EXPLORATION OF THE ROAD DATABASE FOR NAVIGATION SYSTEMS

Lichun SUI, Liqiu MENG

Department of Cartography, Technical University of Munich
Arcisstr. 21, D-80333 Munich, Germany
sui@bv.tum.de, meng@bv.tum.de

Commission II, WG II/6

KEY WORDS: Data mining, Database, Navigation system, Information gain, Rough set modelling, Attribute dependence

ABSTRACT:

Some new methods for analyzing geo-referenced statistical data are presented in this paper. These methods have combined the techniques of exploratory data analysis with algorithms for data mining. They have been integrated in a prototype software system developed at the Technical University of Munich in cooperation with Navigation Technologies (NavTech) GmbH. The system serves the purpose of value-adding the road database maintained by NavTech. In the original database, each road element is described by more than 110 attributes. A number of algorithms on the basis of information gain and rough-set modeling have been implemented to rank the individual attributes and detect the dependencies among attributes based on their values in an arbitrarily selected region. Other algorithms are developed on the basis of road geometry and devoted to the quantitative description of spatial patterns such as routes and urban structures. With the knowledge of relative importance of the individual attributes, users are given the flexibility to buy a local road database with truncated attribute list. By observing the ranking list and correlation matrix calculated for different regions, information that reflects the regional differences of a road network can be extracted. Likewise, the changes in ranking list and correlation matrix of the same region after removing or adding a route imply the relative importance of this particular route.

1. INTRODUCTION

Spatial data have their intrinsic flexibility to be directly downloaded upon ordering. Since years, geo-data suppliers have been concentrating themselves on the tasks of constructing portal sites to attract buyers, filling *holes* and removing redundancy in their *data warehouses*, updating and versioning the data items, developing compression methods as well as data structures for efficient transmission and so on.

The effectiveness of spatial queries is strongly influenced by the accessibility and transparency of the available spatial data warehouses on the Internet. The accessibility requires that (1) the database as a whole be well-tagged with a summary containing the relevant key words; and (2) the individual data items be explicitly indexed with attributes and metadata. The transparency requires further methods to (1) discover the spatial concepts that are otherwise hidden in the database; and (2) describe the discovered concepts using an easily understandable language. An accessible and transparent database allows flexible aggregation and segregation, hence the personal division of the information space. However, personalizing large data inventories is complex and unintuitive. Spatial data suppliers would go insane trying to determine what to offer to whom, especially when they themselves have lost an overview of their own databases. Therefore, such tasks should be performed by automatic spatial data mining systems (Meng，2001).

An appropriate visual representation of spatial data, such as a map, can be isomorphic to space and thus capable of preserving all spatial relationships. This representation is only perceivable by human eyes and can therefore be used by human analysts only. Although the human eye can immediately grasp most spatial properties and relationships properly reflected in a map, our analytical capabilities are very limited in terms of the volume of data and the complexity of the relevant information that may be hidden in the data (Andrienko et al. 2001).

The importance of data analysis has been widely recognized in statistics. The goal of exploratory data analysis is to enhance understanding and seek information from hitherto untouched or insufficiently understood data. An important category of data dealt with in statistics is the category of spatially referenced data. Many statisticians who developed techniques for data analysis have been concerned about proper ways of visualizing such data. In data analysis, it is the task of the human analysts to uncover important characteristics of the data. Data mining methods are developed to automate such a cognitively effortful process.

Our paper is organized as follows: In the next section, the tasks and goals of data analysis with data mining methods are introduced, while the following section describes the selected data mining methods based on the information gain and rough-set algorithm. In the selection 3 we use examples to demonstrate our implemented programs and our developed system. In all the examples we use the test data of the city Munich of Bavaria.

2. TASKS AND GOALS

Our first task lies in the adjustment and optimization of the attributes describing road objects. The voluminous and extensive attributes that already exist in current NAVTECH's road database are resorted according to their relative importance. This task represents primarily the interests of a data supplier who needs a continuous improvement of the insight in his data and an awareness of reducing the expenses of data acquisition and maintenance in the long run.

Another task is the possible quantification of the road structures, which can be examined on different hierarchical levels. The derived measures can be used to quantify and differentiate complex structures (e.g. the overall road clusters representing urban areas).

