Application of Data Mining
In Micro-scale Urban Feature Analysis

A. Sokmenoglu[1], G. Cagdas[2], S. Sariyildiz[3]

[1]Istanbul Technical University & TU Delft
ITU Mimarlik Fakultesi Taskisla Taksim Istanbul, Turkey
+90 532 3420885
ahusokmenoglu@yahoo.com

[2]Istanbul Technical University
ITU Mimarlik Fakultesi Taskisla Taksim Istanbul, Turkey
+90 212 2931300
cagdas@itu.edu.tr

[3]Delft University of Technology
Design Informatics Faculty of Architecture,
TU Delft, Julianalaan 134, 2628 BL Delft, The Netherlands
+31 (0)15 27 85997
I.S.Sariyildiz@tudelft.nl

1. Introduction

This abstract introduces an ongoing research project addressing multi-dimensional and relational complexity of urban environments by the application of data mining as a methodology of knowledge discovery in micro-scale urban feature analysis. This research is an attempt to establish a link between knowledge discovery methodologies and automated urban feature analysis. After presenting our motivation, research questions and our methodology, an application of data mining of urban features will be briefly introduced in this abstract.

2. Motivation

By the beginning of the 1960's, as planning as a design-led practice seemed to fail to explain how urban processes occur, many urban theorists started to criticize the analysis of urban system from the perspective of few interrelated factors, without considering the multi-dimensionality of the system in a deductive fashion (Jacobs, 1961, Lefebvre, 1970, Harvey, 1973, Alexander, 1979). Hence, in the scope of this research, main motivation is that, in urban analysis, there is a need to advance from traditional one-dimensional (Marshall, 2004) description and classification of urban forms (e.g. Land-use maps, Density maps) to the simultaneous consideration of multi-dimensional aspects of urban systems. For this purpose, data mining is proposed as an analysis methodology for urban feature analysis. When applied to discover relationships between urban attributes, data mining can constitute a methodology for the analysis of multi-dimensional relational complexity of urban environments (Gil, et al., 2009). There are several recent studies of data mining applications in the domain of urban and geographical research such as works

of Demsar, 2006, Reffat, 2008, Behnisch and Ultsch, 2008, Liu and Seto, 2008, Cheng and Wang, Cheng and Anbaroglu, 2009 Christopoulou, 2009, Gil, et al., 2009.

3. Research Questions and Methodology

This research aims to address multi-dimensional and relational complexity of urban environments by applying data mining as a methodology of knowledge discovery in urban feature analysis, with a particular interest in exploring the patterns and relationships of micro-scale data in Beyoglu (a historical neighbourhood of Istanbul) as an application area. Two main research questions are formulated:

- What knowledge can be extracted from existing conventional urban analysis maps of Beyoglu, by the application of data mining methodologies? How this knowledge can be represented?
- Could data mining of urban attributes can produce valuable results and assist architects and urban planners at design, policy and strategy levels?

Within the scope of this research, a methodology is developed specifically for formulation and analysis of an urban database of Beyoglu. This methodology consists of the application of data mining into a GIS based urban database built out of official real data of Beyoglu, operating in three stages; Database formulation, Database analysis and Database evaluation. This methodology, applied in Beyoglu, is illustrated in Figure 1.

