a balcony, and a dining room with a separate kitchen connected to a western style room. Notice that the last two graphs 2-66 and 2-57 contain as their subgraphs the EPs having one edge shown in the upper part of the figure. In addition, we observe that as the number of edges increases, gr tends to become larger.

Then, let us look at fig.6 which shows EPs for the lower half (class L). The graphs 1-21, 1-23 and 1-0 which consist of a single edge respectively indicate a dining room directly connected to a hall without any partition, a dining room connected to an entrance, and a dining room connected to a Japanese style room through a door (1-0). The graphs 2-37, 2-48 and 2-103 which consist of two edges respectively indicate a dining room connected to a hall through a fusuma and to a Japanese style room through a door , a dining room connected to an entrance without any partition and to a western style room through a door (2-48), and a dining room connected to a hall without any partition and to a Japanese style room through a fusuma. Each of these EPs having two edges contains as its subgraphs at lease one of EPs shown in the same figure consisting of a single edge. Grs of EPs for the lower half (class L) are larger than those for class H. This might imply that the existence of a small portion of room layout which has a crucial negative influence on the rent.

2) EPs for the Uppermost Quarter and the Lowermost Quarter

In the former analysis, residences whose rent is close to the average may obscure the influence of subgraphs on rent. Thus, in order to see more clearly the influence of subgraphs on rent, we select the upper half of the class H (denoted by class HH) and the lower half of the class L (denoted by class LL). The numbers of residences in the classes HH and LL are 206 and 187, respectively. Since the sizes of classes HH and LL are small and simple subgraphs with one or two edges have small gr, EPs are sought without limiting the number of edges involved. When we consider the class HH (resp. LL) in the analysis, the other class we consider consists of residences other than HH (resp. LL) (which is denoted by $\overline{\text{HH}}$ (resp. $\overline{\text{LL}}$).

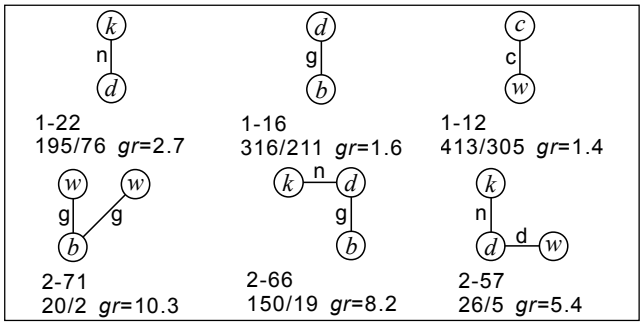

Fig. 5. EPs for the class H. The number a-b attached is an id of the subgraph while the meaning of the fraction c/d is that c and d represent the number of residences in classes H and L, respectively.

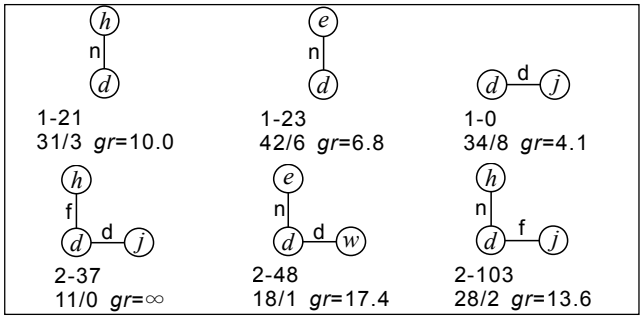

Fig. 6. EPs for the class L. The meaning of the fraction c/d is that c and d represent the number of residences in classes L and H, respectively.

Fig.7 shows top two EPs for class HH. These two graphs 4-502 and 6-989 respectively indicate a dining room facing a balcony, and a dining room connected to a western style room and a hall through doors. Both EPs have common characteristics such as a separate kitchen and a dining room connected to a Japanese style room through a fusuma.

Fig.8 shows top three EPs for the class LL. The graphs 2-49, 4-935 and 8-778 in the figure respectively indicate a dining room directly connected to a hall, a dining room directly connected to an entrance without any partition and to a western style room through a fusuma, and two Japanese style rooms facing a balcony.

From these observations we obtain the conclusion that the segregation level of a dining from an entrance and a hall, the presence or absence of a separate kitchen, and the position of a Japanese style room seem to greatly influence on the rent.

