Technical Report
N° 16/2004

Visualization and ontology to analyse categorical attributes in geographic metadata

Riccardo Albertoni,
Alessio Bertone,
Monica De Martino

Istituto di Matematica Applicata e Tecnologie Informatiche,
Consiglio Nazionale delle Ricerche
Via de Marini 6, 16149 Genoa, Italy

1 ABSTRACT

The paper addresses issues in the analysis of geographic metadata.

Metadata, data about data, are used to maintain, organize, and provide information about geographic data and their analysis is an important task in geographic data searching activity. Due to the complexity and the diversity of the attributes of geographic metadata, the analysis of metadata can lead to the problems of data missing and of unfamiliarity with attributes.

We propose an approach to analyse geographic metadata based on visual data mining techniques to facilitate the navigation in unfamiliar spaces and ontology to support user in the search criteria. In particular the analysis of categorical attributes of metadata is considered. Techniques and concepts of information visualisation, graphic interaction, data mining and ontology are analysed and exploited: the main idea is to integrate some of these techniques and concepts to facilitate data searching activity.

Keywords

Geographic metadata analysis, categorical data, ontology, visual data mining.

2 INTRODUCTION

Geographic Information have a mass of data which is dramatically growing. Demands for digital geographic data grow larger with the increases in GIS technology and World Wide Web (WWW) accessibility. As the quantity of geographic information on the WWW multiplies rapidly, it will become increasingly difficult to share them and retrieve information with reasonable precision. To manage the huge amount of information that the actual society produces, mechanisms and systems which organize data and provide information about where to find and access the searched data have been developing.

Metadata, data about data, are an effective way to satisfy these demands. The use of metadata combined with the use of exploration processes has the potential to improve retrieval of these information resources.

The large set of geographic data, its heterogeneity as well as the large amount of geographic providers determine the generation of large metadata database. This requires the development of approaches that allow to perform the metadata analysis during the activities of geographic data searching.

Metadata are characterised by different attribute representations: numerical, descriptive and categorical. Whereas many approaches deal with the analysis of numerical and descriptive attributes, less are proposed for the categorical ones. Categorical data, sometimes referred to as nominal data (data that can be named), are data that can be separated into different categories distinguished by some non-numeric characteristic. The collection of categorical data involves the counting of occurrences that can be named and enumerated, and it is analysed using a number of statistical methods, including contingency tables, regression models, conditional inference [1], and correspondence analysis [2][3].

The paper reviews important data analysis tools and approaches used to explore geographic metadata in order to define an alternative approach working on categorical metadata attribute. In particular a new approach based on the integration of the potentialities offered by *Visual data mining* [4] and *Ontology* [5] is proposed. The visual data mining is a novel approach in the knowledge discovery process, which combines Information Visualization [6] and Data mining [7]. Information Visualization is used as a communication channel between the computer and the user, whereas Data mining is used to analyse data and to extract patterns from data. Ontology refers to an engineering artefact, constituted by a specific vocabulary used

to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of vocabulary words [5].

In particular the approach is mainly based on interactive visualizations to discover and validate new and interesting patterns among data. Ontology concept is adopted to represent the semantic relationships among data.

This research started within the European research project INVISIP (Information Visualisation for Site Planning) whose main results have been illustrated in detail in [8][9]. From our experience it was evident the importance of working on categorical data and to give more attention to the analysis of semantic relationships among data stored in the metadata database. In this paper we present a new step of our research oriented to the semantic aspects of the analysis.

The paper is organised as follows: in the first part an overview of metadata analysis approaches is introduced and discussed. Then the approach is described in details: a short overview of the main concepts of ontology and similarity is introduced; the use of visual data mining and clustering are described, and how to exploit the semantic relationships is presented. The last part is dedicated to the discussion and analysis of its potentialities and limits.

3 Metadata

Metadata are data about data [10] Metadata help a user to understand what data are about and what processes occurred in producing those data. The importance of metadata lies in their ability to maintain, organize, and provide information about geographic data.

Therefore even for geographic data a metadata standard has to be adopted. There are a variety of metadata standards. The two main organizations providing standards are the International Standard Organization (ISO), which published ISO19115 [12], and the Federal Geographic Data Committee (FGDC) which published the Content Standard for Digital Geospatial Metadata [13]. There are some efforts to harmonize them, but they are still in progress [14].

Any metadata standards is defined by many attributes which are represented by different data types:

- Numerical data: numbers,
- Categorical data: nominal values not containing a natural order or data with an implicit order,
- Descriptive data: free text describing the attribute.

