SEMANTIC DATA INTEGRATION ACROSS DIFFERENT SCALES: AUTOMATIC LEARNING OF GENERALIZATION RULES

Birgit Kieler

Institute of Cartography and Geoinformatics, Leibniz University of Hannover, Germany
birgit.kieler@ikg.uni-hannover.de

Commission II/6

KEY WORDS: Interoperability, Integration, Data Mining, Generalization, Matching, Automation, Cartography

ABSTRACT:

In this paper we present an approach realizing the integration of data sets of different origin and with different resolution levels. The underlying idea is to reveal semantic correspondences between object classes of different geo-ontologies only by analysis of spatial and geometrical characteristics of instances of the data sets. As a result we derive transformation rules with Data Mining methods, which subsequently allow the semantic connection between data sets. For our case study we use data sets with similar thematic focus, but different semantic and geometric resolution: on the one hand building objects from cadastral data in a scale about 1:1K (detailed data set) and on the other hand settlement areas from topographic data in a scale about 1:25K (less detailed data set). To derive links between instances from the detailed data set to the more general one by geometric overlay it is required that the data sets are available in the same geographical extent. Then we generalize the detailed data set by using the object boundaries in the less detailed data set as a constraint for the generalization and generate an 'intermediate data set' that is in a similar spatial resolution. We enrich the given information with additional attributes representing spatial relations and implicit or intrinsic given instance properties (e.g. object size), in order to derive transformation rules. These rules can be further used for classification of settlement areas of unknown regions in the target data set.

1 INTRODUCTION AND OVERVIEW

With the growing automation in spatial data capture and the increasing accessibility of geodata via web services, a huge number of geospatial data sets are available. A combination of all these data is desirable, in order to improve the quality of single data sets, enable interdisciplinary analysis procedures, and thus promote interoperability. However, data integration can be difficult, when the data not only differ in thematic focus, resolution and quality, but also have different data models with often a poor documentation and no explicit semantics. So it is not possible to identify semantic correspondences in the ontologies without expert knowledge. Since this knowledge is not always available, our approach derives these semantic relations by the analysis and exploitation of geometrical and topological properties of object instances themselves. Using these and additionally determined geometrical characteristics transformation rules between the data sets on schema level will be derived automatically by means of Data Mining methods. Already Volz (2005) has used instances of different road data sets in order to derive corresponding object classes in different schemas using statistical analysis. Also Duckham and Worboys (2005) analyse instances of data sets, but in contrast by methods of lattice theory. The rules derived from our approach can be applied for classifying other, unknown regions of the data sets, where the classification is still missing. First results of applying this approach to data of similar scale have been produced using frequency analysis (Kieler, 2007). In this paper we focus on identifying semantic correspondences *across different scales*. First thoughts on this topic are already included in Kieler et al. (2007). Now the challenge is that it is no longer possible to match individual object instances, as they differ in semantic and geometric granularity. Therefore generalization operations have to be applied to bring both data sets to a comparable level of detail.

The paper is structured as follows. In the next section the background of the research is outlined and references to the used methods are given. Then, the used data sets are briefly introduced. In section 4 the approach that enables the identification of semantically similar object classes in two different geo-ontologies and the derivation of powerful classification rules across scales are presented. After presentation and discussion of first results in section 4, the paper concludes with an outlook on future work.

2 RELATED WORK

2.1 Semantic data integration

There is a lot of research dealing with semantic data integration, semantic annotation of geodata (Klien, 2007) and especially with detection of semantic similarities in different ontologies. Kokla (2006) presents a paper, that analyzes and compares existing integration approaches and describes the principal directions to perform semantic integration of geographic ontologies. One option to identify semantic correspondences is to do it manually by careful inspection of given object catalogues or ontologies. But such a manual process is no longer feasible, if we aim at an integration of arbitrary data sets that can be loaded via the internet. Another option is to determine semantic similarity measures, in order to establish the degree of potential semantic interoperability between data of different origin. Because a common theory on semantic similarity does not exist, Schwering (2008) summarizes the existing approaches for the measurement of semantic similarity. In addition to the presentation of the five different classifications of similarity measures: geometric, feature, network, alignment and transformational measure, also the different mathematical foundations, knowledge presentations and notions of similarity are presented in detail. Especially the semantics of geospatial objects are complex and should not be detected solely by a term comparison. These objects are typically described by spatial and geometrical properties, attributes and relations and therefore it is desirable to use all these information improving the derivation of semantic relationships of different data sets, using similarity measures for a successful data integration.