C. EPs for Old/New Residences

In order to see the difference of room layouts according to the building age, we again computed EPs. Here we consider the two classes O and N where class O consists of 70 residences which were constructed before 1979 while the class N consists of 519 residences which were constructed after 1990. Though old dwellings have often been remodeled, we regard them as old ones.

Fig.9 shows three EPs for class O. Subgraphs 4-583, 7-165 and 1-3 respectively indicate a dining room connected to a hall through a fusuma, contiguous two Japanese style rooms, and a Japanese style room connected to a western style room

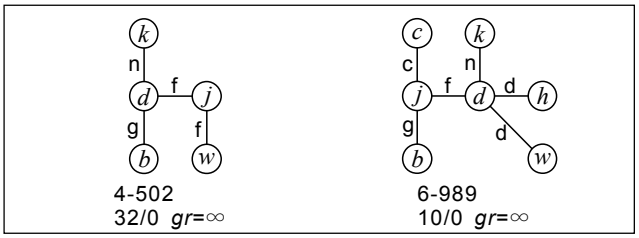

Fig. 7. Top two EPs for the class HH. The meaning of the fraction c/d is that c and d represent the number of residences in classes HH and $\overline{HH}$, respectively.

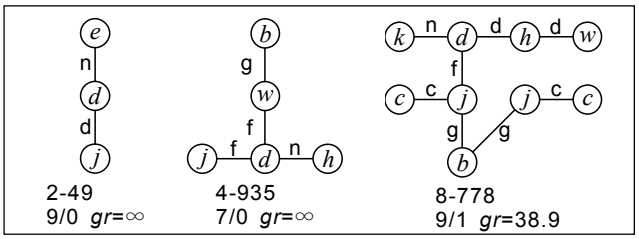

Fig. 8. Top three EPs for the class LL. The meaning of the fraction c/d is that c and d represent the number of residences in classes LL and $\overline{LL}$, respectively.

through a door.

Fig.10 shows two EPs for class N. Subgraphs 7-1079 and 5-1608 respectively indicate a dining room and two western style rooms branched from a hall, a dining room connected to a hall through a door, and a dining room connected to Japanese and western style rooms through fusumas, and to a separate kitchen.

V. Rent Prediction

A. Preparation

In this section, rent prediction models are created from the tabular form data and subgraph data. Three different models are created from three different datasets: tabular data (Model 1), subgraph data (Model 2), and both of tabular and subgraph data (Model 3). In addition, categorical variables such as the nearest station are transformed to boolean ones, and numeric variables are normalized to the range [0, 1].

Since the number of subgraphs discovered by FSG goes up to 8,556 which is much larger than the number of residences, we select significant subgraphs. For this purpose we apply the correlation-based feature subset selection (CFS) method [11] with best-first search whose terminate condition is five consecutive non-improvements on backtracking.

Feature selection is carried out by Weka 3.5.5. with JRE 6.1 on the PC whose CPU is Intel Core2Quad Q6600, memory is 4GB, HDD is 500GB, and OS is Windows Server 2003 Standard R2 x64 Edition. As a result 54 subgraphs are selected, and running time takes 446 seconds. These selected subgraphs are used for rent prediction. The number of subgraphs with respect to the number of edges is shown in fig.11.

In order to construct a rent prediction model we adopt ridge regression [12]. Ridge regression is an extension of classical linear regression such that it avoids multicollinearity by

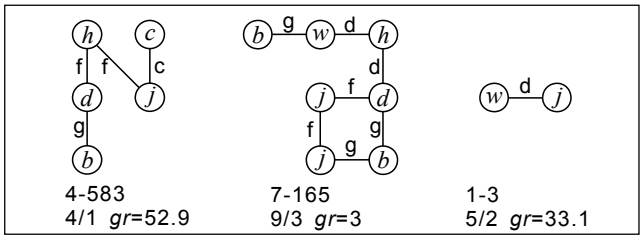

Fig. 9. Top three EPs for the residences constructed before 1979. The meaning of the fraction c/d is that c and d represent the number of residences constructed before 1979 and after 1980, respectively.

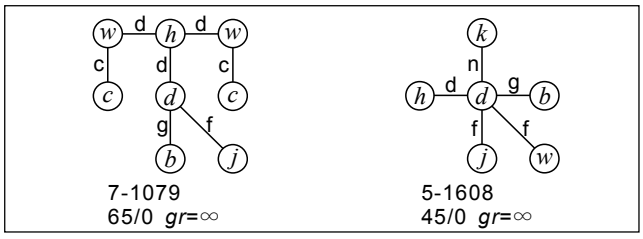

Fig. 10. Top two EPs for the residences constructed after 1990. The meaning of the fraction c/d is that c and d represent the number of residences constructed after 1990 and before 1989, respectively.