The paper focuses on the categorical attribute of ISO 19115 metadata standard. Categorical attributes play an important role both since they are numerically relevant (more than twenty metadata attributes are define as categorical) and they represent important information (e.g. maintenance represented in MaintenanceandUpdateFrequency attribute; progress as status attribute; type of spatial representation as spatialRepresentationType attribute , resolution as SpatialResolution attribute, theme classification as TopicCategory attribute). The ISO 19115 metadata standard defines a *domain* of values for each categorical attribute: it is the set of all possible values the attribute can assume.

4 Metadata Analysis

The vast collections of geographic data determine the generation of large sets of metadata. The complexity of geographic data leads to metadata having several attributes. Instruments of metadata analysis to support the user in data searching and selection are needed.

The analysis process to search data is affected by the following problems:

- *Unfamiliarity with attributes*: the user usually has only a partial knowledge of all the available attributes and must perform his selection in an unfamiliar information space. The searching criteria that he is able to perform might not be enough to successfully end its selection activity. Therefore he needs to refine his criteria using some attributes he is not familiar with.
- *Data missing*: the metadata database might not contain the data the user is looking for; hence he is forced to define new criteria to find data similar to those he is looking for.

Although non visual approaches to accomplish the metadata analysis tasks have been proposed [15],[16], an approach based on data visualization appears more fruitful. Considering the multidimensionality and the quantity of metadata, the search results expressed in a textual format overcome user capabilities of comprehension. On the contrary, an approach based on data visualization helps to face with large volumes of information. [17] describes several information visualization applications; [18] describes how the integrated use of multiple, concurrent visualization techniques can improve the user's understanding of the complex and highly inter-related information.

Different approaches may be adopted according to the data type representation of the analysed attributes. Some approaches focus on the analysis of geographic data expressed as numerical values [19][20]: they provide a visual spatial analysis, and a dynamic and interactive manipulation of the weights of the selected attributive criteria, in order to support the user in his tasks. Others focus on the analysis of descriptive data. They use information visualization techniques to visualize text (such as [21],[22],[23]). The aim is either to find relevant information by refining or filtering user queries combining different visualizations into a so called SuperTable [24], or to browse textual and numerical metadata by also using the domain ontology [25]).

On the contrary, less interest has been posed on categorical data. The exploration of categorical attributes is challenging because the values that they assume provides pieces of information which can be easily understood by a human agent but cannot be trivially managed in automatic way. In order to make the categorical values “machine understandable”, their semantic relationships (such as generalization, dependency and so on) have to be explicitly represented.

An approach for the metadata analysis should satisfy the following requirements:

- *To provide a compact representation of data*: Due to the volumes of metadata that could be generated, visualizations providing an overview of data increase the understanding of the available data and help to face with the unfamiliar attributes.
- *To support in the mining of the selection criteria*: The user is not always aware of all metadata attributes, new search criteria need to be mined to face with unfamiliar attributes and data missing problems.
- *To exploit the semantic relationships among categorical attributes*: the use of semantic relationships facilitates the exploration of categorical attributes especially if the searched data are missing.
- *To ease the compliance to the meaning of the terms given in ISO metadata standard*: ISO metadata standard defines some terms (literals) and how they must be interpreted, therefore their meaning have to be taken into account otherwise their analysis can lead to misleading results. To make accessible the intended meaning of terms is useful in solving unfamiliarity problem.

Several approaches to visualize categorical attributes have been proposed. Some of them are specifically designed for nominal values such as sieve diagrams, mosaic displays, and fourfold displays [26], correspondence analysis maps [27], treemaps and CatTrees [28]. Others use ordering techniques to map nominal values to numbers (e.g. [29] uses another variable to provide the order, [30] obtain the order from a three steps algorithm. All approaches provide a compact representation of the data). They allow to understand the relationships among data and to perform an explorative analysis, then to mine the selection criteria. Unfortunately, they have not been applied to categorical metadata, therefore they do not provide any compliance to standards, and they do not take into account the semantic relationships among categorical attributes. ([30] use the domain semantics to build the natural clusters that order categorical data in the first step of their algorithm. However this is not enough to fulfil

the requirement, since the semantics refer to the behaviour of the elements of the cluster, i.e., grouping hosts that emit events at the same time).

Hence we propose an approach to facilitate the exploration of the categorical attributes in compliance with the ISO 19115 metadata standard. It provides a compact representation of data and exploiting the ontology concept to solve both the problem of unfamiliarity with attributes and the problem of data missing.

5 Overview of Proposed Approach

A metadata analysis framework is proposed whose primary design goal is to allow users to move through large information spaces in a flexible manner without feeling lost. The approach is characterised by:

- A reasoning based on the application of Visual Data Mining to perform the exploratory metadata analysis with the attempt to find interesting structures unknown a priori.
- A reasoning based on the application of the ontology concept to support user in the semantic exploration of categorical data.

In the following we introduce some concepts needed to define the approach, then we give details of the two reasoning.

