Using formal ontology for integrated spatial data mining

Sungsoon Hwang

Department of Geography
State University of New York at Buffalo
105 Wilkeson Quad, Buffalo, NY 14261, U.S.A.
E-mail: shwang5@buffalo.edu

Abstract. With increasingly available amount of data on a geographic space, spatial data mining has attracted much attention in a geographic information system (GIS). In contrast to the prevalent research efforts of developing new algorithms, there has been a lack of effort to re-use existing algorithms for varying domain and task. Researchers have not been quite attentive to controlling factors that guide the modification of algorithms suited to differing problems. In this study, ontology is examined as a means to customize algorithms for different purposes. We also propose the conceptual framework for a spatial data mining (system) driven by formal ontology. The case study demonstrated that formal ontology enabled algorithms to reflect concepts implicit in domain, and to adapt to users' view, not to mention unburdened efforts to develop new algorithms repetitively.

1 Introduction

No single spatial data mining method is best suited to all research purposes and application domains. However, determining which algorithm is suited to a certain problem, and how to set the values of parameters is not a straightforward task. Rather it requires an explicit specification of domain-specific knowledge as well as of task-oriented knowledge. Given this problem, ontology, which is defined as "the active component of information system" [1] in addition to "the explicit specification of conceptualization" [2], can play an important role in organizing the mechanism underlying the spatial data mining phenomenon. Spatial data mining can be thought of as an information system where different kinds of ontologies serve as active components. In this context, spatial data mining system is driven by formal ontology.

This study is concerned with endowing algorithms with semantics and adapting algorithms to users view rather than developing new algorithms for different domains and problems. The focus is placed on the role of ontology in customizing existing algorithms. Thus, the purpose of this study is to illustrate how formal ontology can be used to re-use existing algorithms suited to varying domain and task. To make this clear, we propose a conceptual framework for spatial data mining system driven by formal ontology. The framework will clearly show how ontologies can be incorporated into spatial data mining algorithms. The conceptual framework is

Hwang, S (2004), Using Formal Ontology for Integrated Spatial Data Mining, In: Laganà A et al (ed) *Computational Science and Its Applications – ICCSA2004*, LNCS Vol. 3044, Springer-Verlag, pp. 1026-1035

implemented in finding hot spots of traffic accidents. We evaluate whether using ontology is beneficial in spatial data mining by comparing ontology-based method and existing methods. A case study shows that spatial data mining methods using ontology can take into account domain-specific concepts and users view that have not been handled well before. In short, results (i.e. pattern discovered) are both natural and usable.

The rest of this paper is organized as follows. It begins by describing the conceptual framework for ontology-based spatial data mining in Sect. 2. In Sect. 3, the case study using the proposed method is illustrated, and results are analyzed. We conclude by summarizing the study.

2 Conceptual Framework for Ontology-based Spatial Data Mining

2.1 Relation between Data Mining and Ontology Construction

Data mining enhances the level of understanding by extracting a high-level knowledge from a low-level data [3]. Ontology construction makes implicit meaning explicit by formalizing how the knowledge is conceptualized. Here we first discuss different notion of knowledge between data mining and ontology construction. Second, we discuss how they are related. Third, we examine how data mining can be used for ontology construction, and vice versa.

In the context of data mining, knowledge is *discovered*. In the context of ontology construction [4], knowledge is *acquired*. Wherein different notion of knowledge can be noted; the knowledge discovered from data mining is, to a large extent, data-specific whereas the knowledge acquired for ontology construction is data-independent. This fact arises from different approaches taken. Data mining is a bottom-up (i.e. data-drive) approach while ontology construction is a top-down approach. In terms of cognitive principle, data mining is similar to induction while ontology construction is similar to deduction. The knowledge discovered from data mining is applied in small scope with high level of detail whereas the knowledge acquired for ontology construction is applied in large scope with low level of detail.

The next question is, how are they interrelated? Data mining and ontology construction is bridged by the varying level of abstraction. Fig.1 shows that two kinds of knowledge are at the continuum that is manipulated by the level of abstraction. Where the knowledge is the result of data mining, and becomes the source of ontology construction. Not surprisingly, high level of abstraction in ontology construction entails human intervention such as expert knowledge or concept hierarchy. In contrast, data mining can be, to some extent, automated by machine learning programs.

