Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

VISUAL AND AUTOMATIC DATA MINING FOR EXPLORATION OF GEOGRAPHICAL METADATA

R. Albertoni, A. Bertone, M. De Martino,(1)

U. Demšar, H. Hauska (2)

(1)Institute for Applied Mathematics and Information Technologies,

Consiglio Nazionale delle Ricerche, Genoa, Italy.

albertoni@ge.imati.cnr.it, bertone@ge.imati.cnr.it, demartino@ge.imati.cnr.it

(2) Department of Infrastructure, Royal Institute of Technology (KTH), Stockholm, Sweden.

urska.demsar@geomatics.kth.se, hans.hauska@geomatics.kth.se

1. INTRODUCTION

The paper focuses on issues related to the exploration of geographic metadata with the aim of searching for and selecting available geographic data. It describes the first results of research performed within the project INVISIP: “Information Visualisation for Site Planning”, supported by the European Commission within the Information Society Technology programme (IST 2000-29640).

Search and selection of available geographical data is one of the main tasks in geographic applications. They are characterized by different aspects, such as the large amount of data, the heterogeneity of data and the skill level of the user. Nowadays, huge amounts of geographical data are stored in databases, data warehouses, geographical information systems and other repositories, and these data collections are rapidly growing. Geographical data have many heterogeneous characteristics and can be used for various analyses in a wide range of applications. Considering these aspects and that the user can have either an incomplete knowledge about the available geographic data or the goal of his search is vague, it is necessary to provide tools to assist the user in the task of data selection.

The concept of metadata was applied to geographical data as an aid to the user whose aim is to determine the availability and suitability of a geographical dataset for an intended use. Metadata or "data about data" describes the content, the quality, the condition and other characteristics of data. Metadata standard such as ISO 19115 has been defined with the aim of providing a common set of terminology and definitions for the documentation of digital geographic data [1]. The standard establishes data characteristics such as: the names of data elements and of the compound elements (groups of elements), the definitions of these compound elements and data elements, the information about the values to be provided for the data elements.

The search of available geographical data is performed by exploring and analysing a repository of geographical metadata. Some of the problems that affect this activity are: the organization of the repository, low level of knowledge of the user about metadata characteristics and missing data values in some metadata attributes. In this paper we

address the three problems. Approaches based on a combination of automatic and visual data mining techniques to metadata are proposed. Repositories of geographical metadata are usually very large and rich datasets and traditional automatic approaches based on statistical techniques do not perform well on datasets of such characteristics [2].

An approach based on a combination of traditional data mining techniques and visualisation methods referred to as visual data mining [3, 4, 5] can be a solution. This approach aims to produce novel and interpretable patterns more effectively than automatic traditional techniques and stimulates the hypothesis generation. It has been successfully applied to the analysis of geographical data [6], but a new challenge is to focus on advantages coming from applying this approach to geographical metadata.

This paper gives a short overview of our recent studies on the application of automatic and visual data mining techniques to geographical metadata: we present our attempts to apply visual data mining to analyse and organise a metadata repository and to solve the problem of missing metadata.

2. DATA MINING ON METADATA

The exploration process of a repository of geographic metadata is affected by different factors such as:

- Amount of geographical metadata. The metadata repository contains a huge amount of geographical metadata which can disorientate the user in the searching activity: it prevents him from having a realistic overview of the available data.
- Missing metadata attributes. In the metadata standard some attributes are mandatory and others are optional. The search may require an analysis of metadata sets based on a comparison of their attributes. This task is difficult if some attributes are missing.
- User knowledge about available metadata. The geographical metadata standard ISO 19115 [1] includes many attributes. Metadata are usually explored by considering only a few of its attributes, chosen according to the user's requirements. If the user has only a partial knowledge of the available attributes, his requirements are limited to a subset of the attributes that could be used.

Considering these factors, new approaches to assist the user in the geographical exploration are needed. In this paper we focus on the following issues:

- organisation of metadata to easily provide information about the available data,
- completeness of metadata to improve the possibility to compare metadata elements,
- analysis of metadata to enlarge the user's awareness of the available data.