2.2 Data Mining

The term Data Mining refers to extracting or 'mining' knowledge from large amounts of data, and is also known as knowledge extraction, knowledge discovery in databases (KDD) or data/pattern analysis (Fayad et al., 1996). Typical functionalities are concept description, association analysis, classification and prediction, cluster and also outlier analysis (Han and Kamber, 2001). Specific algorithms for analysing geographic data sets are developed and coined as Spatial Data Mining (Miller and Han, 2001). Among these, classification methods are also known in the field of Machine Learning as supervised or unsupervised learning techniques (learning from examples). These are two step procedures. In the first step, a model is built to describe the set of data classes or concepts based on the examples of a training data set. In the second stage, this model is used to predict the resulting class of new data items (Russell and Norvig, 2003). Different methods to derive classification rules and to represent them are known, but one of the most famous is the induction of decision trees, especially the ID3 algorithm (Quinlan, 1986) or its further developments (e.g. C4.5) (Quinlan, 1993). These algorithms divide the examples in a top-down recursive manner into branches with nodes and leafs at the end using an entropy-based measure - also known as information gain - as a heuristic approach to separate the samples into individual classes (Han and Kamber, 2001). Entropy is a fundamental concept in information theory representing "a measure of how much 'choice' is involved in the selection of an event" (Shannon, 1948). That means higher entropy involves more choices and information, and is not good for classification. Recursively therefore, the algorithms of decision trees are trying to discover the lowest entropy, less choices and information, for the best classifications.

3 DATA SOURCES

In the work presented here, we use two geodata sets in vector format describing topographic objects in different resolutions. On the one hand we use settlement areas of ATKIS data (the German Authoritative Topographic Cartographic Information System) in 1:25K and on the other hand building objects of ALK data (the digital German cadastral map) in 1:1K. In Figure 1 a small part of the investigated region is displayed.

The following textual descriptions will illustrate the difficulty in detecting semantic similarities or correspondences across different scales only from the object class definitions.

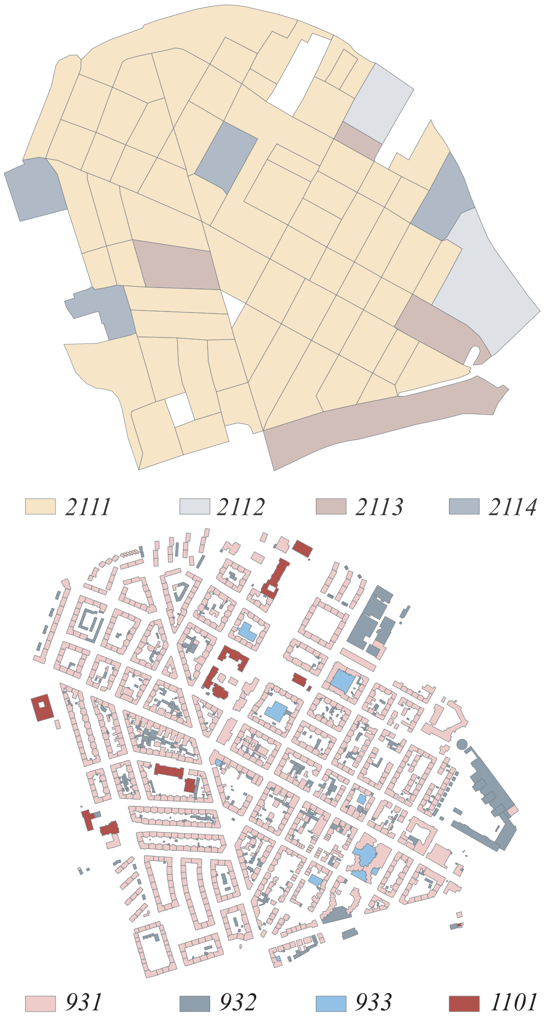

Figure 1: Data sources: ATKIS - less detailed data set with 4 types of settlement areas (top) and ALK - detailed data set with 4 classes of building objects (bottom).