In our experiments we used three of the most widely used data mining methods:

- The method of objects analysis with *information gain;*
- A method that estimates the relevance of given attributes in predicting the values of an independent attribute (*the rough sets method*); and
- The neural nets.

These methods were chosen from the large variety of existing data mining methods for their relative simplicity and their direct applicability to the tasks described in this paper. So far we have finished testing the first two methods (information gain and rough-set modeling). The application of neural nets is still under investigation, therefore, will not be reported in this paper.

3. DATA ANALYSIS WITH DATA MINING METHODS

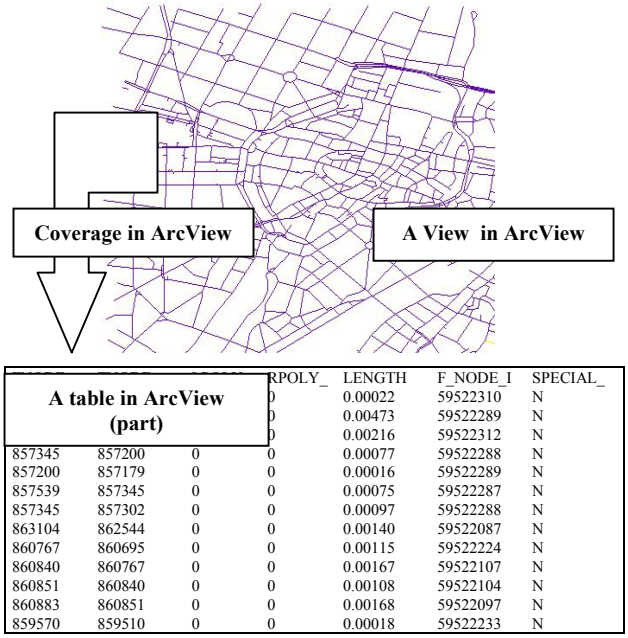
3.1 Data Pre-processing

Our test data are represented in ARC/INFO Export format, which can be transferred for the purpose of analysis. There is a function for the exchange of instruction in the ArcView tool kit. With this function, the data from other ARC/INFO platforms can be read and converted. After the test data have been installed on the platform and the data exchange has been completed, we are able to process the data in the ArcView environment (see Jacobi 1999, Liebig 1999).

For our work of data analysis, all programs begin with an ArcView table. Therefore, we have first selected a subset from the test data and generated an ArcView table. Fig.1 shows this procedure of the data pre-processing. It can be described as following:

- Open a theme or a project in ArcView
- Select a sub-view or an excerpt (sub-theme) from this theme
- Convert this sub-theme to a Shape file
- Add this theme to the view
- Open the theme table to work

A road record in our NAVTECH's data set contains 110 items. Many of them are blank or not defined for a given road segment. There are also many attributes that are not directly relevant for the exploration of the road database with data mining algorithms. So in our work, 24 carefully selected attributes are used for the test and examination. Other attributes can, of course, also be tested according to the same principle. These 24 attributes are tested with our programs of Visual C++ and Avenue through ArcView table.



			RPOLY_	LENGTH	F_NODE_I	SPECIAL_
			0	0.00022	59522310	N
			0	0.00473	59522289	N
			0	0.00216	59522312	N
857345	857200	0	0	0.00077	59522288	N
857200	857179	0	0	0.00016	59522289	N
857539	857345	0	0	0.00075	59522287	N
857345	857302	0	0	0.00097	59522288	N
863104	862544	0	0	0.00140	59522087	N
860767	860695	0	0	0.00115	59522224	N
860840	860767	0	0	0.00167	59522107	N
860851	860840	0	0	0.00108	59522104	N
860883	860851	0	0	0.00168	59522097	N
859570	859510	0	0	0.00018	59522233	N

Figure.1: Data pre-processing with a table in ArcView

3.2 Objects Analysis with Information Gains

According to the terms applied in data mining, classification of a given object means the prediction of the value of an attribute based on a set of its known attribute values. The goal of the algorithm is to find a nearly optimal order of tests in order to

construct a classification tree with possibly pure leaves. Every attribute in the data sets carries its information. In the formation of a group, different attributes, such as road *lane_number*, road *function_class*, road *speed_class* etc. in our data, play different roles. The goal of the introduction of data mining methods is to use all useful attributes to classify and identify the road objects. At first, the relative importance degree of each attribute must be determined.

