Workshop "Semantic Web meets Geopatial Applications", held in conjunction with AGILE 2008
Short Paper

Derivation of Semantic Relationships between Different Ontologies with the Help of Geometry

Birgit Kieler

Institute of Cartography and Geoinformatics, University of Hannover, Germany
Birgit.Kieler@ikg.uni-hannover.de

Abstract. For beneficial data integration, semantic correspondences of geodata of different origin which are mostly unknown, must be determined. This is a prerequisite to a growing interoperability of data and services in Spatial Data Infrastructures (SDI). The approach presented here starts with the assumption that geometric similarity and spatial correspondence are leading to semantic correspondences. In this way we infer semantic relations between heterogeneous geodata from the geometric characteristics of instances using feature matching and Data Mining methods.

1 Introduction and Overview

The increasing availability of geospatial data sets, especially via the Internet, allows a growing interoperability of geodata of different origin as well as information sharing and reuse. But the optimal use of these data is not given in the necessary way, because the semantic relationships between arbitrary data sources, like equivalence-, disjunction- or inclusion-relations are mostly unknown. Thus at present it is possible only with great effort and the experience of experts to identify all relevant geospatial data that answer e.g. the following question 'Find all geospatial data in vector format that describe water objects in the northern part of Germany!'. Neglecting the various syntactic heterogeneities (i.e. data types and formats) the requirement for a successful integration of geodata of different origin, that answer this question, is to identify the semantic relations between the data's ontologies and to determine e.g. all the semantic object classes in the different data sets corresponding to 'water objects', ideally automatically.

Only with the knowledge about the content e.g. that object class 'Gewässer' of a German data set is completely equivalent to object class 'Water Bodies' of an English data set and not to object class 'Vegetation Areas' from the latter, a fast and reliable answer to such questions is possible and therefore suitable for a data integration application on the Web.

However, the semantic relations between various object classes are not always as clear as in the example above. In many cases a linguistic similarity between two object classes is not given, because the label of object classes is only an abstract identifier using characters and/or numbers of which one can derive no affiliation. Also, even if corresponding terms are used, there still may be differences in the actual use of the terms in different user communities and their ontologies. For this reasons we want to disregard the labels and the definitions of the object classes. Instead we want to identify semantic relations only by the exploitation of the geometric characteristics of the instances themselves.

The research hypothesis is that object descriptions belong to the same phenomenon, when they describe either spatial similar objects (i.e. objects in the same geographical position), or if they have similar geometric properties. Starting from this hypothesis we are introducing two scenarios in this paper. In the first case we match identical objects represented in different data sets given in the same geographical extent by a geometric overlay. Subsequently we derive the semantic relations between individual object classes of different ontologies.

Not every geographic area is represented in multiple data sets, and thus the derivation of the semantic relationships across identical objects cannot be made. Therefore the second scenario has to be considered. In this case, semantic correspondence is inferred by geometrically similar objects. These object fulfill pre-defined shape rules of certain object categories derived directly by means of Data Mining methods from the data.

The paper is organized as follows. In the next section the background of the research is sketched and references to existing work are given. Then, the methods for the two integration cases are presented. At first the semantic integration of data in the same geographical extent and then in different extent. A summary and an outlook conclude the paper.

2 Related Work

There is a large number of research work dealing with the semantic data integration and especially with the detection of semantic similarities in different ontologies. Kokla [3] presents guidelines for geographic schema or ontology integration. 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 in the internet.

Rodríguez and Egenhofer [7] summarized, that the general approach of data integration has been to map the local terms of distinct ontologies onto a single shared ontology. Then, the semantic similarity is typically determined as a function of the path distance between terms in the hierarchical structure underlying the single ontology. Kokla and Kavouras [4] developed a method for revealing salient semantic information (semantic properties and relations) from existent geographic ontologies with methods from the Natural Language Processing (NLP) in order to perform concept comparison and reconciliation. It is based on the realization that definitions contain an abundance of semantic information. By comparing terms, semantic elements and their value similarities and heterogeneities between geographic concepts are identified. Rodriguez and Egenhofer [7] calculate semantic similarity using other features, such as attributes, parts and functions. This approach is suitable for comparing categories, when both categories do have such complete and detailed descriptions. However, most existing geographic metadata sources do not provide this sort of information and consequently this approach is not appropriate in every case. The drawback of these approaches, that compare the terms, attribute values or descriptions of the object classes is, that in general, it can not be assumed, that the names or descriptions in different ontologies are the same.