1. The textual catalogue descriptions of the analyzed ATKIS object classes of settlement areas (AdV, 2008):
   - *2111*: Area with buildings, predominantly or solely used for residential purposes. Besides these residential buildings also shops to supply this area, non-disturbing craft producers, facilities for religious, cultural, social and sanitary purposes are allowed.
   - *2112*: Area with buildings, predominantly or solely used for industrial or craft producing purposes. This includes e.g. shopping malls, warehouses / depots, large-scale commercial farms, processing and disposal plants and trade fair facilities.
   - *2113*: Area with buildings without a typical purpose of the buildings. This includes especially areas with a rural character, e.g. agricultural or forestry companies, residential buildings and central areas in a city with commercial buildings and vital economic and administrative facilities.
   - *2114*: Area with buildings of certain purposes. This includes purposes of administration, health and social affairs (hospital), education, research (university), culture (church), safety and order (penitentiary), vacation or weekend homes and national defense.
2. The textual catalogue descriptions of ALK object classes of building objects (VKV, 2008):
   - *931*: Private building (residential building).
   - *932*: Private building (outbuilding).
   - *933*: Subterranean building.
   - *1101*: Public building.

The descriptions are very fuzzy and imprecise regarding the building object classes, that only distinguish very general between private, public or subterranean buildings. Without detailed knowledge about the data sets that for example hospitals or schools have to be classified as public buildings, it is not possible to identify semantic relationships between the object descriptions of both data sets. The problem will be more complex, if the descriptions of the object classes are entirely missing or the terms of the object classes consist only of arbitrary combinations of letters and numbers. Therefore a mapping regarding to terms of the object class is not reliably possible in each case. For this reason the approach presented here relies on the power of the spatial and geometrical properties of data instances themself, in order to establish rules, making it possible to infer AKTIS settlement use types from the ALK building objects.

4 APPROACH

In order to identify semantically similar object classes between the presented data sets only by analysis of geometrical and topological properties of object instances, we have to preprocess the data. The preprocessing is the most expensive and effort consuming step in the knowledge discovery process (Bogorny et al., 2006). Our kind of preprocessing is described in the next subsection in more detail. After this step we use the edited data as training data for learning classification rules by means of the J48 algorithm in the WEKA Data Mining software[1]. A simple example for such a classification rule could be:

IF $parcel\ area \geq\ 10.000\ m^2\ \wedge\ contain\ only$ $private\ buildings$ ***THEN*** $parcel\ type = residential.$

We can apply these derived rules subsequently to a test data set in order to pursue the following two purposes. The first purpose is to find the connection of semantically similar object classes of investigated ontologies and the second to classify unknown regions of the more general data set, referred to as generalization.

4.1 Preprocessing

The derivation of meaningful transformation rules by directly linking the instances across the different scales and subsequently counting and analyzing of the respective combinations is difficult. Therefore, we generated in a preprocessing step an intermediate data set I. The following example illustrates the situation, which has been investigated within this study: Parcels with different building types (private, public, ...) will be aggregated into different types of settlement areas. Depending on most occuring building types in an aggregated area, either an industrial area or a residential area will be created. Thus, a correspondence between individual parcels and the aggregated areas will not yield to unique correspondence types. Therefore, we produce the intermediate data set I with a similar geometric resolution to the less detailed data set by means of generalization. Specifically, we generalize the detailed data set (ALK - building objects) with the object boundaries of the less detailed data set (ATKIS - settlement areas) as a constraint. That means all objects of ALK data that spatially overlap with objects of ATKIS are merged and form intermediate objects. The new intermediate data set I has the same number of instances and the geometry of the ATKIS data set. The ALK instances are considered as attributes (see Figure 2). The following short example should illustrate the linking of instances across the whole data sets, before the preprocessing step: 4590 x *2111* → *931* , 127 x *2112* → *931*, 3830 x *2113* → *931*, ... and afterwards concerning of the blue marked object of ATKIS class *2114* →{3 x *931*, 4 x *932*, 2 x *1101*}(see Figure 2).

