Semantic Analysis of Categorical Metadata to Search for Geographic Information

Riccardo Albertoni, Alessio Bertone, Monica De Martino
Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche, Genova, Italy
{albertoni, bertone, demartino}@ge.imati.cnr.it

Abstract

The paper proposes a semantic-based approach to analyse geographic metadata during the information search activity. In particular, the analysis of categorical attributes of metadata is considered.

Techniques and concepts of Information Visualisation and ontology are analysed and exploited to facilitate the navigation in unfamiliar spaces of geographic data. The approach underlies how semantic information could be exploited by using information visualization techniques to provide the knowledge for the formulation of search criteria.

1. Introduction

The paper addresses issues of metadata analysis to search for geographic information. In particular, the analysis of the categorical attributes is considered.

To search effectively for geographic information in an open and distributed environment is a crucial task. It involves a remarkable interaction between users and information sources situated in different WWW locations. Metadata are used since they provide information about geographic data. Metadata analysis is an important task to support the user in seeking and retrieving information. The large set of geographic data, its heterogeneity as well as the large amount of geographic providers determine the generation of large metadata database. Moreover the complexity of geographic data leads to metadata having several attributes with different representations: numerical, descriptive and categorical.

Users are often unable to formulate pointed questions or express them effectively using conventional information retrieval system, which are based on query-string matching. Metadata analysis tools to support user in the search criteria are needed.

Different approaches to metadata analysis have been defined according to the representation of the metadata attributes. Most of them focus on the analysis of geographical data expressed as numerical values and text [1], [2]. On the contrary, less interest has been posed on categorical data. Categorical data or nominal data are data that can be separated into different categories according to some non-numeric characteristics. Their exploration is challenging because the values that they assume provides information that can be easily understood by a human agent but cannot be trivially managed in automatic way. Some approaches to visualize categorical attributes have been proposed [3], [4]. Unfortunately none of them is based on semantic relations among the categorical values, whereas an explicit representation of the relations existing in the user cognitive space should facilitate the exploration of metadata and improve the effectiveness of search criteria.

The paper proposes a semantic approach to analyse categorical attribute based on visualization and ontology: it aims to improve seeker ability in the formulation of the query, to understand the search results and to reformulate the query. The approach exploits Ontology concept [5] to represent the semantic relations among data, and uses Information Visualization [6] as a communication channel between the computer and the user to provide an interactive visualization of the Ontology: it aids to discover and validate new and interesting patterns among data. The proposed approach is an extension of our research described in [7], where a metadata analysis framework based on Information Visualization is proposed. The paper is organized as follows: in the first part a short overview of our previous research and the main concepts of Ontology are provided. In the second part our new research about semantic exploitation in order to discover new knowledge for the formulation of search criteria is described.

2. Metadata analysis framework

This research activity started within the EU founded research project INVISIP ”Information Visualization for Site Planning”. The result of the activity is the design of a metadata analysis framework to allow the seeker to move through large information spaces in a

1529-4188/05 $20.00 © 2005 **IEEE**

flexible manner without feeling lost. An exploration approach is defined: it is characterized by a reasoning activity based on the integration of visualization techniques, graphic interaction and brushing and linking functionality. The approach is widely illustrated in [7] and an example of how it can be used to solve some problems of geographical information search is described in [8].

The approach has some limitations: it does not provide any knowledge about the semantic relations among the nominal values which represents precious hints for the formulation of new search criteria. We propose to overcome this limitation integrating in the framework a component for a semantic-based analysis. The aim is to support the seeker when the criteria used to formulate the query fail: the system is able to provide a semantic knowledge of similar terms that could be adopted in the refinement of the query. The semantic knowledge is detected by applying similarity criteria among data and is provided to the user through visualization techniques to amplify his cognition and to facilitate the interpretation of the query results.

3. Semantic relations among data

In this paragraph we provide some concepts concerning the representation of semantic relations among data needed to design our approach: the notions of ontology and similarity measure are provided.

The ontology represents an emergent manner to formally describe concepts and relations among entities within a specific domain. It 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 [9].

In this paper ontology is adopted to represent the relations among the nominal values assumed by a categorical attribute of geographic metadata. The ontology is composed by *class entities* (named *classes)* representing the most important concepts of the domain described by the ontology, *instances* representing specific elements of classes, and *slots* which can be *attributes* characterizing the classes or *relations* representing types of interaction among concepts. The classes can be related by *Is-a or part-of* relations.

Semantic similarity facilitates the comparison among the class entities and allows to handle those which are semantically similar. In the paper the Matching-Distance Measure for Semantic Similarity (MDMS) [10] is considered and is defined in terms of slots comparison. Slots are classified according to three different types of features called *distinguishing features*: "function" features which are relations with specific properties describing what is done to or with a class, "part" features (a part-of relation) describing structural elements of a class and "attribute" features which represent class properties. Two entities are more or less similar according with the number of slots belonging to the same kind of distinguishing features they share each other.