In contrast there is another method to automate the integration process, the so called instance-based or extensional method, which uses the instances of the different ontologies in order to determine transformation rules between them (see [1] and [12]). Also Tversky [11] uses for the feature matching approach common and different characteristics between objects or entities to compute semantic similarity. Similar to Volz [12] we want to use the data themselves to derive the semantic relations from the geometric relationships, especially for polygon objects.

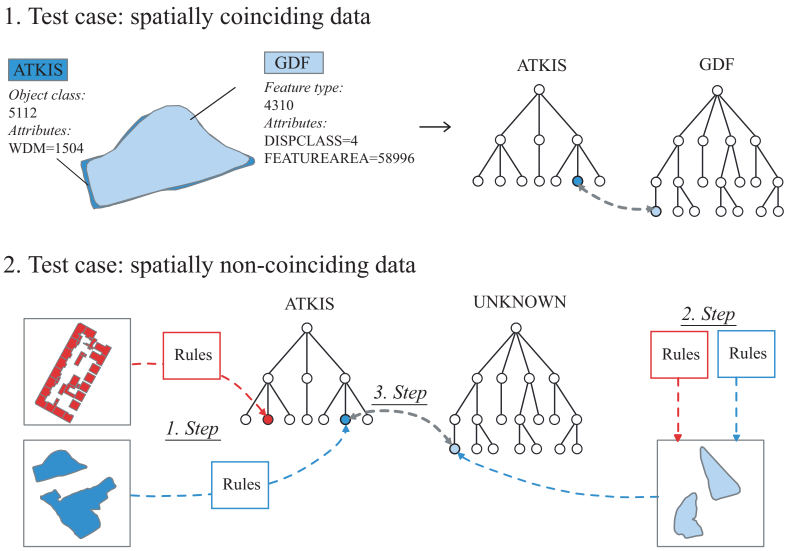
Therefore for our investigation we exploit the use of classification methods that is also known in the field of machine learning as supervised learning technique (learning from examples). The classification is a two step procedure. In the first step, a model is built to describe the set of data classes 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 [8]. Different methods to derive classification rules and represent them are known, but one of the most famous is the induction of decision trees, especially applying the ID3 [5] algorithm or its further developments (e.g. C4.5, C5.0) [6]. 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 to separate the samples into individual classes [2]. Entropy is a fundamental concept in information theory which means 'a measure of how much 'choice' is involved in the selection of an event' [10], i.e. higher entropy involves more choices and information, and is not good for classification. Therefore, the algorithms of decision trees are trying to discover the lowest entropy, less choices and information, recursively for the best classifications.

Sester [9] and Weindorf [13] already used geometric descriptions of spatial data for the derivation of classifications.

3 Our Approach: From Geometry to Semantics

In order to derive semantic correspondences between two geo-ontologies we use the spatial and/or geometric characteristics of single instances of the data sets. For the analysis two geodata sets in vector format describing topographic objects with similar resolution (approx. 1:25K) were used: on the one hand ATKIS data (the German Authoritative Topographic Cartographic Information System) and on the other hand data from TeleAtlas in GDF format (Geographic Data Files). The two data sets are modelled differently: whereas ATKIS uses a three level hierarchy, GDF only distinguishes two hierachical levels. In both cases a further distinction of the lowest class level using special attributes is applied. Another difference between both data sets is the very different granularity with respect either to the object classes and the number of individual instances. Whereas the topographic objects, like water or vegetation objects in ATKIS are modelled in much greater detail than the comparable, more aggregated objects of the GDF data, in turn the road and transportation objects of GDF data are modelled in more detail, because the data was specially developed for vehicle navigation purposes.

In the following sections the two test scenarios are presented in detail and are illustrated in Figure 1. In the first case the two data sets are available in the same geographical extent. The unknown semantic relationships are derived by means of a geometric overlay procedure, in which the spatial correspondences in combination with the geometric similarity are analysed. For the second scenario, we first derive classification rules for all object classes of one data set from the intrinsic characteristics of the objects. Subsequently similar objects in unknown data sets can be identified applying these rules.


1. Test case: spatially coinciding data

2. Test case: spatially non-coinciding data

Fig. 1. Two test scenarios. On the top the first test case using spatially coinciding data sets and at the bottom the second test case using spatially non-overlapping data sets.

