DERIVATION OF IMPLICIT INFORMATION FROM SPATIAL DATA SETS WITH DATA MINING

F. Heinzle*, M. Sester

Institute of Cartography and Geoinformatics, University of Hannover, Appelstr. 9a, 30167 Hannover, Germany - (frauke.heinzle, monika.sester)@ikg.uni-hannover.de

Working Group: TS WG IV/5

KEY WORDS: Spatial, Information, Retrieval, Metadata, Data Mining, GIS, Databases, Internet/Web

ABSTRACT:

Geographical data sets contain a huge amount of information about spatial phenomena. The exploitation of this knowledge with the aim to make it usable in an internet search engine is one of the goals of the EU-funded project SPIRIT. This project deals with spatially related information retrieval in the internet and the development of a search engine, which includes the spatial aspect of queries.
Existing metadata as provided by the standard ISO/DIS 19115 only give fractional information about the substantial content of a data set. Most of the time, the enrichment with metadata has to be done manually, which results in this information being present rarely. Further, the given metadata does not contain implicit information. This implicit information does not exist on the level of pure geographical features, but on the level of the relationships between the features, their extent, density, frequency, neighbourhood, uniqueness and more. This knowledge often is well known by humans with their background information, however it has to be made explicit for the computer.
The first part of the paper describes the automatic extraction of classical metadata from data sets. The second part describes concepts of information retrieval from geographical data sets. This part deals with the setup of rules to derive useful implicit information. We describe possible implementations of data mining algorithms.

1. INTRODUCTION

There is an imagination, a dream, that some day our computer would communicate with us in a meaningful way. Tim Berners Lee (2001) concretised this dream in the range of Internet with the formulation of the Semantic Web. The idea is to let the computer understand not only the words used by humans, but also the context of the expressions and their use in different situations.
Especially when using an Internet search engine, we are often confronted with the stupidity of the computer. Today most of the search engines conduct a query by looking up keywords and comparing them to a precompiled catalogue of all existing web sites. However, there is no analysis of the sense of a query or an interpretation of the combination of used words in web sites. The aim of building a Semantic Web deals with those questions. Linked to the idea of the Semantic Web is the EU-funded project SPIRIT (Jones et al., 2002). SPIRIT (Spatially-aware Information Retrieval on the Internet) is engaged in improving the concept of search engines by evaluating the spatial context of queries and web sites. The inclusion of the context and consideration of the semantic background improves the quality of the results. Often we use spatial concepts to describe something or we keep a spatial situation in mind, when we search for something. In SPIRIT we want to include those structures to define a spatial ontology.
A huge amount of information is stored in spatial data sets. However, usually these data sets are not accessible in the Internet. Most of the time there are neither metadata describing the datasets nor specifications of the intrinsic geometries and attributes. Furthermore these data sets contain a lot of implicit information.
The aim is to make spatial data sets visible in the Internet, especially to enable search engines to get knowledge about the data and publish it or use it in search queries. This requires the definition of metadata that are sufficient enough to describe the significant aspects of the data, but moreover it requires the development of algorithms, which will determine these metadata automatically. The second and more ambitious aim is to even make the contents usable for a search engine. This means to identify spatial phenomena in the data sets and to build a semantic network from implicit information in the data.
Both attempts are described in the following chapters. In section 3 we discuss the first issue, namely the automatic annotation of spatial data sets with a set of important metadata tags. Subsequently we present ideas for the extraction of implicit information to use it for spatial concepts and concentrate on data mining algorithms to derive this information.

2. RELATED WORK

The extraction of information from spatial data sets has been investigated in the domain of interpreting digital images. There, the need for interpretation is obvious, as the task is to automatically determine individual pixels or collections of pixels representing an object. Basic techniques for image interpretation are either pixel based classification methods (e.g. Lillesand and Kiefer, 1994) or structure based matching techniques (e.g. Schenk, 1999). The major applications in photogrammetry lie in the automatic extraction of topographic features like roads (Gerke et al., 2003), buildings (Brenner, 2000) or trees (Straub, 2003). The main challenge is to provide appropriate models for the objects to be found in the images.