In the next sections we present a possible data mining approach for each of the mentioned issues. Particular attention is given to the approach based on the visual analysis of available metadata: the first result of our research has been the development of a visual data mining tool. The aspects of completeness and organisation of metadata are analysed and discussed: some interesting problems and open issues that will be the topic of our future research are outlined.

3. DATA MINING AND ORGANISATION OF METADATA

3.1 Organisation of metadata

Metadata analysis is a difficult activity, especially if a large amount of metadata is available. To limit the drawbacks related to the amount of data, the metadata repository needs to be well structured: data mining techniques are proposed to perform this organisation. Clustering techniques have been extensively studied to organise web document collections relying on the clustering hypothesis: "closely associated documents tend to be relevant to the same request" [7]. The solutions that have been developed are

mainly to facilitate web document collection browsing [8] and to increase the speed of the sifting of results obtained by web search engines [9]. Since document clustering usually bases its similarity criteria on some words or some phrases shared by documents, some techniques [10,11] apply an explicit cluster characterisation using those common words or phrases to provide cluster summaries. This approach would drastically improve the cluster content recognition.

3.2 Application of clustering techniques to metadata

The proposed approach to structure metadata is based on clustering techniques. Clustering techniques group data elements in clusters according to criteria of similarity: the clusters that are obtained are sets of similar elements. Each cluster represents a generalisation of the elements that it contains. The cluster structure represents a simplification of the repository and the analysis of the data can be limited to each cluster instead of to each single metadata element. The goal of our analysis is to study the ability of the clustering techniques to organise metadata in data selections. In the preliminary phase of our analysis, techniques based on hierarchical clustering proved to be the most promising ones. Such clustering produces clusters that are sets of similar elements, but at the same time each cluster is organized in sub-clusters. The result is a structure similar to a mathematical tree. This structure has some interesting properties. A parent node contains the union of elements that appear in its sons. In general, clusters at a high level of the tree contain elements that are less similar than elements that belong to the clusters that appear at a low level. An example of such tree structure is shown in (fig. 1). Metadata elements in this figure are represented as numerically labelled points.

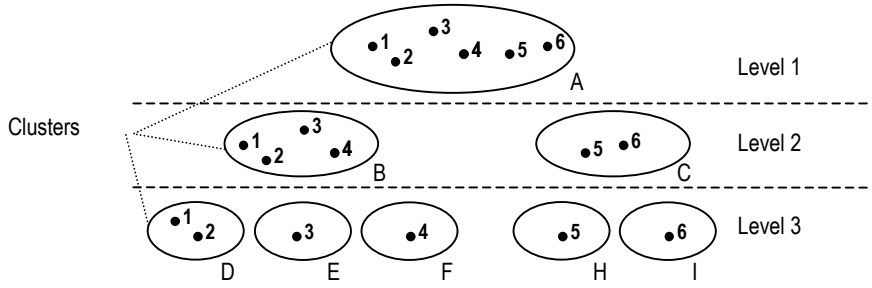

Fig. 1 An example of the tree structure inducted by Hierarchical Clustering.

The node A contains all elements belonging to the nodes B and C. Supposing that the similarity among metadata elements was coded with the distance between the points, it can be noted that the elements contained in clusters at level 1 are further apart than elements contained at level 2. This tree structure provides useful information during the exploration process: during the navigation of the tree structure, it is possible to focus on clusters instead of metadata elements. The applied concept of similarity is stricter if we are further down in the tree. The data search can be performed at different levels of detail.