A formal definition of "global similarity" is based on the definition of two measures: the "slots importance" and the "slots similarity".

Definition 1: function α of "slots importance"
Let us call c_1, c_2, two class entities, d the distance function between the two class entities and lub the immediate super-class that subsumes both classes. α is the function that evaluates the importance of the difference between the two class entities in term of their slots and it is defined by:

$$\alpha(c_1,c_2)=\begin{cases}\dfrac{d(c_1,\mathrm{lub})}{d(c_1,c_2)} & d(c_1,\mathrm{lub})\le d(c_2,\mathrm{lub})\\[2ex] 1-\dfrac{d(c_1,\mathrm{lub})}{d(c_1,c_2)} & d(c_1,\mathrm{lub})>d(c_2,\mathrm{lub})\end{cases}$$

It is important to note that the computation of the distance d consider both is-a and part-of relations to determine the immediate super-class *lub*.

Definition 2: "slots similarity"
Given two class entities c_1 (target) and c_2, (base), t one type of distinguishing features (part, function, attribute) and C_1 and C_2 the sets of features of type t respectively of the class entities c_1 and c_2 . The similarity value of c_1, c_2 is:

$$S_t(c_1,c_2)=\frac{|C_1\cap C_2|}{|C_1\cap C_2|+\alpha(c_1,c_2)|C_1\setminus C_2|+(1-\alpha(c_1,c_2))|C_2\setminus C_1|}$$

Definition 3: "global similarity"
Given two class entities c_1 and c_2, w_p, w_f, w_a the weight of the respective importance of parts, functions and attribute, the global similarity function S between two class entities c_1 and c_2 is the weighted sum of the similarity values of parts (S_p), function (S_f) and attribute (S_a):

$$S(c_1,c_2)=\omega_p\cdot S_p(c_1,c_2)+\omega_f\cdot S_f(c_1,c_2)+\omega_a\cdot S_a(c_1,c_2)$$

The sum of weights is expected to be equal to 1 and the value of each is calculated according to contextual information [10].

4. Semantic analysis of metadata

We propose an approach to analyse categorical attributes of geographic metadata that exploits the semantic relations among the categorical values. The notions of Ontology and MDMS similarity are adopted to express the relations and to make them machine understandable, while Information Visualization is applied to improve similarity cognition of the seeker.

The analysis is performed in three phases:

- definition of the ontology,
- exploitation of ontology to explicit the relations of the categorical values,
- exploration of semantic relation among categorical values.

The ontology definition is obtained by mapping the categorical values of a metadata attribute into class entities. A class hierarchy is built adding Is-a relations. Each metadata description having categorical values represented as class entity is defined as instance of the class. Moreover, as the similarity among classes is defined in terms of slots comparison, a careful definition of slots for each class is needed. Slots grouped according with their functions, parts and attributes should be associated to each class entity also taking into account the intended similarity assessment: in other words, if two classes are expected to be similar they should be forced to share some slots.

The exploitation of ontology to work out the semantic similarity is based on MDMS similarity: it adopts the asymmetric similarity, which seems to be more appropriate to support a query refinement based on the distance among concepts [11]. The MDMS allows to explicit the similarity among categorical values, which is a semantic relationship that usually is not available from the ontology design.

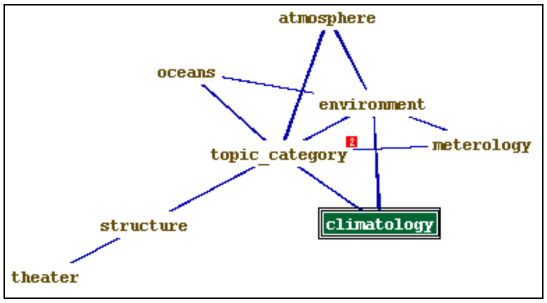

Figure 1: Ontology visualization with Protégé plug-in "TgVizTab".

In the exploration of semantic relations the main issue is to choose appropriate visualization techniques to display the similarity measure providing useful hints in the query refinement. Visualizations such as the cluster maps[15] are developed to support the query refinement but they still do not visualize any information about class similarity. Other ontology visualizations are mainly based on a graph representation where a node can be either an instance or a class entity and edges can be either properties or relations. The graph visualization helps the seeker to analyse and better understand the domain described in the ontology, but it does not provide any support in the query refinement process. For example Protégé [12] offers different visualizations[13]: Figure 1 shows the interactive and graph-based visualization to browse the ontology provided by protégé plug-in TgVizTab [14]. This visualization is inadequate to provide an explicit representation of the semantic relations: it shows the structure of the ontology where each node is a theme and each line is an "is-a" or "part of" relation. It does not provide any interpretation of classes in terms of similarity. It supports the engineers during the ontology design rather than the seeker in the query refinement.