These models are either given by hand or can also be acquired using machine learning approaches (Sester, 2000). The interpretation of vector data sets is a fairly new application. It has mainly been investigated in the context of spatial data mining (Koperski & Han, 1995).

3. METADATA DESCRIPTIONS OF SPATIAL DATA SETS

3.1 Metadata in SPIRIT

In metadata information about spatial data sets can be stored. Metadata are structured data to describe resources and to enable users or agents to select and assess the data. However, there are two major problems:
The expressiveness of metadata highly depends on the used scheme. Many existing schemes define the content more or less strictly. The ISO 19115 standard (ISO/TC-211, 2003) is designed especially for geographical data sets. The metadata used in SPIRIT are highly conforming to this existing international standard. However we identified a set of metatags, which are of essential importance for SPIRIT.
Secondly the enrichment with metadata still is a process, which has to be done manually for the most part. Although there are some tools supporting the data entry by using interfaces and predefined lists of terms, the costs of manpower and time input to enter the data are still almost insurmountable obstacles. This leads to the fact that only few web sites and information resources are enriched with metadata. For this reason tools to generate metadata automatically would be preferably. We will illustrate this ambition on the example of ArcView projects and shape files.

3.2 Automatic Extraction of metadata

For SPIRIT, we considered the following metatags as of high importance: name, spatial extent, keywords, contact and resolution. In this chapter we will illustrate the automatic extraction of metadata from ArcView shape files. Hereby of special relevance is the discovering of keywords regarding the stored spatial elements.
From ESRI shape format the following information can be extracted easily:

- minimum bounding box
- number of geometrical elements
- type of geometrical elements, like point, line, polygon
- information about the attributes and their structure, like name, type

That information is important for the interpretation of the geometrical aspect of a data set. Indeed it does not tell us many things about the semantics of the data. Particularly if the names of the predicates are coded by numbers or like in the abbreviated example given in table 1, the primary information of the shape files is insufficient.

SHAPE	AOBJID	TEIL	OBJART	OART_ATYP
PolyLine	N01CZ70	001	3102	3102
PolyLine	N01CZ1S	002	3105	3105_1301
PolyLine	N20LHCN	001	3106	3106

Table 1. ATKIS-record, Excerpt of the adequate dbf file

From this, it is not apparent, that this data represents a road network, which is displayed in figure 1.

At least it is necessary to know, which data are coded in the set to be able to provide an internet user the right information. Up to date we only know about the type of elements, for example there are lines, but we do not have knowledge whether the lines are streets, pipelines, administrative borderlines or contour lines. To detect this information, we analyse shape files and if there is a legend available, more information can be extracted from the ArcView project file to derive automatically adequate keywords. The following example documents the process.

Figure 1. Road network data set

In figure 2 the automatically extracted metadata are shown.

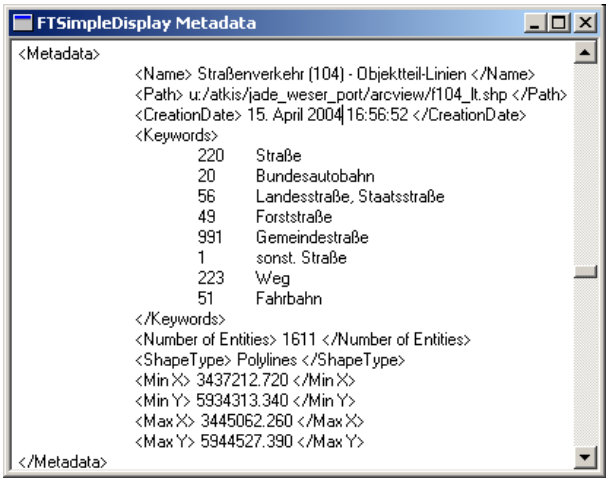

Figure 2. Metadata for the displayed road network data set, distinguishing different types of road (in German)

