Exploring the Patterns and Relationships of Urban Attributes by Data Mining

Ahu Sokmenoglu\textsuperscript{1}, Gulen Cagdas\textsuperscript{2}, Sevil Sariyildiz\textsuperscript{3}
\textsuperscript{1}ITU & TU Delft, Turkey, \textsuperscript{2}ITU, Turkey, \textsuperscript{3}TU DELFT, the Netherlands
\textsuperscript{1}http://www.akademi.itu.edu.tr/sokmenoglu/
\textsuperscript{1}sokmenoglu@itu.edu.tr, \textsuperscript{2}glcagdas@gmail.com, \textsuperscript{3}I.S.Sariyildiz@tudelft.nl

Abstract. This paper introduce an ongoing research project aiming to research data mining as a methodology of knowledge discovery in urban feature analysis. A methodology is developed for formulation and analysis of an urban database by applying data mining methodologies in order to identify the patterns and relationships among multiple urban attributes. The urban database is built upon real and official data of a historical neighborhood of Istanbul.

Keywords. Urban feature analysis; data mining; urban database; urban context; urban attributes.

Introduction

An ongoing research project will be presented in the scope of this paper. This research aims to research 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 neighborhood of Istanbul). The developed methodology for formulation and analysis of the urban database of Beyoglu and an initial data mining application will be presented and discussed. Firstly there will be a presentation of the motivation, hypotheses and the objectives of this research project. Next, there will be an introduction of data mining as a methodology for knowledge discovery in urban feature analysis. Following that, there will be the clarification of the proposed methodology of the research and its application to Beyoglu. Finally there will be a discussing section to state the difficulties of the research followed by a conclusion where the possible contributions of data mining and this research in specific will be listed as well as the further work to come.

Motivation and Hypotheses

The theoretical motivation is derived by the lack of explanatory urban knowledge which is an issue since 1970’s in the area of urban research. This situation is mostly associated with deductive methods of analysis. The analysis of urban system from the perspective of few interrelated factors, without considering the multi-dimensionality of the system in a deductive fashion was not been explanatory enough. (Jacobs 1961, Lefebvre 1970, Harvey 1973) To address the multi-dimensional and relational complexity of urban environments requires the consideration of diverse spatial, social, economic, cultural, morphological, environmental, political etc. features of urban entities.

There are two main hypotheses behind this research. One is that to understand the hidden relationships and patterns of an urban system there is a need of simultaneous consideration of a great
number of independent and dependent features, which is almost impossible to operate manually. Thus there is a need for an automated analysis and discovery method; a computational approach to urban analysis is therefore proposed. The other hypothesis is that, urban system should be handled as complex as it is, meaning that (a microscopic or detailed view) micro-scale analysis is proposed here for a deeper understanding of the system. This approach is in opposition to the abstract, deductive, low-level of detailed methods of urban analysis. The main claim 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 consideration of the simultaneous multi-dimensionality of urban systems.

For this purpose, this research proposes to use data mining methodologies for knowledge discovery in urban feature analysis. Data mining is defined as to extract important patterns and trends from raw data (Witten and Frank, 2005). 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, Montenegro, Beirao and Duarte, 2009) To clarify what data mining is and how it can contribute, there will be an introduction of data mining as a methodology for urban feature analysis, in the following section.

Data Mining as a Methodology for Knowledge Discovery in Urban Feature Analysis

Knowledge discovery approaches are designed to determine unknown patterns and relationships within large amounts of data that could not be uncovered through simple queries or reports. Knowledge discovery techniques are in opposition with the confirmatory techniques of analysis which requires a priori hypotheses that restrict the researcher and prevent the discovery of previously unknown information. (Miller, Han 2001)

Data mining popularly referred as knowledge discovery in databases (KDD) is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases. (Han & Kamber, 2001) It differs from classical statistics as a multidisciplinary field drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high performance computing and data visualization. (Han & Kamber, 2001)

The ultimate goal of data mining is to provide evidence-based insight through deeper understanding of data (in the mind of the analyst) and to produce results that can be utilized at policy and strategy levels. It is widely applied in many fields of science, engineering and business but it has also applications in geographical information systems, remote sensing and many other areas related to spatial data, under the name of spatial data mining. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography.