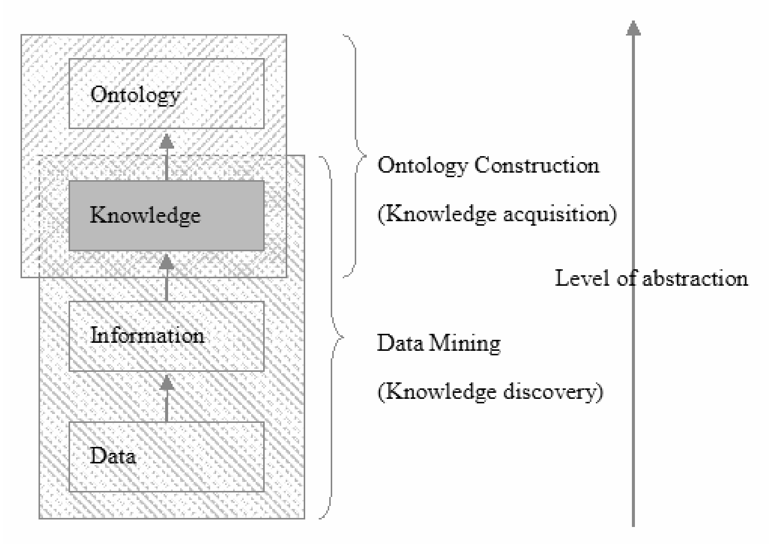


Fig. 1. The relation between data mining and ontology construction seen from varying level of abstraction

The potential role of ontology in data mining is (1) to guide algorithms such that they can be suitable for domain-specific and task-oriented concepts (2) to provide the context in which the information or knowledge extracted from data is interpreted and evaluated [5]. Conversely, the knowledge confirmed in the knowledge discovery process can be seen as candidates for ontology in the long term. In such a way, the interaction between induction (knowledge discovery) and deduction (ontology construction) process can enrich our knowledge base.

Along this line, examining the role of ontology in data mining is significant in explicating the interaction between knowledge discovery and ontology construction. Null hypothesis can be stated as follows: ontology-based data mining method does not improve the quality of knowledge discovery as compared to data mining without using ontology. The study is focused on the first role of ontology. The study is restricted to the spatial data mining. In sum, research questions we attempt to address here are, (1)“Can ontology really enhance spatial data mining?” (2)“If yes, how can it be done?”

2.2 Rationale of Using Ontology for Spatial Data Mining

The same spatial data mining algorithms [6] can yield results inconsistent with fact without considering the domain knowledge. The same data may have to be mined in different ways depending on users’ goals. In sum, the domain and users’ need associated with input data are the major factors that control the mechanism underlying the spatial data mining [7]. Therefore, we need to build a new spatial data mining algorithm that is dictated by domain and task model (Fig.2) [8] [9]. The focus is placed on customizing existing algorithms suited to a certain domain and problem rather than developing new algorithms.

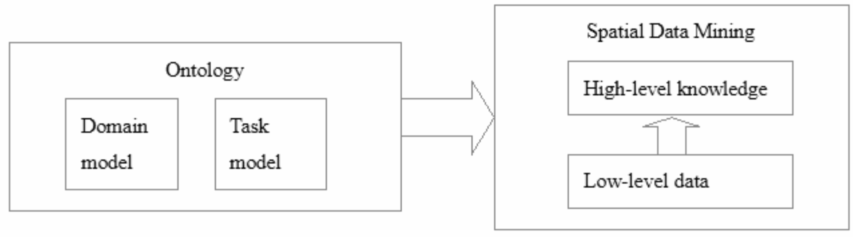


Fig. 2. Using ontology for integrated spatial data mining

To illustrate the domain model incorporated into spatial data mining algorithms, suppose we have two different geospatial data for the purpose of detecting their hot spots: one is traffic accident data, and the other is the location of supermarket. Wherein, traffic accident *is-an* event whereas a store *is-a* physical object. Thus we can conclude they should be handled differently. Moreover traffic accidents can only occur on the road whereas a store is located in the pedestrian blocks. As a result, the topological relation to road network is different: traffic accident occurs *in* road network, but store is located *outside of* road network. To build clustering algorithms, we need to define features to compose similarity measures (or distance function). Different features of similarity measures should be used due to different domain model (or conceptualization) associated with the input data.

To illustrate the task model incorporated into spatial data mining algorithms, let us consider two different tasks of spatial clustering (see [10] for the survey of spatial clustering algorithms): One is to detect hot spots, and the other is to assign customers to market areas. The number of clusters, *k* can be derived from overall data distribution for the former task. In contrast, the number of clusters, *k* will be preliminarily given as a resource constraint for the latter task. In addition to different goals, the number of cluster, *k* will vary with the level of details in which users want to examine. Therefore, different notion of arguments (e.g. *k*) should be adopted depending on task model.

Users are often prompted to specify input parameters without understanding the mechanics of parameters when they use spatial data mining tools. For example, classic *k*-means clustering algorithms prompt users to specify the number of clusters. But the best number of clusters should not be chosen arbitrarily. Rather, the number of clusters should be obtained as a result of learning the data or underlying phenomenon. Likewise, the desired level in hierarchical clustering (i.e. cutting the link in dendrogram) should not be chosen arbitrarily, but rather should be chosen depending on the level of details in which users intend to examine the problem in hand.

Best spatial data mining methods will be achieved by taking into account the factors such as users view, goal, domain-specific concepts, characteristics of data, and available tools. However, if considerations were given to those factors in an arbitrary manner, it would not meet our desire. Given this, ontology can provide the systematic way of organizing those factors. The need for users to specify the input parameters will be reduced if we are able to utilize available ontology in a generic level (e.g. top-level ontology of space and time, domain ontology) [11] and to construct ontology in a specific level (e.g. application ontology) [1]. Therefore, ontology-based spatial data mining can overcome the shortcomings of existing spatial data mining methods.