All available data are analysed to acquire the keywords. Text files are checked to identify street names and designations of regions. Captions often give a glimpse of the character of the stored geographical elements, as well as the names of the attributes in the dbf files.
The spatial extent of the data set is determined by the minimum bounding box. Moreover there are also some indicators to infer the scale or the level of detail of the data set. Analysing only the geometry of features, a simple measure for the scale of a data set can be the distance between the individual points a line or a polygon is composed of. Furthermore, the existence and type of certain geographic elements also give rise to a certain resolution, e.g. typically buildings are only present in large scales; in large scales roads are typically represented as areal objects whereas in small scales they are given in forms of polylines.

Important information can be anticipated already by means of the keywords. However there is still no knowledge about the distribution and location of the geometrical elements, their connections to each other, their accumulation in special places and so on. Those characteristics make the information of a data set complete and allow humans to interpret data. This is the ambition of the next step, namely to extract implicit information from data sets and making them visible in the internet, especially for search engines.

4. EXTRACTION OF IMPLICIT KNOWLEDGE WITH DATA MINING

As mentioned in the above chapter especially the keywords are a first approach to get some semantic information. However these keywords have a big drawback. They are still interpretable only by human beings. Still expressions like "Autobahn", "Aérogare" or "Hospital" are characterless to the computer. We would need a translation in two respects: first a language translation, but moreover a semantic translation. Those catalogues, which describe the meaning of a word and determine its sense depending on the context, are called ontologies.

To enrich the ontology our ambition is focused on teaching the computer to learn spatial concepts and to combine knowledge to higher concepts automatically. They are hidden in the spatial data, less to find on the level of pure geometry, but rather inherent by the combination and interaction of the spatial elements. Spatial data mining is the approach to extract those implicit information.

Needless to say, upon finding those implicit spatial structures still the computer does not know the meaning of "Autobahn". However the concept is learnt, that "Autobahn" is a major road (which has own concepts as well), has less junction points and is situated rarely inside of settlement areas, but rather in peripheral areas.

Next we will introduce those implicit structures and concepts, which could be useful for a search engine. Afterwards we will describe procedures and algorithms to discover inherent information with data mining and will document first approaches and results.

4.1 Implicit Data

As Aristoteles put it: the whole is more than the sum of its parts, the content of a spatial data set is more than only the pure geometry. Cognitive structures of human beings fit to the world, because they were formed by adaptation to the world. Up to now computers do not have this semantic knowledge of the world. The challenge is to reproduce such an adaptation process by learning automatically.

Considering typical queries to a search engine and user scenarios with spatial background, there is a lot of helpful information stored in data sets. E.g. a user would like to search for a hotel in the centre of the city, at least the search engine has to know, where the city centre is located. This knowledge can be discovered in vector data, but it is usually not explicitly stored in an item.

In figure 3 you can see topographic elements of a small village, like roads and houses. However, this is already an interpretation by humans. You have to be aware, that actually you just can spot some lines and polygons, which are differently coloured. That is the prior information the computer is able to get out of the data.

Figure 3. Where is the city centre located?

Indeed we recognise streets and houses and we are able to reason further facts. Humans can locate the church by the special shape of this building. The interaction of the streets and houses and their concentration induces at least the information, that it is a village. We also can identify larger buildings in the upper part and distinguish them from smaller ones in the south. A computer can calculate these facts too. The big challenge is the following reasoning process. Humans interpret the larger buildings as the inner part of the village, because they know about old farmyards and the typical formation of a village (in Germany). The smaller buildings represent a colony of one-family houses. We are able to locate the main street leading through the village as well, because of the structure of the settlement. Therefore humans can detect the city centre approximately without difficulty.