Architects, urban planners, urban geographers, social scientist use both data mining and spatial data mining methodologies for urban knowledge discovery from spatial and non-spatial data. The type of data concerning those fields of research is usually multidimensional, spatially correlated and heterogeneous. Data mining is known to be a suitable methodology for solving non-trivial, complex problems where there is a high volume of diverse data. It is operating by the application of appropriate algorithms for extracting patterns from data. It involves either the fitting of the models to data or the determination of patterns from data where both deterministic and non-deterministic approaches are available. (Fayyad et al, 1996) With the capacity of incorporating various variables without restricting the analyst to a few ones and offering various approaches of analysis, data mining seems to be a promising methodology for knowledge discovery in urban
feature analysis where a large number of diverse and interrelated factors are simultaneously operating.

There are several studies applying data mining in the context of architectural and urban research such as; analyzing multidimensional characteristics of German communities by Behnisch and Ultsch (2008, 2009), identification of Saudi Arabian architectural typologies by Reffat (2008), investigation of the way location affects property prices by Christopoulou (2009), explorations for identifying potential relationships between incident locations and other factors in Helsinki by Demsar (2006), extracting descriptions of street and block typologies by by Gil, et al (2009), modeling urban growth based on spatial data mining by Liu and Seto (2008).

**Objectives, Approach and Methodology**

As stated in the Introduction, this research aims to apply 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 neighborhood of Istanbul). To meet this aim, a methodology is developed for formulation and analysis of an urban database of Beyoglu. The methodology consists of the application of data mining as a knowledge discovery method into a GIS based conceptual urban database built out of official real data of Beyoglu. The power of this open system approach is in its independency from any existing theoretical perspective of urban analysis limited by selected specific properties of urban entities. The data to be included in the urban entity’s database is not limited by the category which it belongs to but the most important limitation could be the unavailability of the high amount of diverse data. This is a data-driven approach not relying on specific theories but only relying on the data itself. There are no a priori assumptions about the system in investigation but this is an attempt to discover previously unknown hidden patterns and relationships within a specific urban database. Hence, this approach also suggests a context-specific analysis.
The proposed methodology has three main stages which will be generally explained here and through its application in Beyoglu, in the next section:

- **Database formulation**: Building a dataset out of micro-scale urban data (data in the form of any facts, numbers, or text that can be processed by a computer) by using different computer applications.
- **Database analysis**: Analysis of the patterns, associations, or relationships among data by applying queries and appropriate data mining methodologies ranging from regression analysis and classification to clustering and exploratory data analysis, by a data mining software application.
- **Database evaluation**: Evaluation of the obtained results in the form of correlation amongst variables, data groupings (clusters) or more complex hypotheses and visualization in the form of maps and diagrams.
- **Below there is a diagram illustrating the overall process of the proposed methodology explained above.**

### Application of the Methodology

One of the reasons of this particular interest in Beyoglu is that, this area was recently analyzed in high level of detail in terms of its urban features as a basis for the Master Plan of Preservation of the Beyoglu.

Urban analysis maps of the 2008 Master Plan of Preservation of Beyoglu provided by the Istanbul Metropolitan Municipality (IBB) are very appropriate for testing the methodology of this research in terms of their richness in covering micro-scale urban features. But most importantly, the reason why Beyoglu is appropriate for this research is that it is a very good sample for a complex urban center. Beyoglu is the center of living, commerce and leisure in Istanbul, which is known for its heterogeneous population characteristics in terms of its inhabitants and daily visitors, very attractive both as an historical and contemporary city center, very diverse in terms of land-use and population characteristics.

Therefore Beyoglu is a perfect example of an urban system where complexity of interrelated micro-scale features of urban entities can be studied by applying data mining. Methodology of the research applied in Beyoglu, will be explained below through its stages.

#### Database formulation

In this 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). Available data of the historical neighborhood 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, and so on.
road attributes, geological attributes and mobility infrastructure and more. There are more than thirty attributes coming from the Master Plan Analysis maps, from which so far, ten attributes are processed in the form of data table, ready for data mining.