The information gain algorithm is mainly applied to calculate the *relative importance* of individual attributes that describe the road segments. The relative importance is expressed by the quantitative measure "*Information gain*". For a given context, i.e. a road net in a particular landscape or a country, all attributes can be ranked in a declining order. This ranking varies from one context to another, which implies the necessity of context-oriented maintenance of road data base.

In order to solve more extensive induction problems, the information gain algorithm provides a strategy to form exactly a "*good*" *decision-tree* with less computing time. A method of criterion selection using two assumptions can serve this purpose. This method is based on information theory. For a set of objects ***U*** that contains ***p*** objects of the class ***P*** and ***n*** objects of the class ***N***, the assumptions are (see Bissantz, N. and J. Hagedorn 1996 and Meng 1998):

A randomly selected object belongs to the class ***P*** with a probability $p/(p+n)$ and to the class ***N*** with probability $n/(p+n)$.

If a decision-tree is used for the classification of an object, an identified class (***P*** or ***N***) is the result. For the decision whether ***P*** or ***N*** is available, an information ***I*** can be calculated with

$$I(p,n) = -\frac{p}{(p+n)}\log_2\frac{p}{(p+n)} - \frac{n}{(p+n)}\log_2\frac{n}{(p+n)}$$

With the attribute ***A*** and its values $(A_1, A_2, \ldots, A_v)$ as the root of a tree, the object ***U*** is subdivided into the corresponding partitions. The total information content of the tree with the attribute ***A*** as the root can be represented by the weighted average

$$Entropie \quad E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

where ***v*** is the total number of attributes.

The information gain ***G*** is defined as

$$G(A) = I(p,n) - E(A)$$

where **E(A)** is the Entropy.

We tested the 24 selected attributes and ranked them with their information gain.

Fig. 2 represents the information gains of a section of city Munich with a diagram. The vertical axis (information gains) represents the degree of importance, e.g., the value ***1*** represents the most important attribute and value ***0*** represents the least important attribute. For example, attribute FUNC_CLASS (Functional class, determine a route for a traveler.) possesses a value of the information gain 0,61 in the diagram and attribute AR_EMER_VEH (Access restrictions – emergency vehicles) has a value of the 0,21 (Sui /Yu 2001). Under this circumstance, it can be understood that the attribute FUNC_CLASS is more important than the attribute AR_EMER_VEH.

The relative importance of the attributes in this test region provides the data supplies with an index indicating the updating priority of the attributes.

Fig. 3. shows the result of the road objects that are grouped based on the information gains. We have chosen the downtown area of city Munich that contains 6351 objects, i.e. road lines. The blue lines (or thick black lines in case of black-white-image) have the higher information gains and thus the higher priority for maintenance than the thin black lines. The number of groups is a free choice of the user. For example, the user can choose to have three or more classes.

In the grouping of objects (here road lines) with this method, all attributes are introduced in accordance with their relative importance. Attributes with higher information gains have the contributions with greater weights.