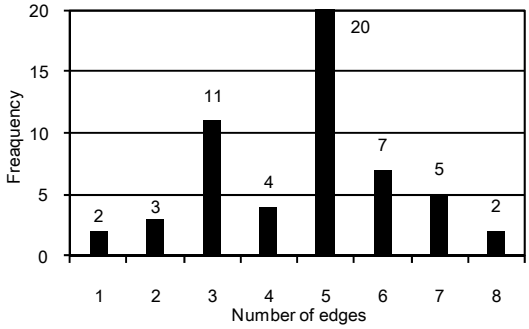

Fig. 11. Histogram of 54 subgraphs extracted in terms of the number of edges.

adding the term of squared sum of partial regression coefficients to the original error minimization criterion with ridge parameter λ which controls the relative importance of both criteria. While explanatory variables which are highly correlated are generally eliminated by stepwise or the other method, ridge regression can use correlated ones at the same time. Thus, we can evaluate the influence of all explanatory variables. This is a major reason why we adopt ridge regression.

Prediction precision is evaluated by multiple correlation coefficients (MCC) and mean absolute error (MAE) between actual and predicted rents derived from leave-one-out cross validation. In this study λ is determined in the preliminary experiment with Model 1 by trial and error: the initial value of λ is set to be 0.01 and is decupled repeatedly until the improvement in precision of leave-one-out is not made.

B. Result

Table III shows the prediction accuracy of three models. Model 3 which uses both of tabular and subgraph data exhibits the highest accuracy such that MCC increases by about 5% and MAE decreases by 700 yen compared with

TABLE III
PREDICTION ACCURACY OF THREE MODELS

	Model 1	Model 2	Model 3
MCC	0.823	0.743	0.876
MAE	5,985 yen	7,421	5,298

TABLE IV
TEN HIGHEST ATTRIBUTES OF PARTIAL REGRESSION COEFFICIENTS ON MODEL 3

Attribute	Coefficient
Occupied area	27100.7
For company only	16288.1
8-780	11716.5
5-397	11593.8
5-1042	8961.4
6-748	8961.4
Area of a family room	8817.4
5-879	7190.4
Area of the smallest western style room	6865.8
5-1436	6721.1

TABLE V
TEN LOWEST ATTRIBUTES OF PARTIAL REGRESSION COEFFICIENTS ON MODEL 3

Attribute	Coefficient
Walking time to the nearest station	-15752.6
Building age	-14373.2
4-893	-12951.4
Tenement	-12686.6
Card key	-10787.4
Distance to a hospital	-10101.7
Piano allowed	-8128.6
5-360	-6574.2
2-49	-6048.7
5-440	-5825.0

those of Model 1. This implies that the room layout has a meaningful effect on rent.

Finally, we examine the attributes influencing on rent positively or negatively by partial regression coefficients of Model 3. Tables IV and V show attributes having ten highest or lowest coefficients respectively. Among ten attributes of table IV, six attributes are related to subgraphs. Let us look at Table V. While attributes concerning distance, building age and so on have large, negative coefficients, coefficients of some subgraphs also take large, negative values. From these results we can confirm the influence of room layouts. Moreover, the number of edges of subgraphs in Table IV is more than five, while that number of subgraphs which appear in Table V is less than five. This result implies that house renters evaluate a residence by demerit system such that they devaluate a residence if it has a little fault of the room layout. Fig.12 and 13 show top three subgraphs which appear in Tables IV and V respectively.

C. Comparison between EPs and coefficients of regression

In section IV we analyzed EPs for four rent classes. Meanwhile, in this section we extracted primal 54 subgraphs that strongly influence on rent by CFS method and observed the strength of the influence as coefficients of multiple regression models. In order to examine the relationship between EPs and coefficients of each extracted subgraph, the

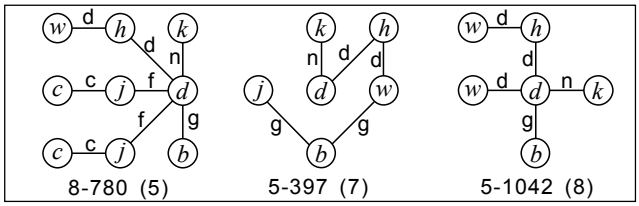

Fig. 12. Upper three subgraphs used in model 3. The number in the parenthesis denotes the number of relevant residences.