To apply clustering on a repository of geographical metadata a similarity measure between metadata items that describes how to choose suitable dataset(s) needs to be defined.

$$s(t,a) = \sum_{i=1..n} c_i * s_i(t_i, a_i) \tag{1}$$

Considering two generic metadata items represented as two tuples $t = (t_1, .., t_n)$, and $a = (a_1, .., a_n)$, a generic similarity measure $s(t, a)$ can be expressed as in (1), where c_i represents a coefficient that expresses the importance of the i-th attribute in the similarity measure and s_i is the similarity criteria on the single i-th attribute defined according to its

format. Starting from the definition above, a similarity measure can be defined as the choice of the coefficients c_i and the similarity functions s_i. The simplest way to define a similarity measure is to consider the number of attributes that two metadata items share. This is easily obtained from the above formula by giving to all coefficients c_i the same value and defining s_i as the function that returns a constant greater then zero when two attributes have the same value and zero otherwise. Considering that all metadata attributes do not have the same relevance for each user, we suppose that the user dynamically chooses the attributes which he considers to be the most important by selecting the coefficients c_i himself. The clustering is performed after the coefficients of the similarity measure have been set. The dynamic choice of coefficients implies that the algorithm performance becomes particularly relevant to reduce the time bursting.

To improve the clustering approach, more sophisticated concepts about "what can be used instead of" can be developed by taking into account some contextual knowledge about geographical metadata. This knowledge has to be elicited by an expert in the geographical domain to determine rules that describe for each attribute "if" and "when" it satisfies a condition to be present or absent in the clustering.

For example, nowadays many GIS formats are interchangeable; we can consider two metadata attributes having different GIS data format (map info, arc view, etc) to be "closer" than two metadata attributes having different spatial representation (vector, row map, etc). A similar reasoning may be applied to keywords, to identify datasets which are more suitable, to identify the analysis to perform and so on.

The problem of representing the background knowledge in the clustering algorithm has different solutions. An interesting approach to semantic clustering has been proposed in the semantic web context, ontology is becoming a buzzword [12] and some interesting similarity measures based on this way to represent knowledge have been proposed [13].

In the future research we will focus on the problem of dependency between the user goal and the similarity criteria by exploring attributes of metadata that are more frequently involved in data search. The categorical attributes will be analysed to identify which are the expected values for each attribute and to check if these attributes support some kind of intuitive similarity concept. Further on, if techniques based on ontology are suitable to represent the background knowledge, their application to define similarity criteria might be investigated.

4. DATA MINING AND COMPLETENESS OF METADATA

4.1 Incompleteness of data

Standard statistical methods and data mining tools are produced to analyse rectangular datasets in the form of a matrix or a table. The rows of the matrix represent data items and the columns represent attributes of the dataset. The entries in the matrix can be as different as continuous numerical values, categorical numerical values (which may be in turn ordered or unordered), textual information, etc. Analysis of incomplete multivariate data can be defined as analysis of such a matrix when some of the entries are left blank. Completeness of such data depends on the application: it may refer to the presence or absence of whole data items (rows), missing attributes (columns) or blank entries within existing records. [14].

Usually the algorithms for data mining require a complete data matrix as input and some of them lose efficiency even if only a few values are missing. Methods for inputting missing data are therefore needed, to minimize the effect of incomplete datasets on the algorithms. Studies which analyse the impact of missing data on the algorithms include Zheng et al. [15], which analyses four recent classification techniques for robustness against missing data and Liu et al. [16] describing how a decision-tree based classification algorithm reacts to missing data values.

4.2 Methods for analysing incomplete data

Methods to analyse incomplete data focus on either ignoring the missing values or substituting them with plausible values, obtained in various ways. In this section we list several methods for dealing with incomplete data. Most of these methods make the assumption that the data are missing completely at random [14].

Two methods that ignore missing data are the complete-case analysis and the available-case analysis. The *complete-case analysis* includes only those data items that have all the attribute values present, while the *available-case analysis* includes those data items where the attribute value of interest is present. Both have the disadvantage that the loss of information might be significant due to the discarding of incomplete cases [14].