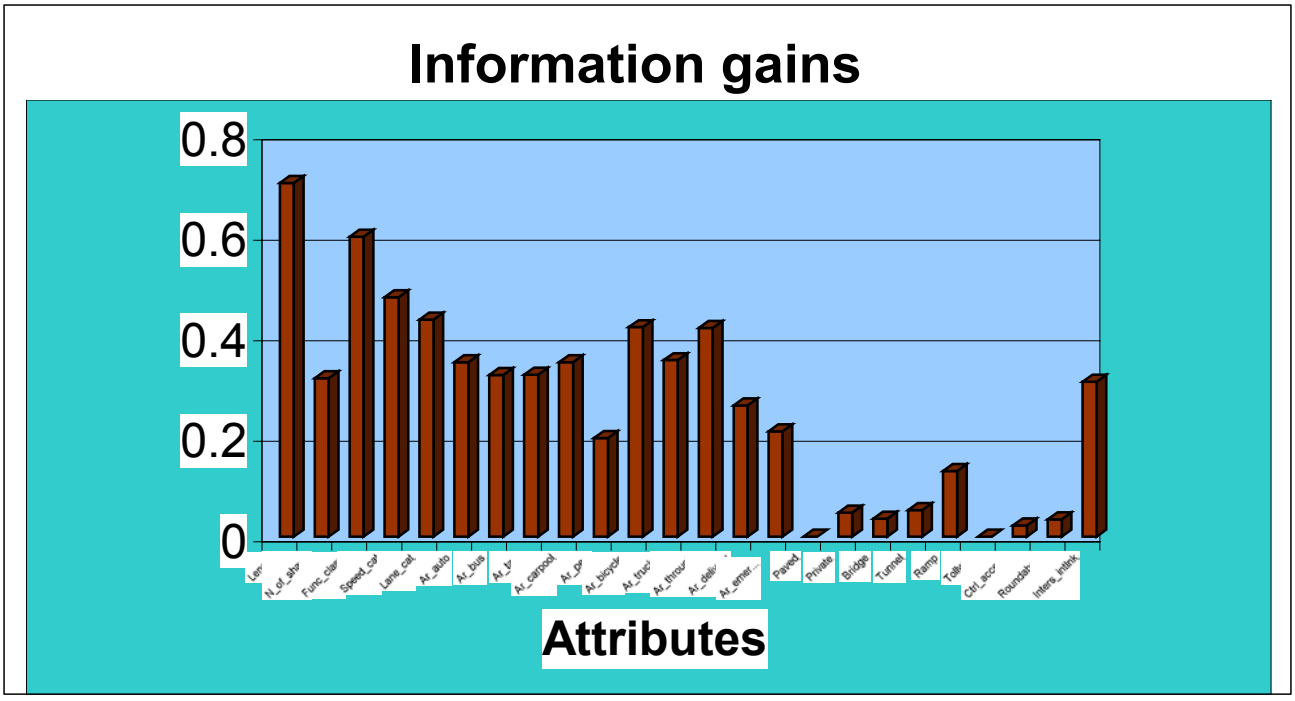


Figure 2. Information gains of the chosen attributes

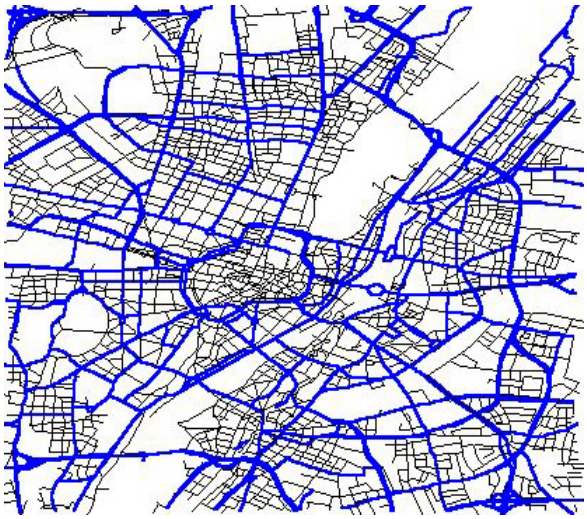


Figure 3. Grouping of the objects according to weighted information gains

3.3 Object Analysis with the Rough Sets Algorithm

Rough set modeling technique has its root in rough set theory. A rough set is an extension of the standard mathematical set. It uses a pair of the standard sets, the lower approximation S_- and the upper approximation S^+, to represent uncertainty and vagueness in a database. The difference between upper and lower approximation S^+ - S_- is called a boundary region, or the area of uncertainty of the rough set (Meng 1998).

Assume that an information system is defined as a pair (U, A) where U is a finite set of objects (interpreted as, e.g., cases, states, processes, subjects, or observations) and A is a finite set of condition attributes. Every attribute $a \in A$ is associated with a set of its values V_a. Each condition attribute ***a*** determines a relation $R_a : U \rightarrow V_a$. In practice, we are mostly interested in discovering dependencies in a so-called decision table. It is a pair $(U, A \cup \{d\})$, where $d \notin A$ is a distinguished decision attribute. Each value of ***d*** corresponds to a decision class. That is, ***d*** determines the partition of ***U*** into ***k*** disjoint decision classes $(D_1, D_2, ..., D_k)$ (where k is the number of different values of ***d***).