There is a plenty of examples and ideas, which would be useful in SPIRIT. At least we would like to concentrate on some concepts mentioned below:

- classification of more or less important cities
- sphere of influence of cities
- detection of the centre of a city
- determination of tourist areas and attractive destinations
- possibilities of suburban or industrial settlement, urban development, quality of housing

The information available in the data set, which we consider to exploit in those concepts together with the necessary operations to extract and combine the information is described in Heinzle et al. (2002).

Some characteristics of the elements can be determined with simple GIS functionality like to calculate an area/size or to count the existence of special objects. The evaluation of other properties, like density, distribution or neighbourhood, is more complicated. The analysis of distances is an essential part to get knowledge of these aspects. However, the handling of threshold values or absolute numbers is less helpful, because it depends on the context, if an attribute or a characteristic is really specific and outstanding. Most of the time those values are of interest and shed light on something, which distinguish themselves and excel at special properties in contrast to the rest of the data. Clustering algorithms can be used to identify groups of elements respectively their neighbourhood. Among clustering algorithms those are preferable that do not need threshold values (Anders, 2003).

Moreover the combination of properties and their calculated values raise a problem. Logic operations have to be extended by weighting and quantifiers, which depend on the importance, relevance, quality of the attribute values and significance of elements.

4.2 Automatic Derivation of Implicit Data

As mentioned above there are rules implicit in spatial data, however there are two different ways of approaching the goal of extraction of implicit knowledge. These two kinds of extraction models are on the one hand to define the rules a priori (association rules) and to apply them to the data, on the other hand to let the computer find the rules by itself by exploring the data. Both ways lead to more knowledge, but in the first case it is knowledge, which we were especially searching for, like the concepts of chapter 3.1. The second case brings up unknown knowledge or inherent information, which may be useful to learn more about the data set, but can be not useful as well. Both methods are usually known as data mining (Witten and Frank, 2000) and will be described and examined. They are discerned into supervised and unsupervised classification.

4.2.1 Supervised Classification: implies knowledge discovery on the basis of predetermined models respectively spatial association rules. Supervised classification starts from a set of classified examples for a concept to be learnt. From this set classification schemes for the concepts are derived, e.g. using machine learning approaches (Michalski et al., 1998), or also Maximum Likelihood classification (Lillesand and Kiefer, 1994). In principle every kind of knowledge representation can be used to form a classification scheme, especially rule-based systems or semantic networks. We will depict the process by the help of decision trees. Every branch symbolises the existence of a distinctive classification feature. Depending on the result of the inquiry the adequate branch will be followed further. In the end the model leads to a classification into different categories of one issue. However the scheme includes some essential problems. The sequence of the validation of a distinctive classification feature is one determining factor. The use of such a step by step algorithm without the possibility to go backwards holds the endangerment of abandoning important elements or a proper solution at an early stage. The determination of thresholds respectively stop criterions can lead to problems. Therefore the need of high quantitative and qualitative data is necessary to be able to calibrate the model.
The concepts of "the centre of a city" can be implemented by using such supervised methods. For example, we could determine a point as a city centre, if it fulfils following conditions:

- major streets will meet in the centre
- the buildings in the centre are larger in comparison to areas outside
- non-existence of industrial areas
- etc. etc.

The weak point of such specifications can easily be recognised:

- the descriptions are given in natural language, which is not directly usable by a computer
- the specifications are vague
- not all conditions might be needed in all cases
- some conditions can be more important, some less important
- there is no guarantee, that the model composed by humans is accurate, proper and especially complete
- possibly there are much more criteria, which we have ignored and did not take into account. On the contrary, we could have included distinctive features, which do not correspond to the reality, and have only been valid for a small test data set.