5.1 Representation of Semantic relationships of categorical attribute.

The ontology represents an emergent way to formally describe concepts and relations among them. In the following, the basic notions about ontology and the similarity measure among its concepts are introduced.

5.1.1 Ontology.

An ontology is a formal explicit specification of a shared conceptualisation. A conceptualisation is an abstract, simplified view of the word that we wish to represent for some purpose. “Explicit” means that all the concepts and the used constraints are explicitly defined, “formal” means it should be machine understandable, “shared” indicates that the ontology captures consensual concepts resulting by an agreement in a specific community [31]. The ontology needs to be defined by a community of expert in the application domain.

Ontologies are used to provide a concrete specification of the names and the meanings of terms. However to make the specification comprehensible for people who are not member of the community, it is important that concepts and relationships illustrated in the ontology are carefully documented [32].

Different level of details can be coded in an ontology according to the application context [33]. In general an ontology is composed by *classes*, *relations*, *instances*,

functions and *axioms*. *Classes* represent concepts of the application domain, *relations* represent types of interaction between concepts of the domain; *instances* represent specific elements of classes (for example if you considered the class “City”, then “Washington D.C”. would be an instance of the City class); *functions* are special kinds of relations; *axioms* model sentences that are always true.

In this paper classes, IS-A relation and instances are considered to explain the core of our approach. The ontology is characterised by classes hierarchically organized according to the IS-A relationship. The hierarchy has a unique root class. Since only IS-A relations are considered, the hierarchy appears as a taxonomy. Moreover it is possible to define *abstract classes*, defined as classes without instances which represent concepts useful to complete the hierarchy of the ontology.

5.1.2 Similarity Measure

A measure of *similarity* among classes and instances needs to be defined to make explicit the semantic relationship among concepts. The paper informally introduces the definition of *Taxonomy Similarity* presented in [34].

Let define:

- C the set of classes belonging to the ontology,
- $H^C \subseteq C \times C$ a direct and transitive relation called *Concept/Class Taxonomy*. $H^C(C_1, C_2)$ means that C_1 is subclasses of C_2,
- I the set of instances,
- $Inst : C \to 2^I$ a function called *Concept Instantiation* that given a class provides its instances,
- $Class : I \to C$ the function which given an instance returns the class it belongs to.

Definition of Upwards Cotopy (UC):

$$UC(C_i, H^C) := \{C_j \in C \mid H^C(C_i, C_j) \vee C_i = C_j\}$$

The Upwards Cotopy of a class C_i on a Class taxonomy H^C represents a set of classes containing C_i and all classes which have it as subclasses in H^C.

Since the Class Taxonomy has always a root class at the top of the hierarchy, $UC(C_i, H^C)$ can be also thought as the set of classes which composed the path to

reach the root of hierarchy from C_i. Starting from the definition of the Upward Cotopy (UC), it is possible to define the Concept Match (CM).

Definition of Concept Match (CM):

$$CM(C_1, C_2) := \frac{\left|(UC(C_1, H^C) \cap UC(C_2, H^C))\right|}{\left|(UC(C_1, H^C) \cup UC(C_2, H^C))\right|}$$

The Concept Match captures the similarity between two classes considering the number of classes that are in common in the hierarchy. Of course the relevance of having n classes in common increases of importance with the decreasing of classes needed to join C_1 and C_2 to the root of hierarchy.

Definition of Taxonomy Similarity (TS):

$$TS(I_1, I_2) = \begin{cases} 1 & if \quad I_1 = I_2 \\ \dfrac{CM(Class(I_1), Class(I_2))}{2} & otherwise \end{cases}$$

Taxonomy Similarity provides a mean to measure the similarity between two instances of an ontology. In the following, it is adopted to ease the metadata analysis.

5.2 Metadata exploration by Visual Data Mining

In our approach the metadata exploration is characterised by the integration of multi visualization techniques, interaction functionalities and data mining concepts. It is characterised by the following two activities which can be repeated during the exploration process:

- Visualization of metadata attribute with graphic interaction and brushing and linking
- Organization of the metadata set in cluster according with similarity criteria.

5.2.1 Visualization and brushing and linking

Figure 1 shows a schema of the approach. Different visualizations can be simultaneously exploited and they are used to display one or several attributes. In particular according to the number of attributes that the visualizations can display, we can distinguish between single attribute and multi attribute visualizations: a single attribute visualization (Pie Chart, Histogram) is proposed to provide the knowledge of the available values and quantitative information of metadata attributes, on the

contrary a multi attribute visualization (Table Visualization, Parallel Coordinates Plot) to provide the knowledge about metadata attributes and the existing dependencies.