As seen in Figure 1, the raw data is gathered in the form of maps (one map for each attribute) in pdf format. Every map is separately transformed from image to a drawing file (dwg) with layers, in a vector editor graphics application. Next all these drawing files are separately cleaned in a CAD application by erasing the irrelevant entities. Then all these drawing files are one by one taken in a GIS application where they are joined into one file, all the attributes in the form of layers. This file contains all the data coming from each map of analysis. Available quantitative and qualitative data is associated with the geographical space with the help of a GIS platform and represented both as a cartographic map and as an urban database. In GIS, 2D Polygon geometry represents urban entities which are basically individual buildings. There are 14,364 buildings in the Beyoglu preservation area, all included in the urban database. The attributes of these buildings are represented in the attribute table available in GIS. Figure 2 represents the list of urban entities, their ten different attributes and the range of their values.

Attributes which are listed below are obtained directly from the Master Plan Analysis maps. Later on, within the scope of the research, not only the rest of the attributes coming from the Master Plan Analysis maps will be included but also more attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Urban Entity Level</th>
<th>Values</th>
<th>Value Type</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt.1</td>
<td>Land Use_Ground floor</td>
<td>Building floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.2</td>
<td>Land use_1st Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.3</td>
<td>Land Use_2nd Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.4</td>
<td>Land Use_3rd Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.5</td>
<td>Land Use_4th Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.6</td>
<td>Land Use_5th Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.7</td>
<td>Land Use_6th Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.8</td>
<td>Land Use_7th Floor</td>
<td>Building Floor</td>
<td>[Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (8 categories)</td>
</tr>
<tr>
<td>Alt.9</td>
<td>Neighborhood Name</td>
<td>Neighborhood</td>
<td>[Urban, Residential, Business, Shopping, Social, Infrastructure, Technical, Infrastructure, Accomodation, Open Space, Empty, Other]</td>
<td>nominal (10 categories)</td>
</tr>
<tr>
<td>Alt.10</td>
<td>Density (Persons/ha)</td>
<td>Building Block</td>
<td>0-100, 100-200, 200-300, 300-500, 500-750, 750-1000, 1000-1500, 1500-2000, 2000+, non person living</td>
<td>nominal (10 categories)</td>
</tr>
</tbody>
</table>
will be calculated with the help of GIS tools. One of the most important properties of this methodology is that it proposes an open-ended formulation of an urban database where any new and relevant entities and their attributes can be added as they become available.

Figure 3 is a section from the computer representation of Beyoglu in GIS, in the form of data table. Same section of Beyoglu is represented in Figure 4, in the form of cartographic map, associated with the data table in Figure 3. Buildings are the entities with unique ID’s and attributes as seen for ID.8971 and ID.8972 in Figure 3 and 4.

The urban database consisting of 14,364 buildings and their ten attributes is imported for analysis in a data mining application software.

Database analysis by data mining and Evaluation of the results

Beyoglu urban database is to be analyzed by Rapid Miner [1] open-source software. The analysis will be done to;

- 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)

The process of building the database is still running at this stage of the research. Therefore there are so far only some initial experiments regarding the phase of data mining. Naïve Bayesian Method of Classification was applied to investigate possible relationships among some of the features of buildings in Beyoglu. Bayesian classifiers are statistical classifiers which can predict the membership probabilities such as the probability that a given sample belongs to a particular class. (Han, Kamber, 2001) Bayesian classifiers are comparable in performance with decision tree and neural network classifiers and exhibited high accuracy and speed when applied to large databases. (Han, Kamber, 2001) Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This is called class conditional independence and this is made to simplify the computations involved in. It is in this sense that the method is called “naïve”. (Han, Kamber, 2001) Therefore, in this case, even if some attributes of an urban entity depend on the existence of others, a naïve Bayes classifier considers all of these attributes to independently contribute to the probability that this entity is in a specific class. One of the experiments with
interesting results will be briefly introduced here.

This experiment of data mining conducted in Rapid Miner returns interesting results. This is a test for predicting the land use value of ground floor (Att.1) of the buildings by using the land use value of first floor (Att.2) neighborhood (Att.9) where the building is located and density of person (Att.10) living in the building. Shown in Figure 5 there is a Rapid Miner screenshot illustrating the process of data mining which consists of applying a Naïve Bayesian learning operator and a cross-validation in order to estimate the performance of the learning operator.