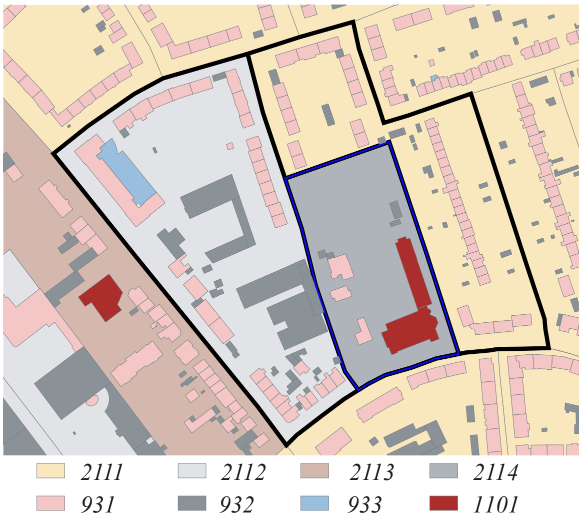

Figure 2: Intermediate data set I formed by spatial overlay of both data sets.

For the further analysis using Data Mining, we have to determine relevant properties of corresponding instances, because these are used to infer the rules later on. The strength of the different properties and the effect of the attributes to the quality of the results will be discussed in the next subsections.

For the evaluation of our approach we use a training data set, in the following referenced as $Train$, for training and extracting the classification rules and a test data set referenced as $Test$ for testing and validating the definitions. The $Train$ data set covers an area of 17 km^2 and the $Test$ data set encompasses 11 km^2. The distribution of each object class in both data sources regarding to the $Train$ and $Test$ sets are listed in Table 1.

	Class Type	#	Class Type	#
	ATKIS	Instances	ALK	Instances
$Train$	*2111*	335	*931*	8740
	2112	21	*932*	4209
	2113	257	*933*	185
	2114	121	*1101*	454
$Test$	*2111*	365	*931*	6483
	2112	41	*932*	5317
	2113	15	*933*	51
	2114	31	*1101*	161

Table 1: The total number of instances regarding to the object classes separated by $Train$ and $Test$ data sets.

4.2 Determination of spatial relations and geometrical attributes as an indication for semantic similarity

For the derivation of correct and strong transformation rules, consisting of geometrical instance properties, we use the J48 algorithm, because J48 can classify numeric attributes and is not restricted to nominal attribute types like the ID3 algorithm. The algorithm requires an attribute list for each instance of I as input. This list must have at least one nominal attribute, for example the object class type as a classifier. As a numeric attribute we introduce the spatial relation 'containment' by determining the number of buildings lying within one parcel. Also the kind of building type may be of particular interest. The mixture of different building types in a parcel provides information about the land use type. The statement, that a residential area always consists exclusively of private buildings is not true in each case. But what is the global definition of a residential area? The mere presence of one single building type does not give rise to the correct type of settlement area. Additional hints for a relation are provided by

[1] Waikato Environment for Knowledge Analysis; http://www.cs.waikato.ac.nz/ml/weka

geometric properties of the instances, for example the size or the shape of buildings and parcels.

In our analysis we calculate for each instance in the intermediate data set I the following itemised attribute values according to the different source data sets. This attribute list does not raise the claim of completeness, but is sufficient for our current purposes.

1. ATKIS - settlement areas
   (a) *ATKIS Class Type*$_N$: The settlement type $N =$ *2111, 2112, 2113, 2114*.
   (b) *ATKIS Area*: The area size of the parcel in square meters.
   (c) *# AKTIS Neighbors Type*$_N$: The total number of neighbors separated for each settlement type N.
   This attribute can only be considered if the neighbors are already classified.
2. ALK - building objects
   (d) *# ALK Type*$_M$: The total number of buildings for each building type $M =$ *931, 932, 933, 1101*.
   (e) *ALK Area Type*$_M$: The total area in square meters for each building type M within a ATKIS parcel.
   (f) *Ø ALK Area Type*$_M$: The average area for each building type M within a ATKIS parcel.
   (g) *ALK Overlay Ratio Type*$_M$: The overlay ratio in percent between the total area for each building type M to the area of the ATKIS parcel.