The exploration task is performed using both interaction functionalities with the element displayed in a visualization, and brushing and linking with highlighting [35]: brushing is an interactive selection process, while linking connects the selected data from the current visualisation to other open visualisations. If the user has several different visualisations open (of one type or of more different types) and selects an objects in one of them, the graphical entities that represent this object and correspond to the same subset of selected data objects are highlighted in each visualisation. Hence the user is able to evaluate the effect of a selection before performing it, and to realize what the actual data are and what data he would have if he performed the selection. Once the selection is performed, all other graphical entities disappear from each open visualisation ([8] [36]).

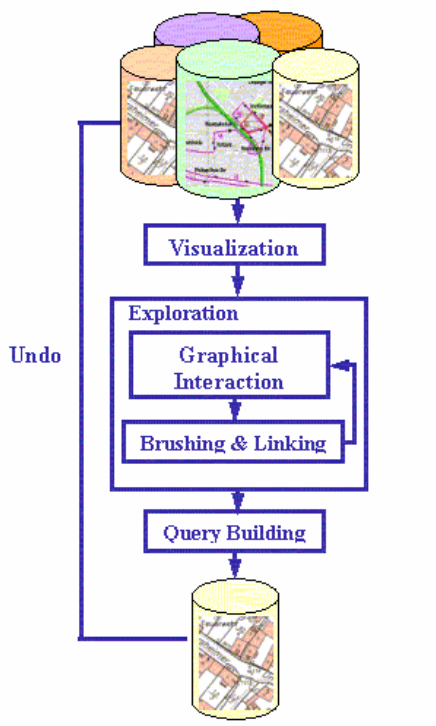

Figure 1: Visual Data Mining approach

5.2.2 Organization of metadata

Another task characterising the proposed approach is the organisation of large sets of metadata in similar clusters in order to successively analyse them.

The organisation of metadata is performed by using a hierarchical clustering [37],[38]. Clustering techniques [39] 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. Each cluster is a set of similar elements, but at the same time it is organized in sub-clusters. The result is a structure similar to a mathematical tree (Figure 2), where 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.

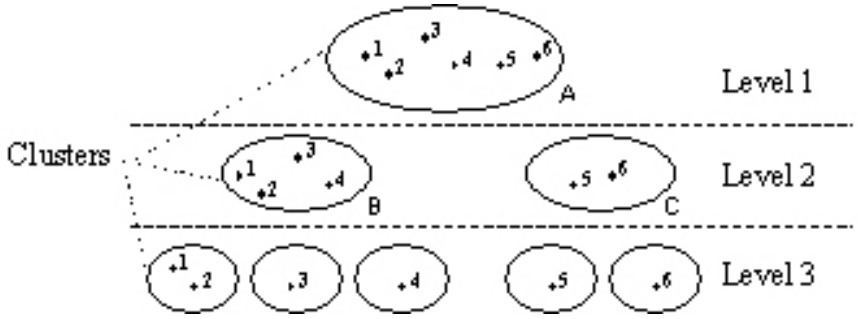

Figure 2: The tree structure inducted by Hierarchical Clustering

The kernel of the clustering is the concept of similarity. To define a criterion of similarity among categorical data is particularly difficult.[40] propose to adopt the syntactical equality as similarity in case of categorical values and they define the distance function among categorical attributes belonging to ISO metadata items as 1 whenever two categorical attributes are equal, 0 otherwise. We propose to improve the similarity criterion by adopting the *taxonomy similarity*, which was originally introduced in [34] for clustering resources on semantic web. Due to the mapping between categorical attributes and ontology proposed in our approach, the taxonomy similarity can be applied to categorical values. Thus the semantic relationships among categorical values are taken into account providing clusters that are more meaningful for the user exploration.

Once a clustering technique has been applied, a visualization of the obtained structure is needed. A dendrogram, that is a hierarchical tree, is the resulting structure of the clustering: it shows the nested groupings of patterns and the similarity levels at which the groupings change.

There exist several solutions to display this structure such as Magic Eye View [41], Cone Tree [42], Treemap [43], H-Blob [44], Snowflake graph [45]. Independently of the visualization, one of them is integrated into the process of brushing and linking and the selected graphical entities are linked to all the other visualisations, in order both to organize metadata and to explore metadata.

5.3 The semantic exploration of categorical data

This paragraph illustrates how to exploit the semantic relationships among the values belonging to the domain of a categorical attribute. The notions of ontology and Taxonomic Similarity are adopted to express these relationships and to make them machine understandable.

The approach is characterised by:

- the definition of semantic relationship,
- the exploitation of the semantic relationship.

The definition of semantic relationship is obtained by mapping categorical data into an ontology and to compute the similarities among concepts by the taxonomy similarity in order to make them machine understandable.