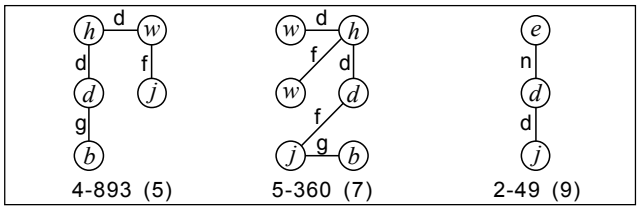

Fig. 13. Lower three subgraphs used in model 3.

TABLE VI
CONCORDANCE RATE ON MODEL 3

l	HH	H	L	LL
$CR(l,l')$	0.89	0.89	0.89	0.94

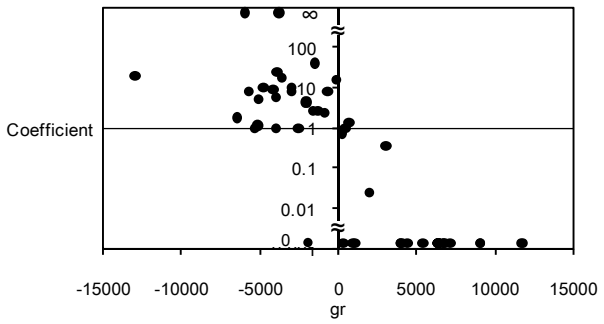

Fig. 14. Gr from $\overline{\text{LL}}$ to LL and coefficient of regression on Model 3.

following concordance rate is defined. Let (l,l') be a pair of rent class labels of the dataset D on EP (i.e. $(l,l') = (\text{H,L}), (\text{L,H}), (\text{HH},\overline{\text{HH}}), (\text{LL},\overline{\text{LL}})$), ES denote the set of extracted subgraphs for the regression model, and a_X denote the coefficient of the extracted subgraph $X \in ES$ in the multiple regression model. The concordance rate $CR(l,l')$ between the gr from $D_{l'}$ to D_l and the coefficient is defined as:

$$CR(l,l') = \sum_{X \in ES} f(a_X, GR^l_{l'}(X)) / |ES|,$$

where

$$f(a_X, GR^l_{l'}(X)) = \begin{cases} 1, & \text{if } (l \in \{\text{HH,H}\} \text{ and } a_X > 0 \text{ and } GR^l_{l'}(X) > 1) \text{ or} \\ & (l \in \{\text{LL,L}\} \text{ and } a_X < 0 \text{ and } GR^l_{l'}(X) > 1), \\ 0, & \text{otherwise.} \end{cases}$$

For example, an extracted subgraph whose gr to HH or H is more than 1 is expected to have a positive coefficient. In case of class LL or L, the coefficient is expected to be negative. If this relationship holds for all extracted subgraphs, $CR(l,l')$

becomes 1. Table VI shows $CR(l,l')$ of each pair of labels on Model 3. $CR(l,l')$ whose l is HH, H or L are the same and that of LL is the highest. This may be related to the comment such as devaluation of apartment mentioned in the previous section. Fig.14 shows the scatter plot of gr from LL to LL and coefficient of regression on Model 3. Grs in the positive coefficient area tend to be 0. Meanwhile, grs in the negative area tend to show various values within the range from 1 to infinity. Although we generally pay attention to the EP whose gr is infinity, this result implies that other EP whose gr is not so large may become a useful factor for regression task. In summary, from these observations we can see the good correspondence of EPs and coefficients of the regression model.

VI. Conclusion

In this paper, we examined the relationship between room layout and rent by discovering subgraphs in the graph representing the room layout based on a graph mining algorithm. Frequent subgraphs of room layout were discovered by graph mining algorithm from the data of family-oriented apartments in the suburb of Kyoto City. Next, discriminative subgraphs which are strongly correlated with rent or building age were extracted by Emerging Patterns. Three different regression models for rent prediction were then constructed from the tabular form data and subgraph data. Through these analyses, typical room layouts which have a strong influence on rent were revealed, and it became clear that room layout significantly contributes to the rent of apartments. Finally, the relationship between EPs and coefficients of the multiple liner regression model was analyzed, and good correspondence was seen.