4.3 Derivation of transformation rules using J48 algorithm and discussion of the results

For an effective classification of AKTIS settlement types as function of several attributes, the attribute list has to be converted in the ARFF-format, which is necessary for using WEKA. For the minimal example of three ATKIS instances (black bordered in Figure 2) one input file for WEKA has to be structured as presented in Figure 3. In order to preserve the clarity, we use, for this example, only three of the above introduced attributes, namely the *ATKIS Class Type*$_N$ as nominal attribute and classifier, as well as the *ATKIS Area* and the *# ALK Type*$_M$.

```
@relation Train

@attribute ATKIS Class Type_N {2111, 2112, 2113, 2114}
@attribute ATKIS Area real
@attribute # ALK Type_931 real
@attribute # ALK Type_932 real
@attribute # ALK Type_933 real
@attribute # ALK Type_1101 real

@data
2111, 24391.04, 40, 26, 0, 0
2112, 35974.70, 34, 28, 1, 0
2114, 16162.61, 3, 4, 0, 2
```

Figure 3: Example for input-file in ARFF-format.

In the following the change of the accuracy by altering attributes has been examined. At this point it should be noted that the accuracy of the classification of the $Test$ set is influenced by the precision of the learned rules from the $Train$ set. The quality of these rules is in turn depending on the chosen parameter set and has to be set for each classifier algorithm. This especially includes the confidence factor (in our case 0.25) and the specification of the minimum number of instances per leaf (10) in the decision tree. Moreover, the quality of the used attributes is important for the final quality.

The overall accuracy A_O, calculated from the number of correctly classified instances regarding to the total number of instances of all object classes, is summarized for the $Train$ and $Test$ sets in Table 2 and presented for the analyzed attribute combinations.

Attributes	$Train$ [%]	$Test$ [%]
(a),(d)	64	78
(a),(d),(e)	74	71
(a),(d),(e),(f)	76	77
(a),(d),(e),(f),(g)	76	83
(a),(b)	53	47
(a),(b),(d)	69	40
(a),(b),(d),(e)	75	68
(a),(b),(d),(e),(f)	78	75
(a),(b),(d),(e),(f),(g)	76	82
(a),(b),(c),(d),(e),(f),(g)	84	85

Table 2: The overall accuracies A_O of the $Train$ and $Test$ sets for different attribute combinations.

Consequently the highest accuracy value for the $Train$ set yielded the combination (a),(b),(c),(d),(e),(f),(g) with 84%. This combination includes all introduced attributes of both data sets, in contrast to the combination (a),(b) with 53% which uses only attributes of the ATKIS data sets. The right mixture of attributes from both data sets influences the accuracy. Also, a high overall accuracy is no guarantee, that the accuracy for the test data set is also high. Comparing the accuracies of Table 2 shows that in 50% of the cases the accuracies for the $Test$ set are lower than those of the $Train$ set. But the A_O is not always an adequate indicator for a good classification for all classes, because the accuracy is depending on the correctly classified instances of all classes. For that reason we have to consider also the accuracies $A_{ClassType}$, with respect to the single object types. For the evaluation on the $Test$ set $A_{ClassType}$ are represented for the same attribute combinations in Table 3.

Attribute combinations	*2111*	*2112*	*2113*	*2114*
(a),(d)	90	0	13	71
(a),(d),(e)	75	32	40	81
(a),(d),(e),(f)	84	32	40	81
(a),(d),(e),(f),(g)	88	59	27	74
(a),(b)	56	10	20	10
(a),(b),(d)	41	10	33	71
(a),(b),(d),(e)	72	32	47	81
(a),(b),(d),(e),(f)	80	32	40	81
(a),(b),(d),(e),(f),(g)	88	49	33	74
(a),(b),(c),(d),(e),(f),(g)	91	59	27	74

Table 3: The single accuracies $A_{ClassType}$ for the $Test$ set and for different attribute combinations.