Another traditional group of methods for analysis of incomplete data is *single imputation*. This group includes methods that impute (fill in) the missing data values. Each blank field in the data matrix is filled in with exactly one missing value, which is derived from a subset of information on the original database. There are many methods that fall into this group of methods, based on the way the missing data value is derived. Here we list some of them: hot deck imputation (recorded units from sample are substituted into blank fields) [14], mean imputation (means of sets of recorded values are substituted) [14], cold deck imputation (an external constant is imputed) [14], regression imputation (missing values are predicted from a regression of the missing item) [14], stochastic regression imputation (missing values predicted by regression plus residual) [14], the most common attribute value imputation (the value of the attribute that occurs most often is set as the value for all the missing values) [17], KNN-impute algorithm (a weighted average of K nearest neighbour values is imputed) [18] and singular value decomposition (missing values are substituted by a linear combination of the first k eigenvectors of the data matrix) [18].

An advantage of single imputation is that any standard algorithms for data mining of complete data can be used once the missing values have been replaced by imputed values. Its disadvantage however is that the extra variability due to the unknown missing data is not taken into account, which can lead to a significant bias of the statistical measures [14,17].

A group of methods that corrects the major flaws of the single imputation while keeping its advantages is known as *multiple imputation*. In this group of methods each missing value is substituted by a set of m > 1 plausible values derived from a subset of the original dataset. Again there are several ways to derive these m values and consequently several types of multiple imputation: all possible values imputation (each missing value is replaced by the set of all values that occur for one attribute) [17], statistical multiple imputation (each missing value is substituted by a set of the most probable m values drawn from a predictive distribution of the values of each attribute) [14], partial values imputation (a partial value is a set of k values such that exactly one of the values in the set is the true value of the missing value, the set is obtained by logic programming) [19,20].

Both single and multiple imputation methods are prone to estimation errors that increase with dimensionality and incompleteness. This is due to the fact that when a large amount of the entries are missing, the attributes can only be estimated with a low degree of accuracy. Therefore other methods have been developed, which cope with the problem of the incomplete information system without using imputation methods. They are based on the fact that in the real datasets there are usually considerable redundancies and correlations among the data representation [21, 22].

One of these is the *data decomposition method* [21] which tries to avoid the necessity of reasoning on data with missing attribute values. The original incomplete data are divided into data subsets without missing values and classification is applied to each of these sets. Finally a merging method is used on the partial answers from all subsets to obtain the final classification.

Another method that avoids the imputation of values is the *conceptual reconstruction method* [22]. It is based on assumption that the attributes in a dataset are not independent

from one another, but that there is some correlation between them which enables prediction of the missing values in one attribute from values of another one (or more than one).

4.3 Application of incompleteness algorithms to metadata

The metadata collected within the INVISIP project is represented in the same form of database as described above for normal data. It is organised into a table or a matrix that has metadata records as rows and metadata attributes as columns. The geographical metadata standard ISO 19115 [1] is characterised by many attributes which describe the characteristics of data: some of the values of these attributes are obligatory for each dataset but others can be optional. Thus some data are frequently missing, which can cause problems in the analysis task when an analysis of attributes with missing values is performed. Therefore it is important to include a pre-processing step in the analysis that would deal with missing data.

We defined an approach that would enable us to identify some of the missing values of metadata by analysing the set of available metadata. The approach is based on a combination of a classification algorithm applied to the metadata attributes and an algorithm that completes the missing metadata values.

The first step is to use a classification algorithm on attributes to identify relationships among attributes which are completed and attributes with missing data. An example of an algorithm that could be used for similarity classification of attributes can be found in [23]. We suggest to apply a classification algorithm to metadata attributes to identify relationships among complete attributes and attributes with missing data. The attributes with missing data could be used to define the class labels, while the set of complete metadata elements could be used as a training set to find relationships among the attributes and to define the criteria that could be used to fill in the missing data.

