Mining Real-Estate Listings Based on Decision Systems over Ontological Graphs

Extended Abstract

Krzysztof Pancerz[1,2] and Olga Mich[1]

[1] University of Management and Administration
Akademicka Str. 4, 22-400 Zamość, Poland
kpancerz@wszia.edu.pl
[2] University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszów, Poland

Abstract. In the paper, we describe the process of mining real-estate listings based on decision systems over ontological graphs. Such decision systems have been proposed to deal with data in the form of concepts linked by different semantic relations. A special attention is focused on preprocessing steps transforming advertisements in the textual form into decision systems (decision tables) defined over ontological graphs.

Keywords: decision systems, ontological graphs, data mining, real-estate listings.

1 Mining Real-Estate Listings

Text mining is a rapidly growing application of knowledge discovery in data (cf. [2]). A special kind of data in the textual form is constituted by advertisements, for example, real-estate ones. In case of advertisements, data have the form of loosely coupled words (terms, concepts) rather than full, grammatically correct sentences. Moreover, underlying data are of qualitative character. We can distinguish several challenges posed by textual data: understanding data semantics and semantic relations between them, considering the external knowledge in processes of data classification, encoding textual data for classifiers working with numerical data.

The semantic relations between concepts play an important role, among others, in cognitive psychology, linguistics and currently also in computer science. In [4], two general types of relations between words (concepts) were distinguished: *Paradigmatic relations* (relations between words belonging to the same grammatical category) and *Syntagmatic relations* (relations between words that go together in a syntactic structure). Paradigmatically related words are, to some degree, grammatically substitutable for each other. In our research, we are interested in paradigmatic relations. As it will be shown later, in real-estate listings, we do not analyze a semantic structure of sentences, but we try to derive some knowledge about concepts (terms) included in them, for example, whether they

are synonyms, whether one concept can be replaced with another, for example, more general one, etc.

In our research, we use the following taxonomy of types of semantic relations (which is modeled on the project called Wikisaurus [1] aiming at creating a thesaurus of semantically related terms): synonymy, antonymy, hyponymy/ hyperonymy (subclass - superclass), and meronymy/ holonymy (part - whole). In the approach presented in this paper, we are interested in synonymy and hyponymy/ hyperonymy. We will use the following notation: $isSyn$ denotes synonymy, $(u, v) \in isSyn$ means that "u is a synonym of v", $isGen$ denotes hyponymy, $(u, v) \in isGen$ means that "u is a hyponym of v" ("u is generalized by v"), and $isSpec$ denotes hyperonymy, $(u, v) \in isSpec$ means that "u is a hyperonym of v" ("u is specialized by v"). Additionally, we take into consideration a semantic relation called "being an instance". Being an instance concerns an example (instance) of a given concept. This kind of relations is important in mining real-estate listings because they include, for example, instances of places.

The knowledge about semantic relations between concepts is included in ontologies. Our approach is based on the definitions of ontology given by Neches et al. [5] and Kohler [3]. That is, ontology is constructed on the basis of a controlled vocabulary and the relationships of the concepts in the controlled vocabulary. Formally, the ontology can be represented by means of graph structures. Let $\mathcal{O}$ be a given ontology. An ontological graph is a quadruple $OG = (\mathcal{C}, E, \mathcal{R}, \rho)$, where $\mathcal{C}$ is a nonempty, finite set of nodes representing concepts in the ontology $\mathcal{O}$, $E \subseteq \mathcal{C} \times \mathcal{C}$ is a finite set of edges representing relations between concepts from $\mathcal{C}$, $\mathcal{R}$ is a family of semantic descriptions of types of relations (represented by edges) between concepts, and $\rho : E \to \mathcal{R}$ is a function assigning a semantic description of the relation to each edge. Sometimes, we can also consider subgraphs of ontological graphs, called by us local ontological graphs. In Figure 1, exemplary ontological graphs representing the real-estate domain are shown.

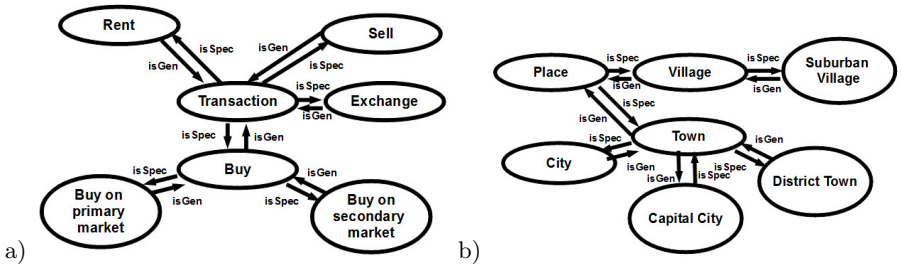

Fig. 1. Exemplary ontological graphs representing the real-estate domain.