In addition to the similar function of ranking the attributes according to relative importance, the rough-set algorithm is firstly used here for the purpose of analyzing attribute dependences.

The methodological basis of the rough-sets (coarse set) is the observation that a coarsening of the illustration accuracy can lead to a better model recognition. Examples of application of the rough-set theory are the optimization of decision tables, the control generation in expert systems and the design of logical counters, also the machine learning, specifically the learning from examples.

The dependence of attributes is the other way of showing their importance. If an attribute is completely dependent with other attributes, it means that its dependence is ***1,0***, so we say that this attribute is of the least importance. That is, this attribute can be completely replaced with other attributes or their combination. In this sense, the dependence of attributes can be applied as a criterion to group road elements.

So far, the attributes can be pair wise compared with our program and their correlation is expressed by a real number between ***0.0*** (independent) and ***1.0*** (dependent). **Fig. 4** represents a diagram of the dependence for the same data set shown in **Fig. 5**. The y-axis illustrates the importance value using the dependence or independence of the attributes, i.e., the value ***1.0*** represents the most independent attribute or most important and value ***0.0*** represents the least independent or least important attribute.

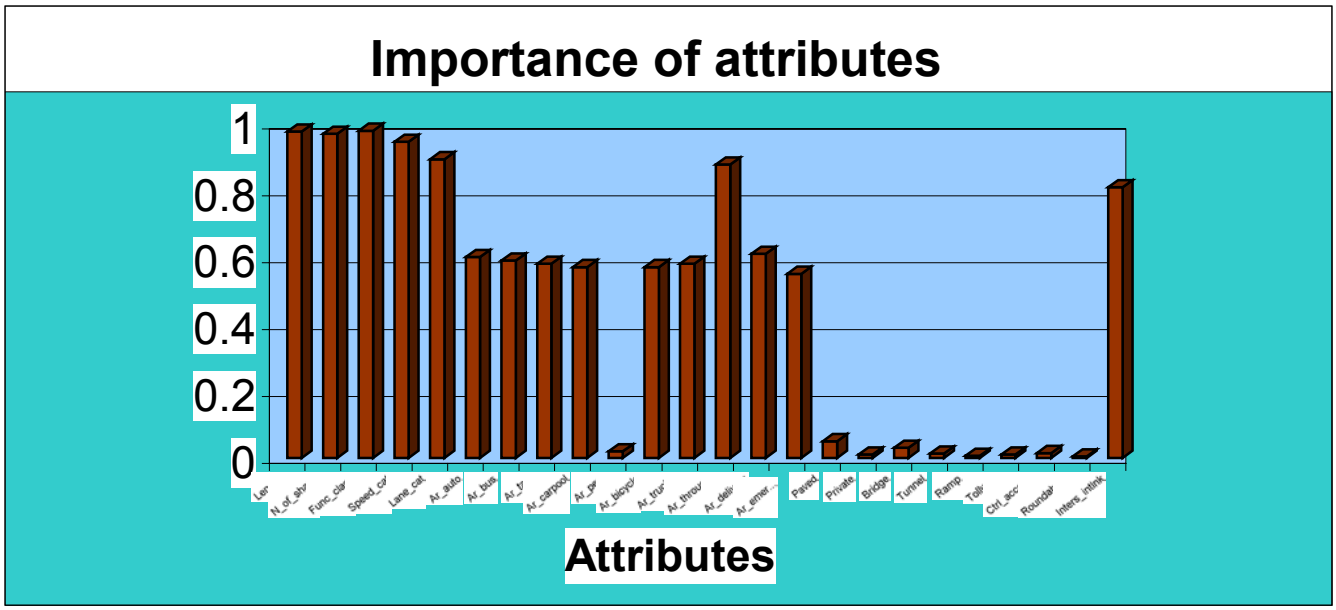


Figure 4. Importance of the chosen attributes indicated by their independence values