Basically we expect to retrieve a special information as a result of predefined inputs. However, the classification model will fail, if the perceptions will not agree with reality. The above mentioned difficulty of combining the criteria and their values is already hidden in the scheme. In the case of inadequate combination and insufficient provision of characteristics misinformation will be generated. On the other hand the quality of deliberately formed models depends highly on the human creativity and ability to reason. Spatial phenomena and relationships have to be recognised by humans a priori to implement them into a supervised classification algorithm.
This implies, that the setup of such models has to be done very carefully, possibly using large test data sets in order to gain the information from and to perform tests for verification of the derived rules. Furthermore, a specific inference scheme has to be designed to apply the rules to the data, that takes the probability or the importance of a condition to a rule into account.

4.2.2 Unsupervised Classification: The method aims at leaving the process of knowledge discovery to the computer itself. That means the computer has to discover rules, separations into categories, similarities in data sets without any predefined restrictions. Koperski & Han, 1995, describe an approach, where spatial associations between objects have been analysed automatically leading to the derivation of a rule stating that "all large cities lie close to a river". Since such rules are induced from a finite set of examples, they cannot be verified, but only falsified. Thus, there has to be a validation of the utility of the detected information. It may happen, that rules will be found, which are obvious and do not give further knowledge. It is another process of learning to distinguish useful and non-useful rules.
One form of Data Mining is clustering in order to find regularities or similarities in data sets. We used it for the following investigation:
A way to analyze geometric objects is to determine their characteristics and to try to find regularities among them. Such regularities then, in turn, can be considered as representatives for a certain class of objects or a class of objects in a certain context or environment. For linear objects or even networks of linear objects the nodes are such a characteristic, including the node degree, i.e. the number of outgoing lines from the node. Furthermore, also the angles of the outgoing lines can be important. Different types of nodes can be distinguished and classified, as shown in figure 4:

Figure 4. Different node types

We made some investigations analyzing the node types of linear networks.
Three examples will point out the process:
1. While investigating the concept of the city centre with supervised models, we introduced the criteria of crossroads in the centre. A crossroad is a node with at least four outgoing lines, which were expected primarily in the city centre, as there many roads come together. The tests turned out in an unexpected result of this investigation. The condition to find crossroads in the city centre depends on the size of the town. There seems to be a rule regarding the relation between the structure of the centre, the spatial arrangement of streets and the size of the city.
In figure 5 typical structures in the city centre are shown, depending on the dimension of the town. In small towns often a big street leads through and mainly TEE-junctions can be

found, whereas in medium size cities the expected structure dominate, meaning that three or more major roads will meet in the city centre. In large cities the opposite trend can be observed: major streets will run around the city but in the centre itself only minor streets or even pedestrian areas will be located.
This could be one useful information for setting up rules, which can be found automatically with data mining mechanism.

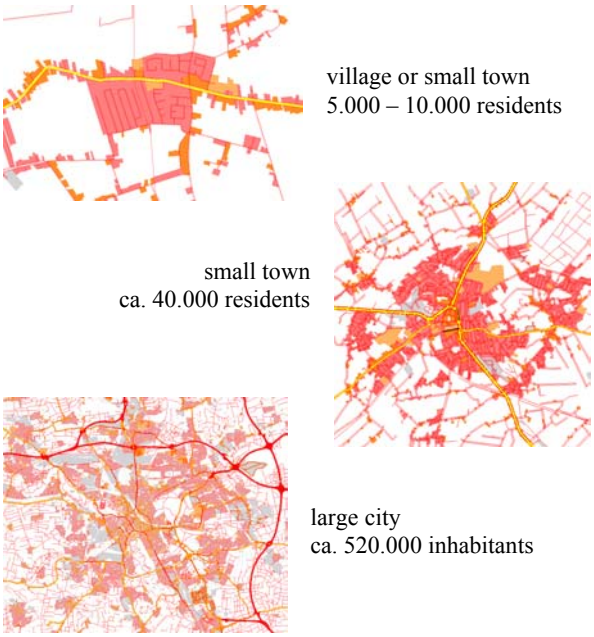

Figure 5. Typical arrangements of streets in the city centre depending on the dimension of the town.