We propose a visualization aimed also to facilitate the seeker in recognizing similarities among the categorical values. It is characterized by the integration of a graph visualization to show the overall structure of the ontology and graphic techniques to represent similarity information.

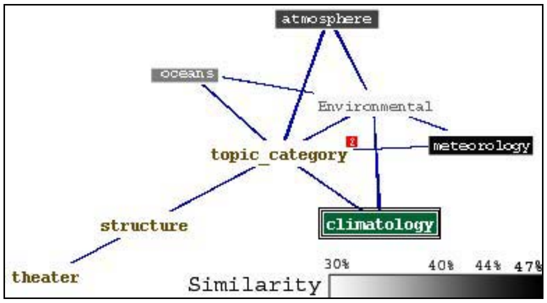

Figure 2:Ontology and Similarity Visualization.

Figure 2 depicts the visualization of an ontology enhanced using similarity information. The target entity (first variable in the similarity function) is the class entity "climatology", the similarities are worked out with respect to it and it is marked with a double-squared rectangle. The other entities considered as base in the similarity (second variable in the similarity function) and whose similarity measures are different from zero are represented by rectangles with different grey level colours. The grey level is brought down proportionally to the decrease of similarity between target and base: the more similar are climatology and the class entity, the darker is the box surrounding the class in the visualization.

4.1. Example

This paragraph aims to clarify the semantic based analysis with a practical example. The example is applied at the categorical attributes of the ISO 19115 metadata standard [16]. In particular, the attribute "topicCategory", which describes a high-level classification for geographic data themes, is analysed.

The first phase of the approach is to develop the ontology of the data themes with the identification of the distinguishing feature for each class entity. In this example we presents a "toy" ontology (Figure 3). The definition of a complete ontology of theme would require a long interactive design process where experts of the domain have to be directly involved. It requires efforts out of the purpose of the paper.

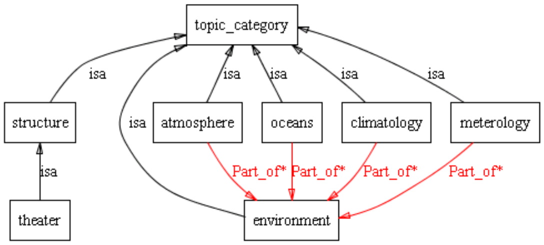

Figure 3: "topic_category" ontology example.

Figure 3 shows a subset of the categorical values defined in the metadata specification as possible themes and their organization in ontology in terms of Is-a and Part-of relations.

Table 1: Distinguishing features of the ontology entities.

Class entity	Parts	Functions	Attributes
Topic Category			
Environment	Climatology Atmosphere Meteorology Oceans	Environnemental assessment Climate phenomena analysis Monitoring environmental risk	Ex_GeographicExtent Ex_TemporalExtent Variable
Atmosphere		Air quality analysis Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Atmosphere layers Temperature
Meteorology		Weather forecast Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Precipitation
Climatology		Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Temperature Precipitation Wind
Oceans	Sea life	Tidal wave forecast Tide analysis Climate phenomena analysis	Ex_GeographicExtent Ex_TemporalExtent Temperature Wind Water composition Sea level
Structure			Material Location
Theatre	Foundation Roof Ticket office	Perform Present Recreate	Material Location Height

Table 1 shows the distinguishing features of each data theme. Note that the entity "topic category" is an abstract class that cannot be instantiated and does not have parts, functions and attributes. It is represented in the ontology mainly for technical reasons, it is an explicit reference to the metadata attribute, which is considered, and it can be useful to contextualize the ontology in the overall metadata schema.

The second phase of the approach concerns the ontology exploitation to explicate the semantic relations among the categorical values. Let us suppose to analyse the semantic relations among the theme "climatology" and the other themes. Table 2 shows the similarity measure between "climatology" and the other entities belonging to the ontology applying the MDMS. The similarity values are calculated considering the ontology graph in Figure 3 and the distinguishing features in Table 1. To provide a simple example the global similarity function S (a,b) have been calculated considering all the weights w_t equal to one third. The result shows that the topic "climatology" is more similar to "environmental" than to "structure" or "theatre". The same happens for "meteorology", "atmosphere" and "oceans". Furthermore Table 2 also quantifies the similarity between the themes: "climatology" is more similar to "meteorology" than "atmosphere", "atmosphere" than "oceans", "oceans" than the generic environment.