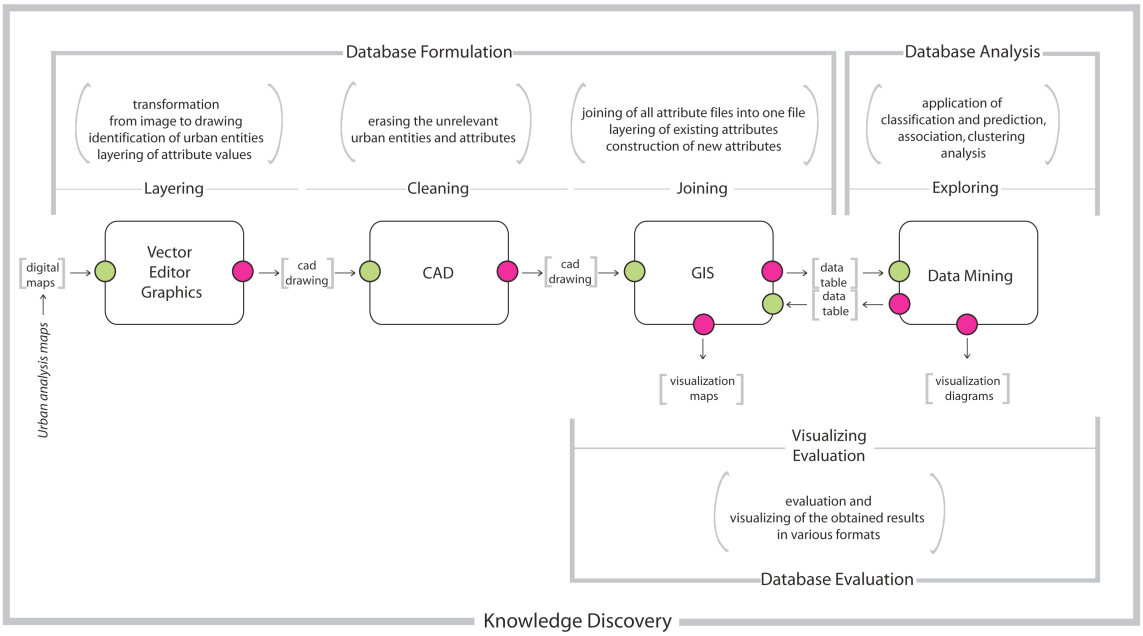

Figure 1. Methodology of knowledge extraction from urban data by data mining

In the following section, an application of the developed methodology will be explained through its stages.

3.1 Database Formulation

In this first stage, micro-scale urban data of Beyoglu is extracted from the various urban analysis maps of 2008 Master Plan of Preservation of Beyoglu provided by the Istanbul

Metropolitan Municipality (IBB). In Table 1, there is a list of urban feature data included in the database as urban attributes.

Attributes			
Att.1-17	Land Use_Ground Floor,	**Att.16**	Land Use_2nd Penthouse
Att.2	Land Use_1st Floor	**Att.17**	Land Use_3rd Penthouse
Att.3	Land Use_2nd Floor	**Att.18**	Neigborhood Name
Att.4	Land Use_3rd Floor	**Att.19**	Density (Person/Ha)
Att.5	Land Use_4th Floor	**Att.20**	Presence in the Bosphorus Silhouette
Att.6	Land Use_5th Floor	**Att.21**	Building Maintenance Conditions
Att.7	Land Use_6th Floor	**Att.22**	Building Construction Style
Att.8	Land Use_7th Floor	**Att.23**	Empty floor ratio
Att.9	Land Use_8th Floor	**Att.24**	Ownership
Att.10	Land Use_9th Floor	**Att.25**	Density of Registred Buildings
Att.11	Land Use_10th Floor	**Att.26**	Factor of Constructable Land (k.a.k.s)
Att.12	Land Use_1st Basement Floor	**Att.27**	Registred Places for Preservation
Att.13	Land Use_2nd Basement Floor	**Att.28**	Ground floor surface area
Att.14	Land Use_3rd Basement Floor	**Att.29**	Distance to Galatasaray
Att.15	Land Use_1st Penthouse	**Att.30**	Distance to Taksim

Table 1. Classification of processed urban attributes of Beyoglu

Available data of the historical neighbourhood of Istanbul covers several scales (from district to block, street, building and building floor) and different forms of classification themes including density, land-use, land value, ownership, material, physical conditions, road attributes, geological attributes and mobility infrastructure and more. There are 11,985 buildings, 700 building blocks, 30 neighbourhoods included in the urban database of Beyoglu preservation area (approx. 3,500,000 m2). The attributes (namely urban features of Beyoglu) of these buildings, building blocks and neighbourhoods are stored in the attribute table available in GIS. So far, in total, there are 30 attributes processed in the form of data table, ready for data mining, 27 attributes gathered from the Beyoglu Master Plan Analysis maps and 3 attributes calculated in GIS are processed in the form of data table.