The result of the process is given in the table (Figure 6) below.

As seen in the table, the model predicts 4030 of the residential as residential and 433 of the residential as false, which gives a 90.30% class recall. More, the model predicts 154 of the accommodation as accommodation and 13 of the accommodation as false which gives a 92.22% class recall. This means that, to some extent the land use value of first floor (Att.2), neighborhood of the building (Att.9) and density of person living in the building (Att.10) is influential on determining the land use value of the ground floor in case of residential or accommodation uses. This hypothesis of course has to be studied by means of other methods to look for strong dependencies between those attributes. The prediction accuracy for the rest of the land use values is not very significant as seen in the table.

![Figure 5](image.png)

*Process of data mining which consists of applying a Naïve Bayesian learning operator and a cross-validation in order to estimate the performance of the learning operator.*

![Figure 6](image.png)

*Results of Naïve Bayesian Classification in the form of accuracy table.*
Discussions

There are some issues and difficulties concerning the research and these are both related to the methodology and to the data-driven nature of the research. Main issues are;

• The inconsistency while converting data from graphics to CAD drawing format which results with the loss of many existing data due to the layering process;
• The time consuming nature of the data formulation stage due to the high amount of data;
• The construction of new attributes from existing ones demands a good level of knowledge of GIS analysis tools and mostly there is need to go further by designing and implementing new tools inside GIS;
• In terms of its application and evaluation, data mining require a good level of statistics expertise both for selecting the appropriate algorithms and for a better interpretation of the obtained results;
• The problem of uncertainty is always an issue. This database represents only some selected aspects of the reality but not the complete reality itself. Reality is infinitely complex and choices are always to make about how much detail to include, how much to generalize or approximate; (Goodchild, 2006)
• Availability of the relevant and temporal data is always an issue. The inclusion of temporal data would be promising in terms of capturing the change in urban data and the dynamism of the urban system in investigation. Unfortunately temporal data is not available;
• To discuss the underlying causes of the discovered relationships and patterns is important for validating the results of data mining process;

Conclusion

To conclude; in the paper, the framework and the first results a research aiming to design an urban database based on micro-scale urban data of a historical neighborhood of Istanbul for further analysis by data mining, is presented. Data mining in the context of urban data is introduced as a computer based, data-driven, context-specific approach for supporting analysis of urban systems without relying on any existing theories. Availability of the many advanced methods of data analysis in data mining allows identifying the patterns and relationships among a great number of independent and dependent urban attributes instead of limiting to a few ones. Data mining in the context of urban data;

• Can help in the design process by providing site-specific insight through deeper understanding of urban data.
• Can produce results that can assist architects and urban planners at design, policy and strategy levels.
• Can constitute a robust scientific base for rule definition in urban simulation applications such as urban growth prediction systems, land-use simulation models etc.

Specifically, the possible contributions of this research will be;

• Construction of a digital urban database for Beyoglu based on official data gathered from the Istanbul Metropolitan Municipality. This urban database specifically designed for this research can also enable usage and further analysis by other researchers, as this is a real world application where data is real and official.
• Using GIS and data mining as complementary applications in urban feature analysis.
• To use and test data mining as a knowledge discovery method for analyzing micro-scale urban data.
• To gain insight about the patterns and relationships among multiple urban features of Beyoglu.
• To underline the importance of collecting and storing urban data for tracking the changes of urban systems. Urban research is a data poor field comparing to many other fields of science. To store urban data is extremely important in
terms constructing the memory of urban space.

- Further work to come in this research will be to complete the data table of Beyoglu by adding the rest of the attributes and constructing new attributes in GIS. Next, to elaborate the process of data mining by experimenting with more attributes and different types of analysis methods available in Rapid Miner. Finally to elaborate the process of visualizing is in the scope of the further work of this research.

Acknowledgments

I would like to thank to Nuffic, the Netherlands Organization for International Cooperation in Higher Education, for funding of this research. I would like to thank Ceyhun Burak Akgul for his support in Data Mining and to H. Serdar Kaya for his support in GIS.

References


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


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

Han, J, Kamber, M 2001, Data Mining: Concepts and Techniques, Morgan Kaufmann.


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