The relationships between the attributes discovered by the similarity classification of attributes could then be used to form the logical rules that represent the connections between different attributes. These rules could be applied in the second step approach that fills in the missing data, similar to for example, the partial values algorithm described in [18]. We suggest to use the partial values algorithm, because it uses logic programming and can be applied to geographical metadata. Other algorithms for filling in the missing values are mostly based on different types of statistical calculations and operate on numerical data. Since geographical metadata consist of very different types of data such as numerical data, textual data, categorical data, etc., the statistical calculations can be difficult to implement. But forming logical rules needed for the partial value algorithm seems to be feasible for all types of data. The feasibility and implementability of this approach is yet to be explored in our future research. One of the possible problems seems to occur if the initial classification is applied to attributes that are not related to each other. In that case the result can be inconsistent: we can discover relationships that do not exist. Consider two standard geographical metadata attributes: format and scale. There is no relation between these two attributes. Maps in different scales can be represented in the same format and the same map can exist in different formats. However, if a classification is applied we may find a connection, which in reality doesn't exist. Therefore a classification technique could be used in the first step of our proposed procedure to complete geographical metadata, but the knowledge about geographical domain must play an essential role in recognition of the resulting relationships among metadata attributes.

5. VISUAL DATA MINING AND METADATA ANALYSIS

A visual data mining approach based on different kind of visualisation and interaction techniques was applied to metadata to discover new patterns [24]. A common problem in the search for geographical data is the vagueness of the goal: geographical metadata are usually represented by many attributes and the user usually has an incomplete knowledge

about them. Therefore, to support the user during this task, it is important to provide tools in order to improve his knowledge and raise his awareness of the available data.

The use of various visualisation techniques and a dynamic interaction with them helps the user to discover new information. In our approach, different visualisation techniques can be applied at the same time to show the user which data are available. These techniques are classified according to the number of attributes they can display: single-attribute and multi-attribute visualisations. A single-attribute visualisation provides knowledge about the values and a quantitative information of a single metadata attribute. A multi-attribute visualisation provides knowledge about several metadata attributes and the existing relationships among them. Several graphical interaction functionalities let the user select metadata according to his needs. Finally, the selection performed in a visualisation window is linked to all other visualisations through a brushing and linking process [4] to enable the user to discover correlations between attributes.

The main advantage of the approach is that even if the geographic data or some of its characteristics that the user is looking for are not available, the knowledge coming from the metadata patterns leads to a compromise between what the user needs and what is available.