3.2. Database Analysis and Evaluation

After the formulation of a micro-scale urban feature database for Beyoglu, this urban database is analyzed by Rapid Miner open-source software and the results are evaluated. The data mining analysis is concerned with the investigation of these generic questions;

- Are there significant recurrence patterns of attributes of the land? (Identification of groups, clusters, strata, or dimensions in data that display no obvious structure)
- How dependent and independent are these attributes? (Identification of associations and links among attributes, factors that are related to each other)

- How influential are these attributes on a particular urban phenomenon? (Identification of factors that are related to a particular outcome of interest (root-cause analysis)

Specifically, an analysis of data mining will be briefly introduced here, as an attempt to investigate second question listed above. Naïve Bayesian Method of Classification is applied for predicting the land use value of ground floor (Att.1) of the buildings by means of other attributes; land use value of first floor (Att.2), density of person (Att.19) living in the building and neighborhood (Att.18) where the building is located, distance to Taksim (Att.30), distance to Galatasaray (Att.29), building surface area (Att.28). In Table 2, below, there is a list of these attributes and their value range, subject to this data mining application.

	Attribute	**Urban Entity Level**	**Values**	**Value Type**
Att.1	Land Use_Ground Floor	Building Floor	{Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other}	8 nominal categories
Att.2	Land Use_1st Floor	Building Floor	{Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other}	8 nominal categories
Att.18	Neigborhood Name	Neigborhood	{Arap Camii, Asmalimescit, Bedrettin, Bereketzade, Bostan, Bulbul, Catmalimescit, Cihangir, Cukur, Emekyemez, Evliya Celebi, Firuzaga, Gumussuyu, Hacimimi, Huseyinaga, Kalyoncu Kullugu, Kamer Hatun, Katip Musafa, Kemankes, Kilicali Pasa, Kocatepe, Kuloglu, Mueyyetzade, Omeravni, Purtelas, Sahkulu, Sehitmuhtar, Sururi, Tomtom, YahyaKahya}	30 nominal categories
Att.19	Density (Person/Ha)	Building Block	{0-100, 100-200, 200-300, 300-500, 500-750, 750-1000, 1000-1500, 1500-2000, 2000+, non person living}	10 nominal categories
Att.28	Ground floor surface area	Building	{0-34 m2, 35-48 m2, 49-61 m2, 62-81m2, 82-114 m2, 115-187 m2, 187-17928 m2} (Quantile Classification Method)	7 numeric categories
Att.29	Distance to Galatasaray	Building	{0-293 m., 294- 451 m., 452-588 m., 588-721 m., 722-872 m., 873-1048 m., 1049-1508 m.} (Natural Breaks, Jenks Classification Method)	7 numeric categories
Att.30	Distance to Taksim	Building	{ 0-450 m., 451-693 m., 694-919m., 920-1178, 1179-1453m., 1454-1728m., 1729-2071m.} (Natural Breaks, Jenks Classification Method)	7 numeric categories

Table 2. Selected urban entities, their attributes and range of attribute values

Below in Figure 2, there is a Rapid Miner screenshot illustrating the process of data mining consists of applying a Naïve Bayesian learning operator and a cross-validation in order to estimate the performance of the learning operator.