The exploitation of semantic relationships is performed by exploring the similarities among concepts along with the approach based on visual data mining which has been previously illustrated. This implies the integration of an appropriate visualization to display the similarity measure among ontology classes in the visual data mining framework. We are evaluating to use some visualization techniques as those proposed in [46], [47].

Details of these two activities are given in the following. How to define the ontology by mapping the categorical attribute in the ontology and an example describing how to compute and exploit the semantic relationships is illustrated.

5.3.1 Mapping Rules between categorical data and ontology

The approach adopts ontology composed by classes, subclass relations (IS-A relation which generates a Class Taxonomy) and instances. For each attribute the following rules to build the ontology are considered:

- *Definition of C set of classes*:

- *Definition of the root of the taxonomy*: the name of the categorical attribute is defined as abstract class that represents the root of the taxonomy.
- *Mapping of literals into the ontology classes*. each literal belonging to the domain of a categorical attribute is mapped into an ontology class.
- *Documentation of classes*: the ISO specification provides a documentation for each literal included in the attribute domain. Such a documentation has to be included as a comment to the class in the ontology definition.

- *Definition of the Class Taxonomy* H^C *:*
- *Introduction of IS-A relationship*: an IS-A relation is introduced stating that a class is a subclass of another one whenever the two correspondent values in the domain of the categorical attribute represent concepts where it is possible to identify that the former is more specific than the latter. Anyway, all the classes are subclasses of the root of the taxonomy.
- *Definition of the set of instances I.* Each metadata item having a specific literal for a categorical attribute is mapped as instance of the class associated to such literal.

Due to the mapping between metadata items and instances of the ontology, in the rest of the paper we refer to a metadata item which has a specific value for a categorical attribute as instance of the corresponding class in the ontology.

In general the literals proposed as domain of a categorical attribute are not enough to define an ontology. Abstract classes can be added to the ontology to complete the conceptualisation above the domain.

For example, in the ontology design, it is important to ensure that all the direct subclasses of a class represent concepts lying at similar level of abstraction. If such a design rule cannot be satisfied by using only the classes corresponding to literals suggested in the ISO metadata, some new abstract classes can be introduced to balance the level of abstraction among the “sons” of a class.

5.3.2 Example: identification and exploitation of semantic relationships.

Let analyse the categorical attribute *topicCategory* defined in the ISO19115 which describes a high-level classification for geographic data theme. The set of values that it can assume is defined in its *domain* MD_TopicCategory and includes “farming”, “biota”, “boundaries”, “society”, “inLandwaters”, “location”, “economy”, “environment”,

“health”, “structure”, “ “climatologyMeterologyAthmosphere”, geoScientificInformation”, “transportation”, “elevation”, “baseMapsEarthCover”, “image”, and so on.
These values have some intuitive semantic relationships that are useful in the data analysis as long as they are perceived by the user. For example, intuitively, the topic “farming” is more related to environmental topics (such as biota, farming, and climatologyMeterologyAthmosphere) than to “boundaries”. According to this semantic relationship it is clear that if data about “farming” are searched, environment topics would be more interesting to be considered, than “boundaries”.

An example representing such an intuitive relationship by adopting an ontology and the taxonomy similarity is now illustrated. Note that this example is ‘minimal’: the definition of a complete ontology about theme classification requires an effort that is out of the purpose of this paper; it would require a long interactive design process where experts of the domain have to be directly involved.
According to the proposed mapping rules, Figure 3 illustrates how some of the values (environment, structure, climatologyMetereologyAtmosphere, utilitiesCommunication, boundary, farming, biota) belonging to the topicCategory domain could be organized in an ontology. The name of considered categorical attribute has been placed at the top of the Taxonomy as the most generic class; the literals suggested in the ISO are included as classes in the ontology (they are represented as rectangles) and some IS-A relationships are included as arrows starting from subclasses and oriented towards the more generic class. Moreover an abstract class floraFauna, depicted as empty rectangle is introduced to state that “biota” and “farming” are both about vegetal and animals