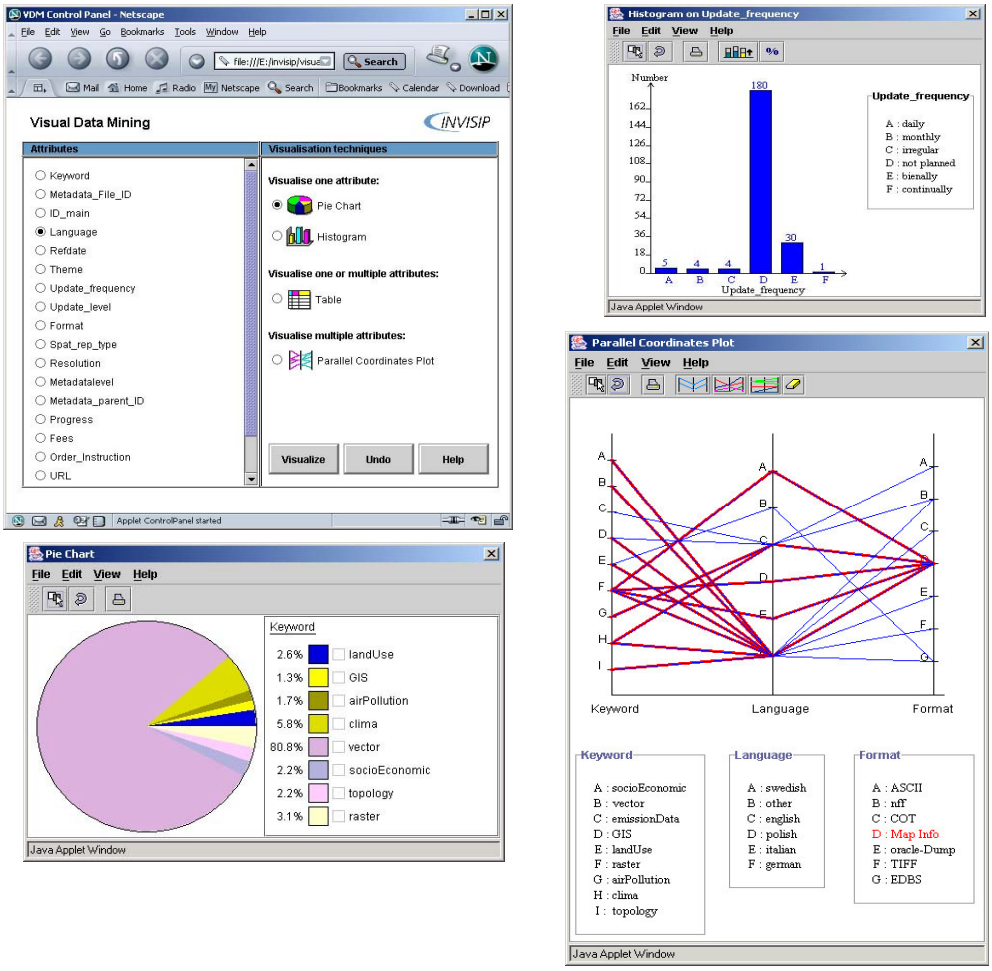

Fig. 2 Tool and its visualisations.

In (fig. 2) is shown a tool developed for this approach and a brief example below describes how this tool can support the user in his exploration. Let us assume that the user wants to acquire geographic data. He performs a search in a repository of geographical metadata. These metadata are characterized by several attributes such as the keyword related to the typology of data (climate, economy, air pollution …), the language of the data, the format of the map, the URL where data are stored, the resolution of the map, etc. Suppose that the user bases his search on some of these attributes: he analyses the keyword, the language and the format, having only a vague idea on how these attributes can be combined. He is interested in data with MapInfo format, in English and dealing with the keyword “emission data”. Using different kind of visualisations he realizes that no data fit his needs. However if he accepts the keyword “air pollution” instead of “emission data”, he can find data that he could use. Using a multi-attribute visualisation, the user can deal with a large metadata repository and with many attributes at the same time. The brushing and linking functionality makes the exploration easier: every time the user performs a selection on a visualisation window, the other views are automatically updated to show which data are currently available. This interaction lets the user explore data, improve the knowledge about data and among the available datasets find those that suit his goal best.

6. CONCLUSIONS AND FURTHER WORK

A combination of automatic and visual data mining techniques used on the geographical metadata provides a friendly way to perform a search for geographic data even for a not very experienced user. In the first phase of our research, an analysis of automatic data mining algorithms and visualisation techniques suitable for geographical metadata analysis was performed. Within this step the first demonstrator tool for visual data mining with some basic visualisation techniques was developed.

Further work includes the evaluation of other automatic data mining algorithms and visual data mining techniques and their suitability to analyse geographical metadata. One of the goals in the development of our tool is to implement the integration between the suitable automatic algorithms and the visualisations provided. There is a possibility to develop other different types of visualisations, such as those that are linked to either spatial or temporal distribution of geographical data described by metadata. The final aim is to provide an instrument that will assist the user with his data exploration throughout the whole site planning process.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the European Commission, IST Programme. They thank all those who contributed to this study and especially the INVISIP partner: University of Agriculture of Krakow, Poland.

8. CONTRIBUTORS

Authors’ names are given in alphabetical order. Riccardo Albertoni, Alessio Bertone and Urška Demšar have written this paper as a joint contribution under the supervision from Monica De Martino and Hans Hauska.

9. REFERENCES

[1] *ISO 19115* «Final Draft: International Standard on Metadata for Geographic Information», Status: approval stage, http://www.iso.org, 2002.

[2] Gahegan M., Wachowicz M., Harrower M., Rhyne T. M., «The Integration of Geographic Visualisation with Knowledge Discovery in Databases and Geo-computation», *Cartography and Geographic Information Science*, Vol. 28, No. 1, 29-44, 2001.

[3] Keim, D., Müller, W., Schumann, H. «Visual Data Mining» *Eurographic STAR proceedings*, Saarbrücken, 2002.

[4] Keim, D. «Information Visualisation and Visual Data Mining», *IEEE Transaction on Visualisation and Computer Graphics*, NO 1, 2002.