Comparing the A_O of the attribute combinations (a),(d) and (a),(d),(e),(f) in the $Test$ area in Table 2 there is only a difference of 1%. Considering also the single $A_{ClassType}$ of these combinations, it is obvious that the two classes *2112* and *2113*, that were classified very bad or not at all for the first attribute combination, increased their accuracies using the latter combination. On the other hand a decrease of the accuracy of class *2111* is the effect of the change of the attribute combination. To gain satisfying results it is necessary to consider the pros and cons of certain attributes for the specific task. Although the attribute combination (a),(b),(c),(d),(e),(f),(g) has the best accuracies it is not suitable for the classification of unknown data sets, because the attribute (c) is included, requiring the knowledge of the settlement type of the neighboring parcels. But this attribute can be used for example for the improvement of the results in a further iterative process, if a first classification has been carried out. Another scenario could be, if a parcel should be newly classified, because there were changes in the construction of a parcel. For that case, the types of land use are known of the neighboring objects and can be taken into account.

The most accurate combination without (c) with respect to the global accuracy is (a),(d),(e),(f),(g). For this case study the single accuracies and the corresponding detailed classification matrix with the instance numbers are displayed respectively in Table Table 4 and Table 5. On the main diagonal of the matrixes the values of the correctly classified instances are highlighted in bold.

	2111	*2112*	*2113*	*2114*
2111	**88**	0	9	2
2112	2	**59**	37	2
2113	53	7	**27**	13
2114	13	3	10	**74**

Table 4: $Test$ set: Classification statistics of $A_{ClassType}$ in % for the (a),(d),(e),(f),(g) attribute combination.

	2111	*2112*	*2113*	*2114*	# Instances
2111	**323**	0	33	9	365
2112	1	**24**	15	1	41
2113	8	1	**4**	2	15
2114	4	1	3	**23**	31

Table 5: $Test$ set: Classification matrix with the instance numbers for the (a),(d),(e),(f),(g) attribute combination.

In the following the classification rules for the attribute combination (a),(d),(e),(f),(g) are shown.

IF $(g)_{931} \leq 8.56 \wedge (d)_{1101} \leq 0 \wedge (e)_{932} \leq 1328.59 \wedge (e)_{931} > 154.16 \vee (g)_{931} > 8.56 \wedge (e)_{932} \leq 828.27 \wedge (g)_{1101} \leq 6.51 \wedge (f)_{931} \leq 625.89 \wedge (d)_{931} \leq 1 \vee (g)_{931} > 8.56 \wedge (e)_{932} \leq 828.27 \wedge (g)_{1101} \leq 6.51 \wedge (f)_{931} > 625.89 \wedge (e)_{931} \leq 4512.69 \vee (g)_{931} > 8.56 \wedge (e)_{932} > 828.27 \wedge (f)_{932} \leq 66.46$
THEN *ATKIS Class Type* = ***2111***.

IF $(g)_{931} \leq 8.56 \wedge (d)_{1101} \leq 0 \wedge (e)_{932} > 1328.59$
THEN *ATKIS Class Type* = ***2112***.

IF $(g)_{931} \leq 8.56 \wedge (d)_{1101} \leq 0 \wedge (e)_{932} \leq 1328.59 \wedge (e)_{931} \leq 154.16 \vee (g)_{931} > 8.56 \wedge (e)_{932} \leq 828.27 \wedge (g)_{1101} \leq 6.51 \wedge (f)_{931} \leq 625.89 \wedge (d)_{931} > 1 \vee (g)_{931} > 8.56 \wedge (e)_{932} \leq 828.27 \wedge (g)_{1101} \leq 6.51 \wedge (f)_{931} > 625.89 \wedge (e)_{931} > 4512.69 \vee (g)_{931} > 8.56 \wedge (e)_{932} \leq 828.27 \wedge (g)_{1101} > 6.51 \vee (g)_{931} > 8.56 \wedge (e)_{932} > 828.27 \wedge (f)_{932} > 66.46$
THEN *ATKIS Class Type* = ***2113***.