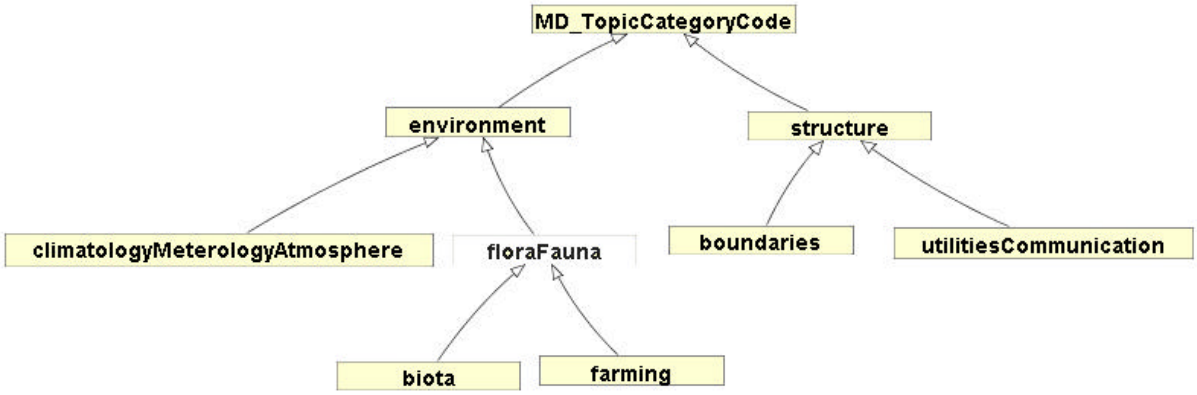

Figure 3: A simple example of ontology about some literals defined in MD_TopicCategory.

Let suppose to have a ISO 19115 metadata item, that has topicCategory value "farming", then to map it as an instance I_1 of the class farming. Other metadata items I_2, I_3 and I_4 are respectively instances of biota, climatologyMetereologyAtmosphere and Boundaries. Thus, computing of the Taxonomic Similarity previously introduced, we obtain:

$$TS(I_1, I_2) = \frac{CM(Class(I_1), Class(I_2))}{2} = \frac{3}{10}$$

$$TS(I_1, I_3) = \frac{CM(Class(I_1), Class(I_3))}{2} = \frac{1}{6}$$

$$TS(I_1, I_4) = \frac{CM(Class(I_1), Class(I_4))}{2} = \frac{1}{12}$$

and $TS(I_1, I_2) > TS(I_1, I_3) > TS(I_1, I_4)$.

Given such a result, it is possible to make machine understandable the fact that the instances of farming are more semantically related to topics belonging to environmental than to boundaries.

6 Discussion and future development

The illustrated approach provides a solution to the problems of data missing and unfamiliar attribute, as:

- the visualizations provide a compact representation of data giving a summarized overview of the values that the attributes assume: this supports in the solving of the unfamiliar attribute problem since it gives an idea about the information the attributes represent.
- the interaction with different visualizations connected by brushing and linking allows to mine the selection criteria applicable to the available data. Along with visualizations, clustering provides a structure which represents a simplification of the data. It eases the explorative analysis allowing to focus on clusters instead of on each single metadata item. Since the selection criteria to solve the data missing and unfamiliar attribute problems depend on user requirements as well as on available data, the combination of these functionalities appear as essential.
- ontologies and the taxonomy similarity measure are applied in order to achieve a machine understandable representation of the semantic relationships among categorical values. To exploit this representation permits either to find a substitute of the data that are missing or to make the selection criteria looser.
- the documentation attached to the ontology definition explains the meaning of literals proposed in ISO specification. This poses the basis to ensure the compliance to ISO metadata standard and to solve the problem of unfamiliar attributes. Any user who is unfamiliar with an attribute can display such a comment to obtain a more precise indication about the meaning of literals.

We have demonstrated the utility of the integration of visual data mining and ontology concepts to analyse geographic metadata, but we have also discovered numerous open issues connected to the development of the approach.

Issues related the use of the ontology:

- different ontologies which deeply differ one another can be defined for the same application context. The construction of an ontology is strongly affected by the domain of knowledge of the community of experts who define it, and the language and the logic used to define it.
- A problem to be solved is what happen when some relationships are not IS-A and how these relationships can be exploited for the exploration.

- the visualization of the ontology is strictly related to the tree structure it is represented with. Novel visualizations should be investigated in order to facilitate the interaction and the navigation in the ontology structure and similarity measure.

Issues related to the mapping of ISO metadata standards:

- Two different data types are defined to express the domain of categorical attribute: <<Enumeration>> and <<Codelist>>. Both of them consist of a list of literals, however, the former is a list the of literals which are specified by the ISO 19115 metadata standard and new values cannot be added, whereas the latter is a list of values (specified by ISO) that can be extended with new items. The approach at the moment does not consider that new literals can be added by the user.
- In metadata standards there are attributes that are not defined as categorical but they are expressed as compound data types (for example, the scale of a map is represented in the domain MD_resolution and defined as union of two values “equivalentScale” and “distance”, the former is the fraction that represents the scale, the latter states the unit of measure of the scale). We can evaluate to extend our approach to those compound attributes that can be categorized.

7 Conclusion

In the paper the importance of defining approaches to analyse geographic metadata has been debated. Two problems that affect the user during the metadata analysis have been highlighted: data missing problem and unfamiliarity of the user with the metadata attribute.