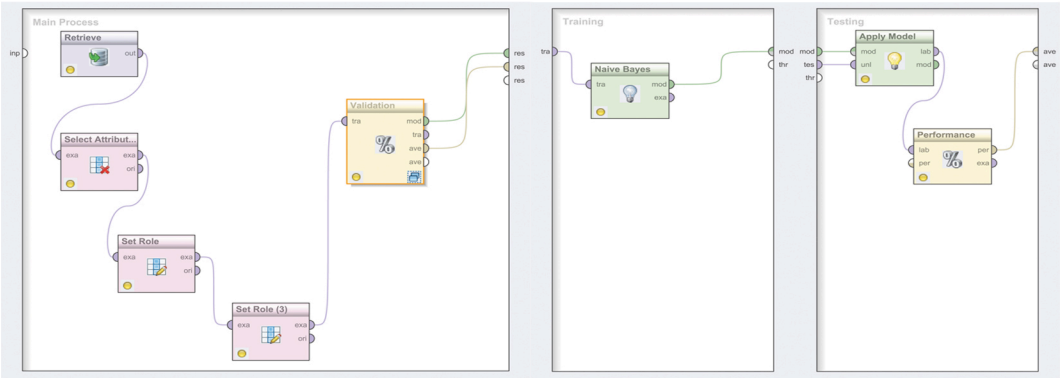

Figure 2. Process of data mining in Rapid Miner

First test is to predict Att.1 by Att.2. Results of this process can be seen in Figure 3, in the form of accuracy table, given by Rapid Miner software.

Table View Plot View

accuracy: 74.63% +/- 1.02% (mikro: 74.63%)

	true Other	true Residential	true Business-Shoppin	true Accomodation	true Sociocultural Infra	true Technical Infrastru	true Empty	true Open Space	class precision
pred. Other	688	138	428	4	409	19	64	1	39.29%
pred. Residential	11	3689	1115	0	8	1	110	0	74.77%
pred. Business-Shoppi	9	18	2923	0	12	1	86	0	95.87%
pred. Accomodation	1	0	17	149	0	0	0	0	89.22%
pred. Sociocultural Infr	0	2	69	0	296	8	3	0	78.31%
pred. Technical Infrastı	0	0	1	0	0	10	0	0	90.91%
pred. Empty	1	21	470	0	0	0	1009	0	67.22%
pred. Open Space	0	0	0	0	0	0	0	139	100.00%
class recall	96.90%	95.37%	58.19%	97.39%	40.83%	25.64%	79.32%	99.29%	

Figure 3. Accuracy table

As seen in the table, the overall accuracy of prediction is 74.63 %, which is significant in terms of claiming a dependency relationship between the land-use values of ground floor and first floors of the buildings in Beyoglu, in general. In case of residential use of ground floor for instance, the model predicts 3689 of the residential as residential and 179 of the residential as false, which gives a 95.37% class recall. More, the model predicts 149 of the accommodation as accommodation and 4 of the accommodation as false, which gives a 97.39% class recall. The model is successful in predicting the land-use values in case of other uses (96.90%), residential (95.37%), accommodation (97.39%), empty (79.32%) and open spaces (99.29%) uses. On the other hand, the model do not return significant results in case of business-shopping (58.19%), socio-cultural infrastructure (40.83%) and technical infrastructure (25.64%). This means that, to some extend in general, land use value of first floor of the building is dependent on the land use value of the ground floor. This hypothesis is especially valid in case of other uses, residential, accommodation, empty and open spaces uses.

After completing all the tests of this analysis with the rest of the attributes (Attributes 18, 19, 28, 29, 30) similar to the test introduced above, briefly we found that to a large extend in general, land use value of first floor is the most influential attribute among others, on determining the land use value of the ground floor. Neighborhood of the building and density of person living in the building are influential on determining the

land use value of the ground floor only in case of residential use with a class recall over %80. Distance to Taksim and Galatasaray (major transportation nodes in Beyoglu) are both not influential on determining the land use value of the ground floor, although in residential and business-shopping cases it can be claimed that there is a small degree of dependency which is over 60%. Surface area of the building is not influential on determining land use value of the ground floor in general, except significantly, in case of business-shopping use there is an accuracy level of 74.50%. These hypotheses must be certainly verified by means of other analysis methods in order to test their validity. Still the results are inspiring enough to expect that this kind of relational analysis methods of urban features could result in valuable site-specific knowledge.