IF $(g)_{931} \leq 8.56 \wedge (d)_{1101} > 0$
THEN *ATKIS Class Type* = ***2114***.

Whereas on the top of Figure 4 the allocation of the *ATKIS Class Type* for the $Test$ data set is illustrated, the correctly and incorrectly classified instances are shown together at the bottom of the figure.

In Table 5 it is noticeable that there is just a small misclassification to class *2112* due to a precise class definition. Most of the misclassifications take place in class *2113*. This leads to the conclusion that the description of this class is very fuzzy (compare the class definition in section 3) and that the attributes are not strong enough to separate this class from the others. Because in class *2113* most of the wrongly classified instances belong to *2111* and *2112* it seems that this class is a combination of these two classes. Comparing the textual definitions this assumption is confirmed, because *2113* holds objects of a mixed use land type, including as well as residential building from *2111* and outbuildings from *2112*.

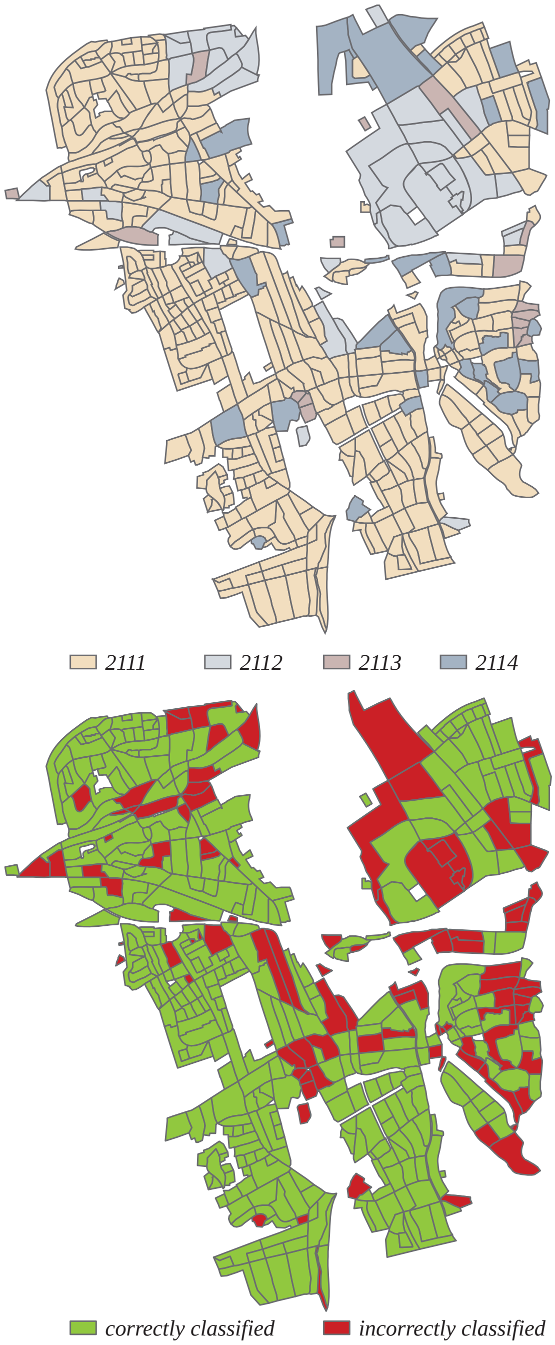

Figure 4: On the top: The overview about the *ATKIS Class Type* in the $Test$ set. At the bottom: The presentation of the correctly and incorrectly classified instances in the $Test$ set using the (a),(d),(e),(f),(g) attribute combination.

The derived transformation rules make a connection between the ATKIS and ALK data sets possible. These can also be used in order to control or to refine the semantic description of the different object definition with more details, for example the proportion of private objects in residential areas.