[5] Fayyad, U. M., Grinstein, G. and Wierse, A. Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, San Francisco, 2002

[6] Takatsuka M., Gahegan M. «GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualisation»*Computers & Geoscience,* Vol. 28, 1131-1144, 2002.

[7] Van Rijsbergen C.J. Information Retrieva l, Butterworths, London, 1979.

[8] Cutting, D. R., Karger, D. R., Pedersen, J. O., Tukey J. W., «Scatter/Gather: a cluster-based approach to browsing large document collections», *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark, 318—329, ACM Press, 1992.

[9] Leuski, A., «Evaluating document clustering for interactive information retrieval», *Proceedings of the tenth international conference on Information and knowledge*, Atlanta, Georgia, USA management, 33—40, ACM Press, 2001.

[10] Zamir, O., Etzioni, O., «Web document clustering: A feasibility demonstration», *Proceedings of the 21st Annual International ACM SIGIR Conference*, pp. 46--54, 1998.

[11] Zamir O., Etzioni O., «Grouper: a dynamic clustering interface to Web search results», *Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999.

[12] McGuinness L.D, «Ontologies Come of Age», *The Semantic Web: Why, What, and How,* MIT Press, 2001.

[13] Maedche A., Zacharias V., «Clustering Ontology-based Metadata in Semantic the Web», *Principles of Data Mining and Knowledge Discovery*, Helsinki, Finland, 2002.

[14] Fogarty D.J., Blake J., «Utilising Recent Advancements in Techniques for the Analysis of Incomplete Multivariate Data to Improve the Data Quality Management of Current Academic Research», *Quality and Quantity*, Vol. 36, No. 3, 277-289, Kluwer Academic Publishers, 2002.

[15] Zheng Z., Low B. T**., «**Classifying Unseen Cases with Many Missing Values», *Lecture Notes in Computer Science*, Vol. 1574, 370-375, Springer Verlag, Berlin, Heidelberg, 1999.

[16] Liu W.Z., White A.P., Thompson S.G., Bramer M.A.,«Techniques for Dealing with Missing Values in Classification», *Lecture Notes in Computer Science*, Vol. 1280, 527-536, Springer Verlag, Berlin, Heidelberg, 1997.

[17] Grzymala-Busse J. W., Hu M.**, «**A Comparison of Several Approaches to Missing Attribute Values in Data Mining», *Lecture Notes in Artificial Intelligence*, Vol. 2005, 378-385, Springer Verlag, Berlin, Heidelberg, 2001.

[18] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B., «Missing value estimation methods for DNA microarrays», *Bioinformatics*, Vol. 17, No. 6, 520-525, Oxford University Press, 2001.

[19] McClean S., Scotney B, Shapcott M., «Using Background Knowledge in the Aggregation of Imprecise Evidence in Databases», *Data & Knowledge Engineering*, Vol. 32, 131-143, Elsevier, 2000.

[20] McClean S., Scotney B, Shapcott M, «Using Background Knowledge with Attribute-Oriented Data Mining», *Proceedings of IEEE Colloquium on Knowledge Discovery and Data Mining*, London, Digest No: 98/310, 1-4, 1998.

[21] Latkowski R., «Incomplete Data Decomposition for Classification», *Lecture Notes in Artificial Intelligence*, Vol. 2475, 413-420, Springer Verlag, Berlin, Heidelberg, 2002.

[22] Aggarwal C. C., Parthasarathy S., «Mining Massively Incomplete Data Sets by Conceptual Reconstruction», *Proceedings of the 7th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, San Francisco, California, 2001.

[23] Ankerst M., Berchtold S., Keim D. A., «Similarity Clustering for an Enhanced Visualization of Multidimensional Data», *Proceedings of Infovis, IEEE Symposium on Information Visualization*, North Carolina, 52-61, 1998.

[24] Albertoni R., Bertone A., De Martino M., «Information Visualisation And Interactive Geo-Data Mining In Site Planning Process», *Eurographic Italian Chapter*, Milan, 2002.