2. Junctions of roads have been investigated regarding the existence of nodes with four outgoing lines. The intention was to look into detail, if there are reasoning mechanisms to cut settlement areas into partitions, especially if the lines will meet approximately orthogonal (CRS type).
Among other things it came up, that highways will be represented by separate clusters with solely one edge (ELL-junctions), with the exception of the access roads. Naturally there are only a few intersections with highways, the parts between are direct polylines without branches.
As shown in figure 6 it could be one of a criteria to determine highways respectively to distinguish their access roads from their carriage-ways in data sets. It can be very helpful to validate further structures like the neighbourhood of settlement areas in the vicinity of an highway access.

Figure 6. All red lines are “one edge cluster”, the highway is easy to locate in the middle.

The analysis of junction or node types can also help to distinguish between different features on a geometric level: when looking at different linear networks, it gets clear, that certain junction types only occur with certain objects – or do not or only rarely occur with certain objects (figure 7). E.g. the 4-junction mentioned above mainly can be found in road networks – and hardly ever in river networks, as in nature it is very rare, that four streams will meet in the same place. Another extreme example are lines which typically do not intersect at all or only at (very rare) saddle points.
It does not lead to new knowledge, but to new information to the computer. This investigation can shed light on the content of a data set, especially which line elements belong to the road network. In this context the obvious rule can turn into a very helpful information.

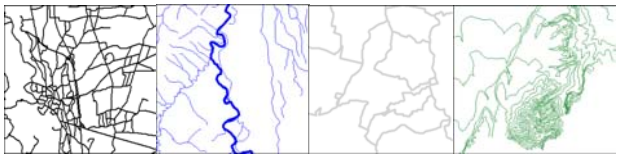

Figure 7. Appearance of different line elements: roads, rivers, administrative boundaries, contour lines

Furthermore the investigation into the nodes with four outgoing lines led to following conclusion regarding a partitioning: these CRS nodes can be of a separating nature, especially along the major roads. It is similar to a Voronoi diagram, which here, however, does not exist on the basis of geometric distance, but rather on the topographic detail of the intersection of four lines. Figure 8 documents the results of the analysis in two different data sets, one French and one German data set. Especially along the major roads the data set is segmented into different partitions. In the figure on the left side you can see, that the data set is split in two main sections each on the left and on the right side of the picture. In the middle a valley with major roads and a town is located.

Figure 8. Clustering of road networks by analyzing the CRS nodes. Left: French data set. Right: German data set.

Other measures we are going to investigate is the “straightness” of a linear object, i.e. a collection of polylines that can be traversed more or less straightly. A method to derive these so-called strokes is described in Thomson and Richardson (1999) and has been used for network generalization and classification by Elias (2002).
On the basis of described processes we are able to examine the data sets and the above mentioned results in more detail, whereby supervised and unsupervised models can hardly be kept apart at this stage. The following factors could be decisive for further (supervised and unsupervised) interpretations: size of a single mesh, length of segments between nodes, frequency of

occurrence of high-order respectively significant nodes, in-depth study of the type/shape of the nodes. To get on without predefinition of thresholds or the preliminary fixing of minimal and maximal values it is our aim to continue the search for broader regularities, for example in the combination between above mentioned criteria.

Regarding this we could imagine many hypotheses and to find the following or similar structures with data mining:

- capitals are always located at large rivers?
- in general all big cities are located at large rivers?
- in the city centre are larger buildings than in outskirts?
- in tourist areas are more bicycle tracks than in non-tourist areas?
- industrial areas are situated mostly along big traffic routes?
- winding roads are always in regions with heavy differences in elevation?
- villages are embedded mostly in agricultural crop land, very rare they are located in forest?
- 90 per cent of all junctions of traffic lines are situated in settlement areas?

We will concentrate on both ways, supervised and unsupervised methods. Both can support knowledge discovery and during the implementation of algorithms, both data mining models will influence each other.