3.1 Spatially Coinciding Data Sets: Analysis of Spatial and Geometrical Characteristics of the Objects

For the derivation of semantic relationships between different data sets that are given in the same spatial extent, in a first step identical objects have to be found.

For this purpose our method performs a geometrical overlay of the data. In the analysis we assume that the data sets are organized in layers each representing an object class. The instances within a layer are modelled in a tessellation, but instances of different layers within a data set may overlap each other. A typical example are administrative objects that often encompass larger areas and generally overlay all the other objects, we call them 'container objects'. Due to the data organization in layers a simple spatial intersection of the data sets without considering further characteristics returns more than one matching candidate to an object, which could cause difficulties for the further analysis. But not only the layer structure, also the geometric discrepancies at the object boundaries themselves, cause an increasing number of possible matching candidates, because adjacent objects may partially overlap. In order to reduce the number of matching candidates an exclusion of neighbouring or minimally overlapping objects is done with the consideration of the geometric criterion *area* during the analysis process. The overlay ratios R_i between the object area O_i and the intersection area I are calculated in both directions with $R_i = \frac{I \cdot 100\,\%}{O_i}$ with $i = 1, 2$. In Figure 2 the analysed 1:1-equivalence ($R_1 \geq 80\,\% \wedge R_2 \geq 80\,\%$) and 1:n-inclusion ($R_1 \geq 80\,\% \vee R_2 \geq 80\,\%$) instance relations are illustrated with a simplified schematic diagram.

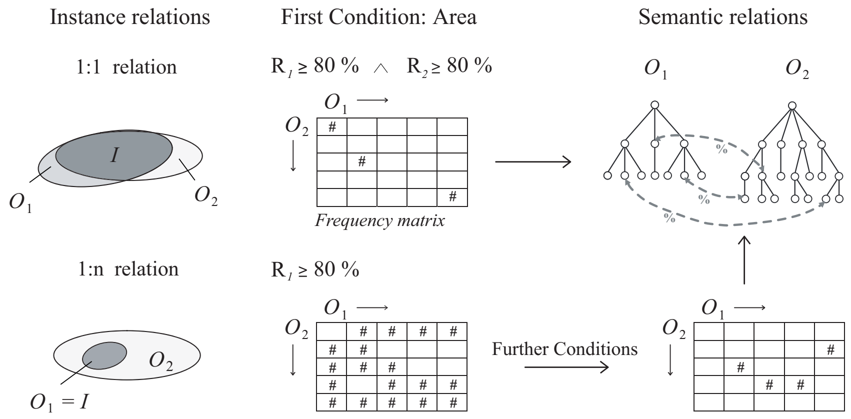


Fig. 2. Process of derivation of semantic relations between two geo-ontologies from instance relations basing on defined conditions concerning the geometry.

The result for each relation is set up in a frequency matrix, that contains the number of possible matching candidates, that fulfill the pre-defined conditions. In our example the search for 1:1 relations did not yield many candidates, because as described above GDF contains more aggregated objects. In contrast, in the frequency matrix of the 1:n instance relations also all relations to 'container objects' are included, as the 80% ratio between the object areas has to be met at least in one direction. But these results are not desired and for this reason further geometric conditions have to be used. For example by comparing of the mean widths or the elongations of two objects, one can exclude the relation to an 'container object'. Depending on the type and the strictness of the conditions chosen, reliable object relationships with information about the quality between differently modelled objects are detected. Analysing the final matching candidates within the individual frequency matrices yields to semantic relationships between the object classes in the ontologies. A strong indicator for semantic correspondences between object classes is probably if all kinds of instance relations do exist. Therefore the reliability of the existence of semantic correspondences can be derived from the relative frequency of the individual instance relations regarding the total number of instances of the object classes. For example, if in a simple case all instances of two object classes from different data sets meet the 1:1 instance relation condition, then a 1:1 relation between these two object classes can be assumed.

3.2 Spatially Non-coinciding Data Sets: Using Intrinsic Geometrical Characteristics

For the second scenario we relax the restriction of the geographical extent. Consequently the geometrical overlay procedure does not succeed in the first test case. For this reason we have to identify geometrically similar objects by exploiting the intrinsic characteristics of the objects themselves. Subsequently by means of these geometrical properties we can derive the semantic correspondences between object classes in different ontologies. Therefore we have to define generally valid description rules consisting of geometrical parameters from a training data set representing all the object classes present in one data set.