In [12], information (decision) systems were proposed as the knowledge representation systems. In simple case, they consist of vectors of numbers or symbols (attribute values) describing objects from a given universe of discourse. In our research, we are interested in mining textual data in the form of concepts (words,

terms). Therefore, in [7] and [10], ontologies were incorporated into information (decision) systems, i.e., attribute values were considered in the ontological (semantic) space. Information (decision) systems over ontological graphs can be created in different ways. In [7], two approaches were mentioned. In the first one, attribute values are concepts from ontologies assigned to attributes - a simple information (decision) system over ontological graphs. In the second one, attribute values are local ontological graphs of ontologies assigned to attributes - a complex information (decision) system over ontological graphs. In case of real-estate listings, whether a client was interested in a given property can be considered as a decision attribute.

The main goal of this paper is to show how to transform real-estate listings into simple information (decision) systems over ontological graphs. This is a very important preprocessing step that can be generally depicted as it is shown in Figure 2. We can distinguish three main steps:

- *Stemming* - defining basic grammatical forms (roots) for particular words existing in advertisements, for example, using a quite popular Porter stemming algorithm [13].
- *Attributation* - assigning concepts (built from words) existing in advertisements to proper attributes as their values, according to defined ontological graphs.
- *Deinstantiation* - replacing instances existing in advertisements with the most specific concepts (with respect to the hyponymy / hyperonymy relation) whose instances they are.

Deinstantiation is an important step if we are interested in a more general knowledge derived from real-estate listings, for example, some client is interested only in houses in a village (not a particular one).

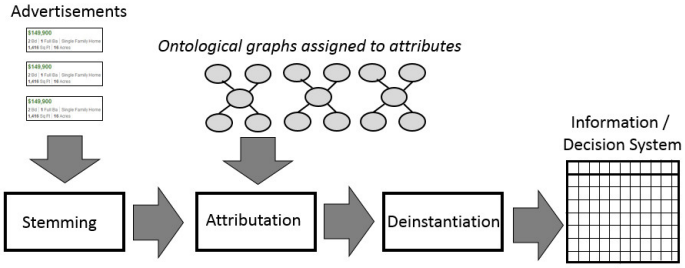

Fig. 2. A procedure for transformation of real-estate listings into simple information (decision) systems over ontological graphs.

Let us consider, as an example, the following advertisement: "For sale, Warsaw, Poland, Villa, bedrooms: 4, 437 m^2". Let us also assume that we take into consideration three attributes: *Transaction*, *Place*, and *Property* with proper

ontological graphs assigned to them. After performing our procedure, the advertisement becomes one object (row) in an information (decision) system:

U/A	$Transaction$	$Place$	$Property$
u_1	Sale	Capital City	Villa

Having an information (decision) system over ontological graphs, we can apply different machine learning and data mining methods to extract some valuable knowledge. In our previous papers, for such systems, we considered: rough sets [9], decision rules [8], decision rules based on the DRSA approach [6], neural networks [11].

References

1. The Wikisaurus Homepage: http://en.wiktionary.org/wiki/Wiktionary: Wikisaurus
2. Kargupta, H., Joshi, A., Sivakumar, K., Yesha, Y.: Data Mining: Next Generation Challenges and Future Directions. The MIT Press, Cambridge, MA (2004)
3. Köhler, J., Philippi, S., Specht, M., Rüegg, A.: Ontology based text indexing and querying for the semantic web. Knowledge-Based Systems 19, 744–754 (2006)
4. Murphy, M.L.: Semantic Relations and the Lexicon. Cambridge University Press (2008)
5. Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.: Enabling technology for knowledge sharing. AI Magazine 12(3), 36–56 (1991)
6. Pancerz, K.: Dominance-based rough set approach for decision systems over ontological graphs. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Proceedings of the FedCSIS 2012. pp. 323–330. Wroclaw, Poland (2012)
7. Pancerz, K.: Toward information systems over ontological graphs. In: Yao, J., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, vol. 7413, pp. 243–248. Springer-Verlag, Berlin Heidelberg (2012)
8. Pancerz, K.: Decision rules in simple decision systems over ontological graphs. In: Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierek, A. (eds.) Proceedings of the CORES'2013, Advances in Intelligent Systems and Computing, vol. 226, pp. 111–120. Springer International Publishing (2013)
9. Pancerz, K.: Semantic relationships and approximations of sets: An ontological graph based approach. In: Proceedings of the HSI'2013. pp. 62–69. Sopot, Poland (2013)
10. Pancerz, K.: Some remarks on complex information systems over ontological graphs. In: Gruca, A., Czachórski, T., Kozielski, S. (eds.) Man-Machine Interactions 3, Advances in Intelligent Systems and Computing, vol. 242, pp. 377–384. Springer International Publishing (2014)
11. Pancerz, K., Lewicki, A.: Encoding symbolic features in simple decision systems over ontological graphs for PSO and neural network based classifiers. Neurocomputing 144, 338–345 (2014)
12. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
13. Porter, M.: An algorithm for suffix stripping. Program 14, 130–137 (1980)