5. CONCLUSIONS

The paper presented attempts in the range of spatial data mining in the context of realising a spatially aware search engine. To solve spatially related queries, the computer has to be aware of semantic aspects. Ontologies are used to represent them. However the information therefore can not completely be acquired manually. Automatic detection and learning processes of the computer are essential to enrich such data collection.
Classical metadata are a first approach to reveal the content of a data set. Our intention is to extract metadata automatically from geographical data sets. An automatic enrichment with specific metadata, e.g. the keywords, was presented.
Further steps are necessary to make semantic of geographical data visible, so that the computer receives background knowledge and can perform logical reasoning procedures. Therefore we use and implement data mining methods. In this article concepts and first attempts were introduced and explained, which have emerged as main focus during our investigations. First algorithms were developed and realised.

6. REFERENCES

Anders, K.-H., 2003. Parameterfreies hierarchisches Graph-Clustering Verfahren zur Interpretation raumbezogener Daten. Dissertation Universität Stuttgart.

Berners-Lee, Tim, James Hendler & Ora Lassila, 2001. The Semantic Web. *Scientific American*.

Brenner, C., 2000. Towards fully automatic generation of city models. *Proceedings of the XIXth ISPRS Congress*, volume XXXIII-B3 of *International Archives of Photogrammetry and Remote Sensing*, pages 85-92, Amsterdam.

Elias, B., 2002. Automatic Derivation of Location Maps. IAPRS Vol. 34, Part 4 "Geospatial Theory, Processing and Applications", Ottawa, Canada.

Gerke, M., Butenuth, M., Heipke, C., 2003. Automated Update of Road Databases Using Aerial Imagery and Road Construction Data, IntArchPhRS, Munich, Vol.XXXIV Part 3/W8, pp 99-104.

Heinzle, F., Kopczynski, M., Sester, M., 2002. Spatial Data Interpretation for the Intelligent Access to Spatial Information in the Internet. *Proceedings of 21st International Cartographic Conference*, Durban, South Africa.

ISO/TC-211, 2003. Text for FDIS 19115 Geogaphic information - Metadata. Final Draft Version. International Organization for Standardization.

Jones, C.B. et al., 2002. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project". *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland pp.387 – 388.

Koperski, K. & Han, J., 1995. Discovery of Spatial Association Rules in Geographic Information Databases, *in:* M. J. Egenhofer & J. R. Herring, eds, 'Advances in Spatial Databases '95', Vol. 951 of *Lecture Notes in Computer Science*, Springer Verlag, Heidelberg, pp. 47–66.

Lillesand, T.M., Kiefer, R.W., 1994. *Remote Sensing and Image Interpretation.* John Wiley & Sons, Inc. New York - Chichester - Brisbane - Toronto - Singapore.

Michalski, R.S., Bratko, I., Kubat, M., 1998. *Machine Learning & Data Mining: Methods and Applications*. John Wiley, England.

Schenk, T., 1999. *Digital Photogrammetry, Volume 1*. Terra Science, Laurelville. 428 pages.

Sester, M., 2000. Knowledge acquisition for the automatic interpretation of spatial data. *International Journal of Geographic Information Science*. Vol. 14, No. 1, pp. 1-24.

Straub, B.-M., 2003. Automatic Extraction of Trees from Aerial Images and Surface Models. ISPRS Conference on Photogrammetric Image Analysis, September 17-19, Munic, Germany, IntArchPhRS, Vol. XXXIV, Part 3/W8, pp. 157-164.

Thomson, R., Richardson, D., 1999. The 'good continuation' principle of perceptual organization applied to the generalization of road networks. *Proceedings ICA '99, Ottawa, Canada, Session 47B.*

Witten, I.H., Frank, E., 2000. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.

7. ACKNOWLEDGEMENT

This work is supported by the EU in the IST-programme Number 2001-35047. We thank the National Mapping Agency of the state Lower Saxony in Germany (LGN) for providing the ATKIS data and the National Mapping Agency of France (IGN) for providing several topographic data sets.