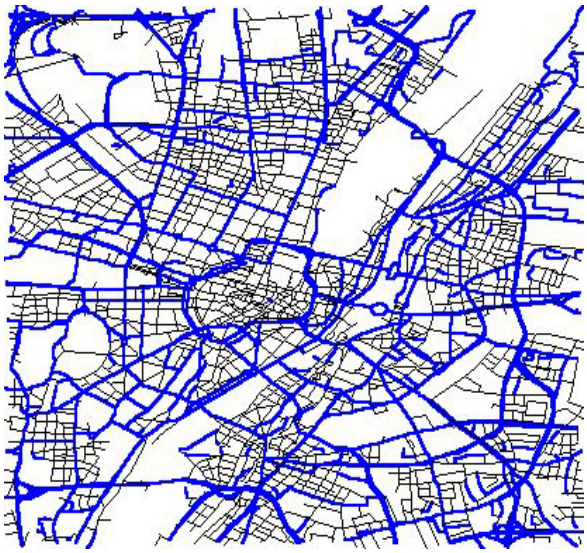


Figure 5. Grouping of the objects based on rough set algorithm

The value of the independence shows the importance of an attribute in another sense. For example, attribute *FUNC_CLASS* has a value ***0.98*** in the diagram and attribute *AR_EMER_VEH* is ***0.55***. So it can be easily understood that the attribute *FUNC_CLASS* is more important than *AR_ EMER_VEH*.

The table of ArcView after attribute analysis with information gains and rough sets method is supplemented with two new attributes, i.e. ***In_gains*** and ***Rough_set.*** **Tab.1** shows the new table of ArcView after our attribute analysis. With these two new attributes all objects can be ranked.

The test results may vary with test areas, which implies the regional differences. But for the same test area, the two data mining methods we have applied have produced the comparable results, which means both methods are relatively solid and robust

4. CASE STUDY AND OUTLOOK

A navigation system provides a new, extensive contemplation of the interplay of the driver, the vehicle and the environment in traffic. Suppose you want to move from one place to another. Then a navigation system offers you a personalized route planning. In order to build an optimal route planning all street lines must be analyzed. For example, each street line can be assigned with a degree of importance.

The two methods that are developed by us offer such possibility to classify each street or route quantitatively. **Fig. 6** shows an example. In this illustration, three arbitrarily selected routes ***A***, ***B*** and ***C*** are marked with different colors. As mentioned above, each object has 110 attributes. After the data mining process, each route is described with the value of information gains or with the value of rough sets. In this example, line ***A***, ***B*** and ***C*** possess the value of ***In_gains*** (calculated by information gains) and ***Rough_set*** (calculated by rough sets) with *30.52, 32.33, 35.44* and *25.23, 26.88, 28.99* (This value is calculated by the addition of all objects within the line ***A, B*** or ***C***). With this method each street line can be identified and described with one value. Only a simple is shown here. This function will be examined and further developed in the context of the spatial analysis.

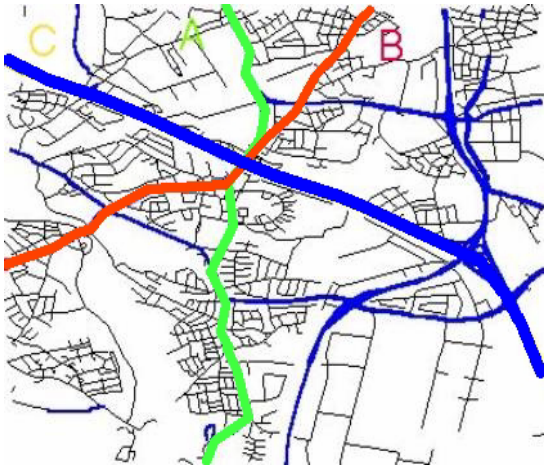


Figure 6. Quantitative classification of objects

FNODE_	TNODE_	LPOLY_	RPOLY_	LENGTH	F_NODE_I	…	***Number***	***In_gains***	***Rough_set***
857200	856059	0	0	0.00473	59522289		***0***	***0.82182***	***0.63018***
857164	856635	0	0	0.00216	59522312		***1***	***0.64697***	***0.46922***
857345	857200	0	0	0.00077	59522288		***2***	***0.64933***	***0.49095***
857200	857179	0	0	0.00016	59522289		***3***	***0.64697***	***0.46922***
857539	857345	0	0	0.00075	59522287		***4***	***0.88973***	***0.77604***
. . .						...			. . .
860840	860767	0	0	0.00167	59522107		***6347***	***0.79195***	***0.60243***
860851	860840	0	0	0.00108	59522104		***6348***	***0.79194***	***0.60240***
860883	860851	0	0	0.00168	59522097		***6349***	***0.79005***	***0.59993***
859570	859510	0	0	0.00018	59522233		***6350***	***0.70096***	***0.51827***