4. Conclusion

Methodology of urban feature analysis applied in this research provides a multi-dimensional study of urban entities meaning that how each attribute of an entity is related to the other(s). Not only one kind of attribute is in interest, many of them are considered in a simultaneous manner. Departing from classical one-dimensional description of urban features' attributes, by means of the computational methods, this research looks for capturing the interrelations among those attributes. Hence, the focus of the analysis is on the relationships that exist within the order of an urban area rather than a conventional description of this urban order. More, microscopic or detailed view of urban system proposed in this research by relying on micro-scale data, provides a way of exploring urban system as complex as it is, allowing a deeper understanding of the system. Finally, data mining seem to provide a promising way of addressing multi-dimensional and relational complexity of urban environments by enabling to explore hidden patterns and relationships among urban features.

5. Acknowledgments

We would like to thank to Nuffic, the Netherlands Organization for International Cooperation in Higher Education, for funding of this research.

6. References

Alexander, C, 1979, *The Timeless Way of Building*, Oxford University Press

Behnisch, M, Ultsch, A, 2008 'Urban Data Mining Using Emergent SOM' in Preisach, C., Burkhardt, H., Schmidt-Thieme L., and Decker, R. (eds.), *Data Analysis, Machine Learning and Applications*, Springer Berlin Heidelberg.

Behnisch, M and Ultsch, A, 2009, 'Urban data-mining: spatiotemporal exploration of multidimensional data', *Building Research & Information*, 37(5): 520-532.

Cheng, T. and Wang, J, 2008, Integrated Spatio-temporal Data Mining for Forest Fire Prediction. *Transactions in GIS*, 2: 591–611.

Cheng T., Anbaroglu B, 2009 Spatio-Temporal Outlier Detection in Environmental Data. *Spatial and Temporal Reasoning for Ambient Intelligence Systems Workshop*; 20- 25 September 2009, France.

Christopoulou, K, 2009 *A Geographic Knowledge Discovery Approach to Property Valuation*, PhD Thesis.UCL

Demsar, U, 2006, *Data Mining of Geospatial Data: Combining Visual and Automatic Methods*, PhD Thesis. KTH

Gil, J, Montenegro, N, Beirao, JN, Duarte, JP, 2009 'On the Discovery of Urban Typologies: Data Mining the Multi dimensional Character of Neighbourhoods', in Çağdaş, G. and Çolakoğlu, B. (eds.),

Proceedings of 27th Conference on Education of Computer Aided Architectural Design in Europe, Istanbul, Turkey, pp. 269-278.

Harvey, D 1973, *Social Justice and The City*, University of Georgia Press.

Jacobs, J, 1961, *The Death and Life of Great American Cities*, Random House INC., New York.

Lefebvre, H, 1970, *The Urban Revolution*, University of Minnesota Press.

Liu, W, Seto, KC, 2008, 'Using the ART-MMAP neural network to model and predict urban growth: a spatiotemporal data mining approach', *Environment and Planning B*, 35(2): 296 – 317.

Marshall, S, 2004, *Streets and Patterns: The Structure of Urban Geometry*, Routledge.

Master Plan of Preservation of Beyoglu, 2008, Istanbul Metropolitan Municipality (IBB).

Rapid Miner Community Edition; http://rapid-i.com/content/blogsection/7/82/lang,en/

Reffat, R, 2008, 'Investigating Patterns of Contemporary Architecture using Data Mining Techniques', in Muylle, M. (ed.), *Proceedings of 26th Conference on Education of Computer Aided Architectural Design in Europe*, Antwerpen, Belgium, pp. 601-608.