5 CONCLUSIONS AND FUTURE WORK

This paper presented a way to include geometrical attributes and spatial relations between instances of different data sets for the semantic connection of the geo-ontologies as precondition for an improved interoperability. Transformation rules are derived by means of a Data Mining algorithm. These rules can also be used for the classification of an unknown data set. With the investigation of our test data sets, we reveal, that the accuracies of the derived rules are influenced by a lot of factors, e.g. attribute lists or fuzzy class definitions. The work regarding to these problems is still in progress. In the future the results can be improved on the one hand by adding additional attributes with respect to the detailed data sets. For example, the dominant building type size, the arrangement of buildings as perimeter block development or also linear buildings seem to be very significant for residential areas and could be useful for a clearly distinction to mixed used areas.

In addition the identified transformation rules between the two data sets with different resolutions can be used for an updating of the more general data set. If objects in the detailed data set change, it is possible to adapt the land use type. It is also possible to expand the presented process to an iterative process. That means, in a first stage, classifications yielding good results could be applied. After this pre-classification also further classifications that take neighborhood information into account, can be computed. Because it is obvious, that the type of neighboring objects can be an indicator for the most probable classification. On the one hand clusters of the same land use types are very clearly visible as shown on the top of Figure 4 and on the other hand the principles of urban planning do not allow an industrial area surrounded only by residential areas.

ACKNOWLEDGEMENTS

This work is part of the project 'Automatic Semantic Transformation between Geo-Ontologies' and is funded by the German Research Foundation (DFG). It is part of the bilateral bundle-project entitled 'Interoperation of 3 D Urban Geoinformation' in cooperation with China (http://www.ipk.bv.tum.de/3dugi/).

References

AdV, 2008. ATKIS-Objektartenkatalog. http://www.atkis.de (2008/04/30).

Bogorny, V., Engel, P. and Alvares, L., 2006. GEOARM: An interoperable framework to improve geographic data preprocessing and spatial association rule mining. In: Proceedings of the 18th International Conference on Software Engineering and Knowledge Engineering, San Francisco, USA, pp. 79–84.

Duckham, M. and Worboys, M., 2005. An algebraic approach to automated geospatial information fusion. International Journal of Geographical Science 19(5), pp. 537–557.

Fayad, U., Piatetsky-Shapiro, G., Smith, P. and Uthurusamy, R. (eds), 1996. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, Californien, USA.

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

Kieler, B., 2007. A geometry-driven approach for the semantic integration of geodata sets. In: Proceedings of 23rd International Cartographic Conference, Moscow, Russia.

Kieler, B., Sester, M., Wang, H. and Jiang, J., 2007. Semantic data integration: Data of similar and different scales. Photogrammetrie Fernerkundung Geoinformation (PFG) 6, pp. 447–457.

Klien, E., 2007. A rule-based strategy for the semantic annotation of geodata. Transactions in GIS 11, pp. 437–452.

Kokla, M., 2006. Guidelines on geographic ontology integration. In: International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences (ISPRS Technical Commission II Symposium), Vol. XXXVI Part 2, Vienna, pp. 67–72.

Miller, H. and Han, J. (eds), 2001. Geographic Data Mining and Knowledge Discovery. Taylor and Francis.

Quinlan, J., 1986. Induction of decision trees. In: Machine Learning, Vol. 1, Morgan Kaufmann Publishers Inc., pp. 81–106.

Quinlan, J., 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, Califonia, USA.

Russell, S. and Norvig, P., 2003. Artificial Intelligence: A Modern Approach, 2/E. Prentice Hall.

Schwering, A., 2008. Approaches to semantic similarity measurement for geo-spatial data: A survey. Transaction in GIS 12(1), pp. 5–29.

Shannon, C., 1948. A mathematical theory of communication. Bell System Technical Journal 27, pp. 379–423 and 623–656.

VKV, 2008. ALK-Automatisierte Liegenschaftskarte. http://www.lgnapp.niedersachsen.de/vkv/allgemein/gesetze/n3727350.pdf (2008/04/30).

Volz, S., 2005. Data-driven matching of geospatial schemas. 3693, pp. 115–132.