We have described an approach to overcome these problems. In particular it focuses on the analysis of a particular type of attributes of geographic metadata: the categorical attribute. The analysis of this kind of attribute also requires to address aspects of semantic data modelling. The approach aims to provide a visual data exploration: visual data mining is adopted to involve the human in the data exploration, allowing to get insight into data, to recognise patterns and directly interact with them, while ontology concept is proposed to discover the semantic relationships among data.

The proposed approach aims to support the user to navigate in an unfamiliar space: the visualization of the search results and the contemporary use of many visualization techniques enable users to have a compact overview of the available data, to achieve a correct interpretation of the result set, to mine properties and relationship among data and to improve the searching criteria. Moreover he can prevent the problem of data missing using the semantic relationships among categorical attributes that he can discover with the support of ontology.

The approach has been partially developed. From the implementation point of view different roadblocks are emerging, we are thinking to evaluate and to solve them in the future development of the work.

8 ACKNOWLEDGMENTS

The authors acknowledge the support of the European Commission, IST Program. They thank all the partners of the INVISIP project who contributed to the starting tasks of this research: a special thank goes to KTH Team in Stockholm, and UoC Team in Krakow.

9 REFERENCES

[1] Lloyd, C. *Statistical Analysis of Categorical Data.* New York: John Wiley & Sons, 1999.

[2] Hoffman, D.L. "Correspondence Analysis: The Graphical Representation of Categorical Data in Marketing Research." *Journal of Marketing Research,* 213-227

[3] Watts, D.D. Correspondence analysis: a graphical technique for examining categorical data *Nursing Research* 46 (4), 235-239, 1997.

[4] Ankerst M., *Visual Data Mining*, PhD Thesis, Ludwig-Maximilians-Universit¨at, M¨unchen, 2001.

[5] Guarino, N. Formal Ontology and Information System. In *Proceedings of the Formal Ontology and Information System (FOIS'98),*(Trento, Italy, June 6-8,1998), IOS Press, Amsterdam, 3-15.

[6] Hoffman P.E. and Grinstein G.G., A Survey of Visualizations for High-Dimensional Data Mining. In: Fayyad U., Grinstein G. G. and Wierse A., editors, Information Visualisation in Data Mining and Knowledge Discovery, pages 47-82, Morgan Kaufmann Publishers, San Francisco, 2002.

[7] Hand D., Mannilla H. and Smyth P., *Principles of Data Mining*, MIT Press, Cambridge,Massachusetts, 2001.

[8] Albertoni, R., Bertone, A., Demsar, U., De Martino, M., and Hauska, H. "Knowledge Extraction by Visual Data Mining of Metadata in Site Planning" SCANGIS 2003-

[9] Albertoni, R., Bertone, A., Demsar, U., De Martino, M., and Hauska, H. "Visual and automatic data mining for exploration of geographical metadata", *Proceedings of the 6th AGILE*, Lyon, France, April 24-26, 2003.

[10] Weibel, S. Discovering Online Resources. The Dublin Core: A Simple Content Description Model for Electronic Resources, Arts and Humanity Data Service, 1997 http://www.ahds.ac.uk/public/metadata/disc_03.html

[11] Martin, J, *Managing the Database Environment*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[12] ISO 19115, *Geographic Information Metadata*, International Standard Organization, http://www.isotc211.org/, 2003

[13] *Content Standard for Digital Geospatial Metadata*, Federal Geographic Data Committee, National Spatial Data Infrastructure, USA, 2003.

[14] FGDC/ISO *Metadata Standard Harmonization*, Federal Geographic Data Committee, National Spatial Data Infrastructure, http://www.fgdc.gov/metadata/whatsnew/fgdciso.html, 2003.

[15] Swoboda, W., Kruse, F.,. Nikolai, R., Kazakos, W., Nyhuis, D., and Rousselle,H. “The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany”, *Meta-Data'99*, Third IEEE Meta-Data Conference, Bethesda, Maryland April 6-7, 1999.

[16] Stein, D. Geospatial Data Sharing through the Exploitation of Metadata, *ESRI International User Conference*, San Diego, California, July 8-11, 1997.

[17] Card, S.K., Mackinlay, J.D., and Ben Shneiderman , *Readings in Information Visualization*. Using Vision to Think. San Francisco: Morgan Kaufmann, 1999.

[18] Risch, J., May, R., Thomas, J., and Dowson, S., *Interactive Information Visualization for Exploratory Intelligence Data Analysis*. In Proceedings of the Virtual Reality Annual International Symposium, 230—238, 1996.

[19] Andrienko, G., Andrienko, N., and Voss, H., GIS for Everyone: the CommonGIS project and beyond, *Maps and the Internet*, Elsevier Science, 2003, 131-146.

[20] Takatsuka, M., and Gahegan, M. GeoVista Studio: a codeless visual programming environment for geoscientific data analysis and visualization, *Computers & Geosciences N. 28*, Elsevier Science, 2002, 1131-1144.