Table 2: Similarity measures between the theme "a", "Climatology", and the theme "b".

b	α	S_p(a,b)	S_f(a,b)	S_a(a,b)	S(a,b)
Environment	0,00	0,00	0,33	0,67	0,33
Meteorology	0,50	0,00	0,67	0,75	0,47
Atmosphere	0,50	0,00	0,67	0,66	0,44
Oceans	0,50	0,00	0,50	0,72	0,40
Theater	0,33	0,00	0,00	0,00	0,00
Structure	0,50	0,00	0,00	0,00	0,00

The third phase of the approach concerns the presentation of similarity information to the seeker. In this example the simple visualization illustrated in the previous paragraph is applied. The TGVizTab visualization depicted in Figure 1 can provide a compact overview about the domain, which is useful to get the contextual information. On the contrary, the visualization proposed in Figure 2 provides also similarity information through grey level colour of the box surrounding the entities. For example "meteorology" has a surrounding black box since "climatology" is more similar to the "meteorology" than "atmosphere", "oceans" and "environmental" whereas the colour of surrounding box of "atmosphere", "oceans" and "environmental" discolours according to decreasing of the similarity decrease. Moreover, entity classes like "structure" and "theatre" do not have any surrounding box since the similarity is equal to zero. Even if it can appear a trivial representation, this kind of visualization provides a useful support in the query refinement. In other terms, the similarity exploitation in the metadata analysis makes machine understandable the fact that when the user is searching for data having "climatology" as theme and he gets unsatisfying results, the system suggests him to refine his query. The proposed visualization provides suggestions of adopting "meteorology", "atmosphere", "oceans" and "environmental" as possible refinement keywords.

5. Semantic visualization in the metadata analysis framework

The ontology and its visualization is integrated in the metadata analysis framework mentioned in the paragraph 2. Adding simple interaction functionalities the seeker can interact at the same time with ontology visualization and the other visualizations provided by the tool to explore the metadata attributes. The general purpose is that the contemporary use of different visualization techniques enable seekers to have a compact overview of the available data, to achieve a correct interpretation of the result set, to mine properties and relation among data. In particular, the ontology visualization prevents him from the problem of missing data [8] suggesting new search criteria.

6. Conclusion and future work

In this paper a metadata analysis approach to facilitate the query refinement process in the geographic information search is described taking into account the semantic information of the metadata attributes. The use of MDMS similarity is proposed to handle similar entities and Information Visualization is applied to facilitate the cognition of similar entities.

On going work will investigate novel visualizations to facilitate the interaction and the semantic navigation of the ontology instances.

7. Acknowledgement

This research is partially supported by the European Commission within the INVISIP project and the Network of Excellence AIM@SHAPE.

8. Bibliography

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

[2] M. Hearst, "Tilebars: Visualization of term distribution in full text information access". In *Conf. Proc. Human Factors in Computing Systems*, ACM Press, NY, 1995, pp. 59–66.

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

[4] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, and M. O. Ward, "Mapping nominal values to numbers for effective visualization", *Symposium of Information Visualization,* IEEE, 2003, pp. 113-120.

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

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

[7] R. Albertoni, A. Bertone, U. Demsar, M. De Martino, and H. Hauska, "Knowledge Extraction by Visual Data Mining of Metadata in Site Planning", *SCANGIS* 2003, pp. 119-130.

[8] R.Albertoni, A.Bertone, and M.De Martino, "Visual analysis of geographic metadata in a spatial data infrastructure", *15th International Workshop on Database and Expert Systems Applications(DEXA 2004),* IEEE, 2004, pp. 861-865.

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

[10] M.A.Rodriguez and M.J.Egenhofer, "Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure, "*Int. Journal of Geographical Information Science*, vol.18, no.3, 2004, pp. 229-256.

[11] M.J. Egenhofer and D.Mark "Naive geography". *Spatial Information theory :A theoretical basis for Geographical Information System*, Int. Conf. COSIT '95, Semmering, Austria. *LNCS 988*, Springer-Verlag, Berlin. pp. 1-15.

[12] M.A. Musen, R.W. Fergerson, N.F. Noy, and M. Crubezy, "Protege-2000: A plug-in architecture to support knowledge acquisition, knowledge visualization, and the semantic Web", *Journal of the American Medical Informatics Association*, 2001, pp. 1079.

[13] A.Ernst, M. Storey, and P. Allen,"Cognitive Support for Ontology Modelling",*Int.J.Human-Computer Studies*, 2004.

[14] TGVizTab, www.ecs.soton.ac.uk/~ha/TGVizTab, 2004.

[15] H. Stuckenschmidt, F.van Harmelen, A.de Waard, T. Scerri, R.Bhogal, J.van Buel, I.Crowlesmith, C.Fluit, A. Kampman, J.Broekstra, and E.van Mulligen,"Exploring large document repositories with RDF technology: the DOPE project",*Intelligent Systems,*vol.19,no.3,*IEEE*,2004,pp.34-40.

[16] ISO19115, *Geographic Information Metadata*, International Standard Organization, http://www.isotc211.org/, 2003.