As a preprocessing step the following critical question has to be answered: Which shape descriptive parameters are necessary in order to describe the geometry of an object precisely and nevertheless objectively, and thus possess the potential to make a clear identification of particular types of objects possible? A fundamental problem is to model the subjective perception of objects as a function of these characteristics. To show the difficulties in describing the objects a few examples are displayed in Figure 3. Humans use the following properties to describe such object groups: buildings are of a small size and have generally parallel borders and right angles, whereas a road network consists of a netlike structure with mostly rectangular junctions. In turn, in most cases lakes and ponds do have a round, organic outline, whereas rivers show a lengthier but nonetheless organic shape with partially parallel borders and a mostly constant width. These natural language descriptions also contain a lot of vague and fuzzy descriptions, which make it difficult to transfer them to a computer. Still, it is not clear, if they are sufficient enough to seperate different object types clearly from each other.

After identifying the appropriate geometrical parameters, we compute the following parameters for all objects of both data sets: Object area, perimeter,

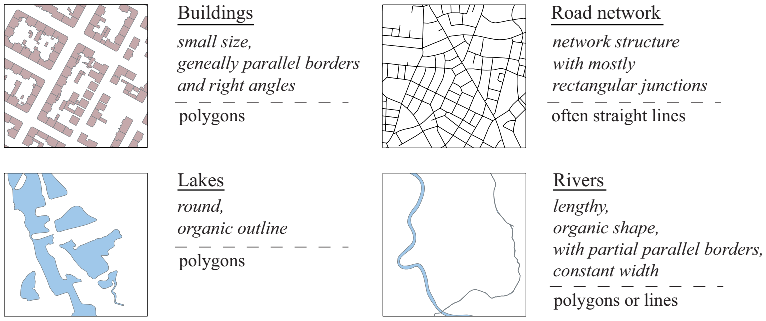


Fig. 3. Possible descriptions of a subjective perception of selected object groups.

compactness, rectangularity, mean width, the approximate elongation, the minimum bounding rectangle and also combinations of these parameters. This list of parameters is a start and not complete. As illustrated in Figure 1 in the first step we use these collected parameter values in combination with the individual object attributes (unique identifier and object class) in a classification process (e.g. decision tree) in order to set up description rules for every object class of the source data set. Subsequently objects in unknown data sets can be identified with these derived rules.

For the first tests we use this procedure for water objects from the already mentioned data sets: ATKIS and GDF. Therefore ATKIS is the source data set and consists of two objects classes (5101 and 5102) representing rivers with a total of 12 objects and one lake object class (5112) with 268 objects. We used the J48 algorithm as classification procedure, which is based on the C4.5 and ID3 algorithm. The source data set was partitioned into a training (70%) and a test (30%) set. As a result we got the following decision tree as illustrated in Figure 4.

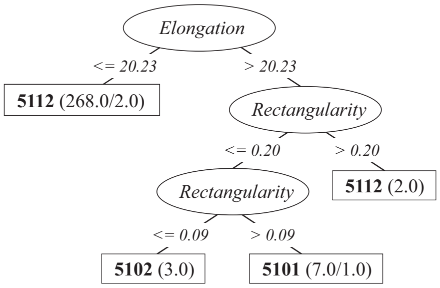


Fig. 4. Decision tree for water objects of the source data set: ATKIS.

The outcome of this are the two following simple description rules: $lake = \{obj|Elongation \leq 20.23 \vee Elongation > 20.23 \wedge Rectangularity > 0.20\}$ $river = \{obj|Elongation > 20.23 \wedge Rectangularity \leq 0.20\}$. 96% of the test objects were classified correctly and only 4% incorrectly. In Figure 5 the results of applying these rules to 647 water objects of the GDF data set are shown. In this case 92% were classified correctly and 8% were not correct. In this process mainly river objects were mis-identified as lake objects. Apart from the small percentage of the incorrectly classified objects, at a first glance the results seem to be very good.

