Structural Relational Graph Based Data Mining Applied to the Multi-Functional Spaces of Properties in "Puebla of the Angels" in the XVI, XVII, and XVIII Centuries

[1]Oscar E. Romero A., [1]Jesus A. Gonzalez, [1]Ivan Olmos, and [2]Rosalva Loreto

[1]National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro No. 1, Sta. Maria Tonantzintla 72840
Puebla, Mexico
{oromero, jagonzalez, iolmos@ccc.inaoep.mx}

Benemerita Universidad Autonoma de Puebla
Social and Human Sciences Institute
Palafox y Mendoza 208, Centro 72000
Puebla, Mexico
rossloreto@yahoo.com

Abstract

There are many interesting applications of knowledge discovery applied to real world structural domains. In this paper, we present the application of the graph-based data mining system Subdue to the "Multifunctional Spaces of the Properties in Puebla of the Angels in the XVI, XVII, and XVIII Centuries. Our experimental results show how important is the knowledge representation to successfully find relevant patterns at different levels of abstraction. We worked very close to the domain expert who used the patterns found to support some of her theories and also proposed new hypotheses to guide her research. This is a successful project where data mining techniques were used for decision support in a history domain.

Introduction

In this research, we use a graph-based data mining system that allows us to work with structural domains (actually with any domain that can be represented with graphs) called Subdue and we apply it to the history structural domain named "Multifunctional Spaces of the Properties in Puebla of the Angels in the XVI, XVII, and XVIII Centuries". One of the greatest challenges in our study was the compilation of this information from the General Archives of the Notary Office of the State of Puebla of the Angels generated from transactions of turnover of these houses and buildings. The main idea of this work is to find patterns that differentiate between the multifunctional spaces of the different parishes (each parish had a different jurisdiction) in the following two aspects: distribution of the space, construction characteristics and material of the spaces. Next we describe the most important parts of our work including our results.

Subdue

Subdue is a graph-based knowledge discovery system developed at the University of Texas at Arlington, by the machine learning group [Cook, Holder, and Djoko 1995]. The input to Subdue is a graph that represents objects and their attributes with labeled vertices and edges, vertex labels give name to objects or their attributes. Relations between objects and relations among objects with their attributes are represented by labeled edges, where the label determines the name of the relation. This allows us working with any domain that can be represented with graphs. Subdue is able to perform the concept learning task and this version of the algorithm is known as SubdueCL. SubdueCL accepts a set of positive and negative examples as its training set and uses a set covering approach to find substructures that describe a sub-set of positive examples but not the set of negative examples. Then, the algorithm learns a set of rules in DNF [Gonzalez, Holder and Cook 2002].

Multifunctional Spaces of Houses in "Puebla of the Angels" in the XVI, XVII and XVIII Centuries

This domain comes from a large project called "Habitar y vivir: Análisis del Espacio habitacional de la ciudad de Puebla. 1690 - 1890".

The geographic division of Puebla City from the XVIth to the XVIIIth centuries depended directly on the churches system [Loreto 2003]. Each district of the city was contained within the jurisdiction of a parish-district. The description of the houses used in our research was obtained from writings of transactions of the General Archives of the Notary Office of the state of Puebla. Each assessment describes a house in base to its construction characteristics and its circulation spaces.

Knowledge Representation

We tested with three alternative representations: 1. circulation spaces, 2. circulation spaces + transitory spaces, 3. circulation spaces + transitory spaces + construction characteristics. An example with part of our graph-based representation of a property is shown in figure 1a. In this figure we can see the circulation sub-spaces: "street", "tailgate", "yard", "storeroom", "stairs", and "hall" and the connections between them (there exists a door between "street" and "tailgate", "yard" and "storeroom" and an open connection in all the other cases). Double contour in the vertices describing spaces represents the second floor. We can also include transitory spaces (internal spaces such as a door connecting two main spaces) as we can see in figure 1b. Finally, we can include all the information about a property that describes not only the main and transitory spaces but also their construction characteristics. An example of this representation is shown in figure 1c.