[21] Hearst, M Tilebars: Visualization of term distribution in full text information access. In *Conference Proceedings Human Factors in Computing Systems*, pages 59–66, New York, 1995. ACM Press.

[22] Wise, J.A. The ecological approach to text visualization, Journal of the American Society for Information Science, 50(13):1224–1233, 1999.

[23] Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. Visualizing the non visual: Spatial analysis and interaction with information from text document. In *Proc. Information Visualization '95*, pages 51--58, October 1995.

[24] Klein, P., Reiterer, H., Muller, F., and Limbach, T. Metadata visualisation with VisMeB. In *Proceedings Seventh International Conference on Information Visualization*, 16-18 July 2003, 600-605.

[25] Seeling, C., and Becks, A. Exploiting metadata for ontology-based visual exploration of weakly structured text documents. In *Proceedings Seventh International Conference on Information Visualization*, 16-18 July 2003, 652-657.

[26] Friendly, M., Visualizing categorical data. In *Cognition and Survey Research*, John Wiley & Sons, Inc., New York, 319-348, 1999.

[27] Greenacre, M.J. *Correspondence Analysis in Practice*, Academic Press, London, 1993.

[28] Kolatch, E., and Weinstein, B., 2001, CatTrees: Dynamic visualisation of categorical data using treemaps http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Kolatch_Weinstein/

[29] Ward, M. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In Proceedings of Visualization '94, 326-333.

[30] Ma, S. and Hellerstein, J.L. Ordering categorical data to improve visualization. In *IEEE Information Visualization Symposium Late Breaking Hot Topics*, 15-18.

[31] Gruber, T.S., Toward principles for the design of Ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, 43, 5 (November 1995), 907-928.

[32] Noy, F. N. and McGuinness, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Technical Report KSL-01-05 Stanford Knowledge Systems Laboratory, Stanford, 2001.

[33] McGuinness, D. L. Ontologies Come of Age in *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. MIT Press, 2002.

[34] Maedche, A. and Zacharias, V. Clustering Ontology-Based Metadata in the Semantic Web. *Lecture Notes in Computer Science (LNCS)* , 2431, Springer (2002),348-360

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

[36] Albertoni,R., Bertone, A., and De Martino, M. A Visualization-Based Approach to Explore Geographic Metadata, WSCG Posters proceedings WSCG'2003, February 3-7, 2003, Plzen, Czech Republic

[37] Fraley, C. Algorithms for Model-Based Hierarchical Clustering, *SIAM J. Sci. Comput.* 20, 1, (1998) 279–281.

[38] Meila, M. and Heckerman, D. An Experimental Comparison of Several Clustering and Initialisation Methods. In *Proc. 14th Conf. on Uncert. in Art. Intel*., Morgan Kaufmann, 1998, 386–395.

[39] Jain, A.K., and Dubes, R.C. *Algorithms for Clustering Data*, Englewood Cliffs, N.J., Prentice-Hall, 1988.

[40] Podolak, I. and Demšar, U. Discovering structure for geographic metadata, accepted for *Proc. 12th Int. Conf. on Geoinformatics - Geospatial Information Research: Bridging the Pacific and Atlantic (Geoinformatics 2004),* University of Gävle, Sweden, 7-9 June 2004.

[41] Kreuseler M. and Schumann H., A Flexible Approach for Visual Data Mining, IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 1., pp. 39-51, 2002.

[42] Mackinlay, J.D., Robertson, G., and Card, S.K. “The Perspective Wall: Detail and Context Smoothly Integrated”, *ACM Conference on Human Factors in Computing Systems (CHI '91)*, 1991, 173-179.

[43] Johnson, B., and Shneiderman, B. "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures." *Proceedings of IEEE Information Visualization ’91,* 1991, 275-282

[44] Sprenger T. C., Brunella R. and Gross M. H., H-BLOB*:* A hierarchical visual clustering method using implicit surfaces, IEEE Visualization 2000, Salt Lake City, Utah, USA, Proceedings. IEEE Computer Society and ACM, pp. 61-68, 2000

[45] Demšar, U. A Visualisation of a Hierarchical Structure in Geographical Metadata. To appear in: *Proceedings of the 7th AGILE Conference on Geographic Information Science*, Chania, Greece, April 2004.

[46] Mutton, P. Goldbeck, J. Visualizing of Semantic Metadata and Ontologies. *In Proceedings of the Seventh International Conference on Information Visualization (IV’ 03).*(London, England, July 16 - 18, 2003), IEEE Computer Society , 300-306.

[47] Fluit, C. Sabou, M. Van Harmelen, F, Ontology-based Information Visualization. *In Visualising the Semantic Web*, Springer Verlag, 2002