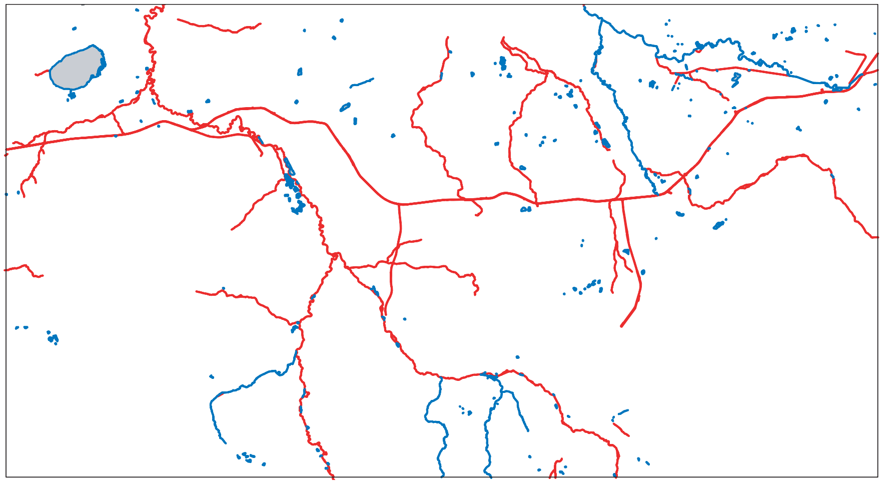


Fig. 5. The result of the application of the derived rules from ATKIS to the GDF data set. The red colored objects are classified as rivers and the blue ones as lakes.

To test the potential of the classification rules, we applied them to another object class of the GDF data set, namely vegetation. 76% of the objects are identified as lakes and 1% as river objects. This result, of course, is not satisfying and reveals that the rules and consequently the chosen attributes are not sufficient. The general idea is however, to use all classification rules derived for the objects of the first data set. They are applied to the objects of the second data set. In cases where an object of the second data set yields only one classification, the result is unique and the reliability of a correct classification is high. In cases in which there are several classifications, additional investigations are needed. We plan to extend the rules with additional criteria, like neighbourhood relations. For example buildings are generally situated near roads or water objects are generally surrounded by vegetation. Finally the last step is to establish links between the related objects classes of the ontologies.

4 Summary and Outlook

The paper describes ongoing work on semantic data integration. The need of integrating data sets of different origin and different granularity is evident, especially for data reuse. But the various data sources are quite different, starting from different data formats to insufficient documentation. For this reason we use the instances of the data themselves in order to infer the semantic correspondences between object classes in the ontologies, being the prerequisite for data integration. For determination of semantic relationships between different geo-ontologies we analyse two different scenarios. In the first approach we use two data sets in the same geographical extent with similar resolution. But how useful is this method, if the resolutions are different? This further case remains still open. In the second scenario we use data sets, that do not overlap. With classification procedures from Data Mining we set up description rules consisting of geometrical properties of the instances themselves in order to use these rules for the identification of objects in unknown data sets. Up to now the presented methods are in an experimental state and future work will be necessary with respect to the quality assessment of the derived semantic relationships. Particularly the methods have to be tested on different data sets from several sources.

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 3D Urban Geoinformation' in cooperation with China.

References

1. M. Duckham and M. Worboys. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Science*, 19(5):537–557, 2005.
2. J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., 2001.
3. M. Kokla. Guidelines on geographic ontology integration. In *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences (ISPRS Technical Commission II Symposium)*, volume XXXVI Part 2, pages 67–72, Vienna, 12. - 14. July 2006.
4. M. Kokla and M. Kavouras. Semantic information in geo-ontologies: Extraction, comparison, and reconciliation. 3534:125–142, 2005.
5. J.R. Quinlan. Induction of decision trees. In *Machine Learning*, volume 1, pages 81–106. Morgan Kaufmann Publishers Inc., 1986.
6. J.R. Quinlan. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, Califonia, USA, 1993.
7. M. Andrea Rodríguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003.

8. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach, 2/E.* Prentice Hall, 2003.
9. M. Sester. Knowledge acquisition for the automatic interpretation of spatial data. *International Journal of Geographical Information Science*, 14(1):1–24, 2000.
10. C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.
11. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
12. S. Volz. Data-driven matching of geospatial schemas. 3693:115–132, 2005.
13. M. Weindorf. *Regelbasierte Interpretation unstruktuierter Vektordaten.* PhD thesis, Fakultät für Bauingenieur- und Vermessungswesen der Universität Karlsruhe, 2002.