Then, we want to add the following remarkable results. In the previous studies of spatial analysis using graphs have considered only the adjacency relationship of rooms but not considered the kind of partitions between rooms. However, our study revealed that a certain kind of partition significantly influences on rent. In addition, subgraphs which have strong influences on rent are not so simple. These results could not be derived until graph mining algorithm was used. We also want to mention the limitation of our current study. Existing graph mining algorithms extract both subgraphs which are in parent-children relationship and have the same support. The more complex subgraph is redundant for room layout analysis. The method which can suppress the output of such redundant subgraphs is expected.

Appendix

We assume that a database is composed of a set of attributes and that the original dataset denoted by T (i.e., a set of transactions) in a database have attribute set denoted by A. Each transaction is associated with a class label. Here we assume there are two class labels l and l'. A transaction is expressed as $\{(A_i, v_i^t) \mid A_i \in A\}$, where v_i^t is the value of attribute A_i in transaction $t \in T$. Each pair (attribute, value) is called an item, and an itemset is a set of items.

Let $I = \{i_1, i_2, ..., i_N\}$ be a set of all possible items. A transaction can be identified with a subset of I. A subset X of I is called a k-itemset when $k=|X|$. We say a transaction t contains an itemset X, if $X \subseteq T$. The support of an itemset X in a dataset D denoted by $supp_D(X)$ is derived from $\#D(X)/|D|$. Here, $\#D(X)$ denotes the number of transactions containing X in D. Given a positive number σ, we say an itemset X is σ-*large* in D if $supp_D(X) \geq \sigma$, and X is a σ-*small* in D otherwise. Let $Large_\sigma(X)$ (resp. $Small_\sigma(X)$) denote the collection of all σ-*large* (resp. σ-*small*) itemsets.

For a given ordered pair of datasets D_l and $D_{l'}$, the growth-rate of an itemset X from $D_{l'}$ to D_l denoted by $GR^l_{l'}(X)$ is defined as:

$$GR^l_{l'}(X) = \begin{cases} 0, & \text{if } supp_{D_l}(X) = 0 \text{ and } supp_{D_{l'}}(X) = 0, \\ \infty, & \text{if } supp_{D_l}(X) \neq 0 \text{ and } supp_{D_{l'}}(X) = 0, \\ \dfrac{supp_{D_l}(X)}{supp_{D_{l'}}(X)}, & \text{otherwise.} \end{cases}$$

Given $\rho > 1$ as a growth-rate threshold, an itemset X is said to be an ρ-emerging pattern (ρ-EP or simply EP) from $D_{l'}$ to D_l if $GR^l_{l'}(X) \geq \rho$.

References

[1] P. Linneman, "Some Empirical Results on the Nature of the Hedonic Price Function for the Urban Housing Market", *Journal of Urban Economics*, 8(1), 1980, pp.47-68.

[2] R. A. Dubin, "Predicting House Prices Using Multiple Listings Data", *Journal of Real Estate Finance and Economics*, 17(1), 1998, pp.35-59.

[3] K. Sumida, "A Hedonic Analysis of Condominium Prices in Kanazawa City", *Journal of Kanazawa Seiryo University*, 36(2), 2002, pp.55-63 (in Japanese).

[4] K. Kurosawa, "A Study on The Applicability of Pattern-Analytical Approach for House Planning", *Journal of architecture, planning and environmental engineering. Transactions of AIJ*, 381, 1987, pp.90-99 (in Japanese).

[5] B. Hillier and J. Hanson, *The Social Logic of Space*, Cambridge Univ. Pr., 1984.

[6] T. Hanazato, Y. Hirano and M. Sasaki, "Syntactic Analysis of Large-Size Condominium Units Supplied in the Tokyo Metropolitan Area", *Journal of architecture and planning*, 591, 2005, pp.9-16 (in Japanese).

[7] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data", *Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases* (PKDD'00), LNAI 1910, Springer-Verlag, 2000, pp13-23.

[8] M. Kuramochi and G. Karypis, "Frequent subgraph discovery", *Proc. of 2001 IEEE International Conference on Data Mining* (ICDM), 2001, pp.313-320.

[9] X. Yan and J. Han, "gSpan, Graph-based substructure pattern mining", *ICDM'02: 2nd IEEE Conf. Data Mining*, 2002, pp.721-724.

[10] G. Dong, and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences", *Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA: ACM Press, 1999, pp.43-52.

[11] M.A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*, Ph.D Thesis, Univ. Waikato, 1999.

[12] A.E. Hoerl and R.W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, 12(3), 1970, pp.55-67.