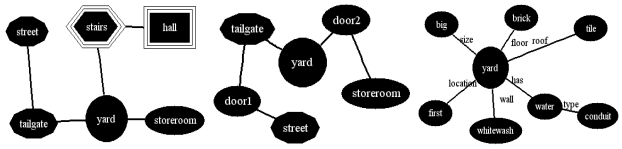

Figure 1. Knowledge Representations. a) Circulation spaces between the objects, street, tailgate, yard, stairs, hall and storeroom. b) Transitory spaces between the spaces street, tailgate, yard and storeroom from figure 1a. c) The yard space (from figure 1a) and its construction characteristics.

Methodology

The experiments were made with the knowledge representations presented before. The properties information (or houses examples) was divided in three subgroups as follows: District of Analco 49 assessments, District of San Sebastian 51 assessments, Districts of Analco and San Sebastian, 100 assessments. We ran Subdue with each of the representations for each subset of properties. We also tested SubdueCL (with the three knowledge representations) using the Analco district as positive examples and San Sebastian as negative examples. We then switched Analco to be our negative examples and San Sebastian to be the positive examples. Some of the results of these tests are presented in the next section.

Results

In this section we present the most significant results for the set of experiments mentioned in the previous section. Figure 2a shows one of the patterns found in the district of Analco using representation 1 with Subdue. In figure 2b we can observe one of the transitory patterns found in the district of San Sebastian using representation 2 with Subdue. The substructure shown in figure 2c is a pattern that reflects some of the construction characteristics of the properties in the district of Analco (this pattern was found using representation 3 with Subdue). This is a very significant pattern for our history domain because it reflects the importance that wood doors had as means of protection of the property (the door of the tailgate) and as means of privacy between different subspaces in the property. Historians were sure that security was not important by that time and this pattern shows the opposite (we found a similar pattern with windows).

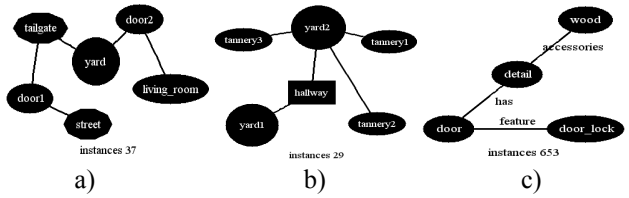

Figure 2. Subdue Results. a) Circulation pattern found at the Analco parish. b) Transitory patterns found at the San Sebatian parish. c) Construction characteristics of doors found at the Analco parish.

We also ran SubdueCL with representation three (construction characteristics) using data from the Analco district as positive examples and data from the San Sebastian district as negative examples. We performed a 10-Fold cross validation to see how well the model distinguishes between the properties that belong to the Analco and San Sebastian parishes. In this test we obtained a 94.85 % of accuracy. From the history point of view, these substructures help us to distinguish the circulation and construction patterns between the two parishes. The differences are due to the ethnic groups living in them and also to their living conditions.

Conclusion

The different knowledge representations used in this work, helped to find significant patterns at different levels of abstraction by means of Subdue. These patterns are being used by our history domain expert to confirm her hypotheses and to generate new theories about how people was organized to populate new areas.

Acknowledges

This project is supported by CONACYT under project number 38257H.

References

Cook, Diane J., Holder, Lawrence B. and Djoko, Surnjani, 1995. Knowledge discovery from Structural Data. Journal of Intelligent Information Systems. Vol. 5, no. 3, pages 229-245.

Gonzalez, Jesus; Holder, Lawrence B; and Cook, Diane J. 2002 "Graph-Based Relational Concept Learning", Proceedings of the Nineteenth International Conference on Machine Learning.

Loreto Lopez, Rosalva, 2003. La Casa, la Vivienda y el Espacio Doméstico en la Puebla de los Angeles del Siglo XVIII. Mexico.