Table 1. A table after objects analysis

This example can be understood differently with following illustration (Fig. 7). Fig. 7 shows an example. We assume that an accident or a jam happens on the street. We have two alternative possibilities to drive through this street, i.e. we drive either on the street A or street B (Fig. 7). Through the analysis of information and attributes of all roads it can be calculated that the importance-degrees of the street A and B in each case are 0,78 and 0,89. So, we can determine that the street B can drive through with the better possibility as the street A. With the introduction of the analysis of information gains or the rough sets analysis, all street lines can be described only with a value. This value summarizes all information of the 110 attributes. Of course this is only a simple example. In fact, the result can be improved by the introduction of spatial information based graphic generalization. However, with the combination of the spatial and attribute information the street lines can be better identified and classified. This function must be also examined and developed in the future.

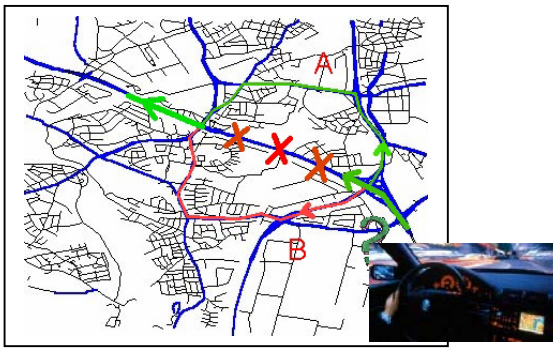


Figure 7. A example of transfer-offers before a breakdown

Another task in the future is the realization of the neural network for objects grouping and classification. For example, the information gains algorithm is used as decision-tree generator to approximate the concept, and this approximation is then further refined by neural network training. Such a training process starts from the results of the object analysis with information gains or rough sets analysis. We can use information gains and attribute-dependence as the initial values for the objects grouping in a neural network.

This project offers us an opportunity to examine the method of data mining in combination with the exploration of road database. The tests have shown that the data mining method can be applied to the analysis of road database.

Another work for the future lies in the combination analysis of both spatial information and attribute information.

5. REMAINING PROBLEMS

The results of the attribute analysis should be further applied to the classification of road lines. Now the quantitative grouping is oriented only to the original fragmental line segments. For a pragmatic navigation, these shorter lines or segments should be first connected into a longer line. With the information of the whole routes it can then be more reliably determined which route is more important.

The possible application of the artificial neural network for the analysis of the road database is being tested. Some work remains to establish a practical system, especially for the network training procedure.

Our programs run very well under ArcView 3.1 and with Windows 98, Microsoft ODBC SP5, SPSS 10. 0 and Merant 3.70. It can be used with Windows NT too.

LITERATURE

Andrienko, N., G. Andrienko, A. Savinov, H. Voss and D. Wettschereck, 2001. Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining. Cartography and Geographic Information Science, Vol. 28, No. 3, pp. 151-165.

Bissantz, N. and J. Hagedorn, 1996. Data Mining im Controlling. Teil A: CLUSMIN - Ein Beitrag zur Analyse von Daten des Ergebniscontrolling mitDatenmustererkennung (Data Mining).

Jacobi, M.H.H, 1999. Programmierung mit Avenue. Freising, im Oktober.

Liebig, W., 1999. Desktop-GIS mit ArcView GIS, Leitfaden für Anwender. 2., neubearbeitete und erweiterte Auflage. H. Wichmann Verlag. Heidelberg.

MENG, L., 1998. Cognitive modeling of cartographic generalization. Project report on "Strategies on Automatic Generalization of Geographic Data" - Stage 2.

MENG, L., 2001. Towards Individualization of Mapmaking and Mobility of Map Use. ICC 2001, Beijing, China.

SUI, L. & Yu, X., 2001. Pilot project - Exploration of the road database with data mining methods. Cooperation between NavTech GmbH and TU Munich. Project report from TU Munich, September.