2.3 Conceptual Framework for Ontology-based Spatial Data Mining

The component of ontology-based spatial data mining systems can be divided into three parts: Input, Ontology-based spatial data mining method, and Output (Fig.3).

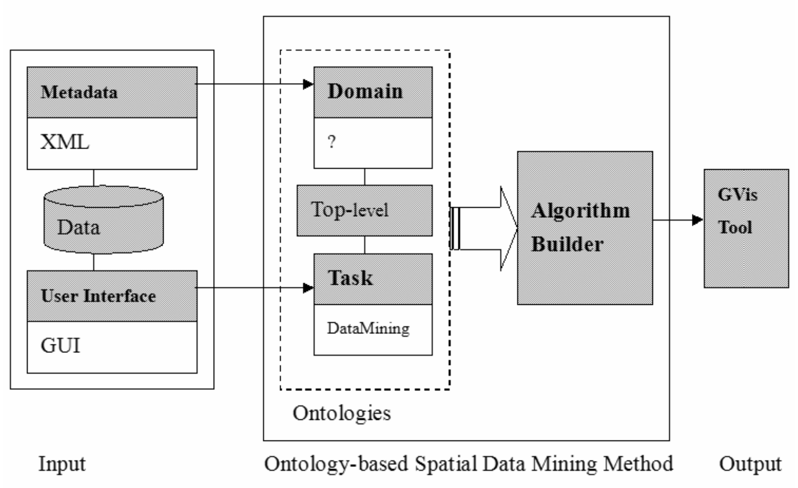


Fig. 3. Conceptual Framework for Ontology-based Spatial Data Mining System

Input component is composed of metadata and user interface that are linked to data. Metadata contains the information on data content. User interface allows users to select a goal to be achieved and related parameters wherever necessary. Ontology-based spatial data mining method is composed of ontologies and an algorithm builder. The natural language given by metadata (domain) and users (task) is translated into corresponding components (domain/task ontologies). Task ontologies define methods adequate to the goal specified. Domain ontologies specify domain-specific concepts, relation, function, and properties. The generic characteristics inherit from the top-level ontologies [12]. Task ontologies interact with domain ontologies to filter the relevant information to activate operations defined in the method. In a spatial data mining algorithm builder, algorithms are dynamically built from the items (method, concept inherent in domain) derived from ontologies. Output component presents results through geographic visualization (GVis) tools. The remainders of this section describe each component in more detail.

Metadata. Metadata is a summary document providing content, quality, type, creation, and spatial information about a data set. A parser program can easily retrieve the theme of data in hand utilizing the tag structure in XML. In such a way, metadata informs domain ontologies of the semantics of data.

User Interface. User interface allows users to effectively explore and select information relevant to a task in hand. Task-centered user interface prompts users to select their goals. Moreover, users can select the level of detail (e.g. jurisdiction level), and the geographic area of interest.

Domain Ontologies. Terms within the "theme" tag in the metadata are used as a token to locate the appropriate domain ontologies. Domain ontologies specify their definition, class (e.g. Accident is a Subclass-Of Temporal-Thing) and properties (e.g. Road has a Geographic-Region as a Value-Type) (see [13] for the ontology of geography). Properties of class inherit from upper-level ontologies.

Task Ontologies. The goal selected by users in the user interface is translated into task ontologies. The method, requirement, and constraints adequate to the user-supplied goal are specified in task ontologies. A certain class requires the interaction with domain ontologies (also known as interaction problem). For example, in order to detect the hot spots of traffic accidents, algorithms need to get the information from domain ontologies, such as spatial constraints (e.g. traffic accident occurs along the road segments) [14] [15].

Algorithm Builder. The method (hierarchical) suited to user-supplied goal ("find hot spots"), requirements (data, scale, detail level), and constraint (road) has been already supplied from the interaction between input components (metadata, user interface) and ontologies (domain, task, top-level). Algorithm builder puts together these items to build the best algorithm. That way, ontology provides the arguments necessary for running algorithms without a need to prompt users to select them. More specifically, the algorithm builder filters the data content through domain ontology and users' requirement through task ontology.

GVis Tool. The geographic visualization tool displays the resulting clusters. Results will be displayed differently depending on goal and scale. Well-designed GVis tool facilitates hypothesis formulation, pattern identification (that was the purpose of spatial clustering task also) and decision-making.

3 Case Study

3.1 Ontology-based Spatial Clustering Algorithm

We used 7413 geocoded fatal accident cases that have been reported in New York State from year 1996 to 2001 (see [16] for data description; see [17] for georeferencing procedures). Fig.4 shows (a) input data and (b) output clusters in Buffalo, NY, where output clusters are generated by ontology-based spatial clustering (OBSC) algorithms. All components (with a focus on spatial clustering) discussed in the preceding section are combined to find the hot spots of traffic accident in the case study.

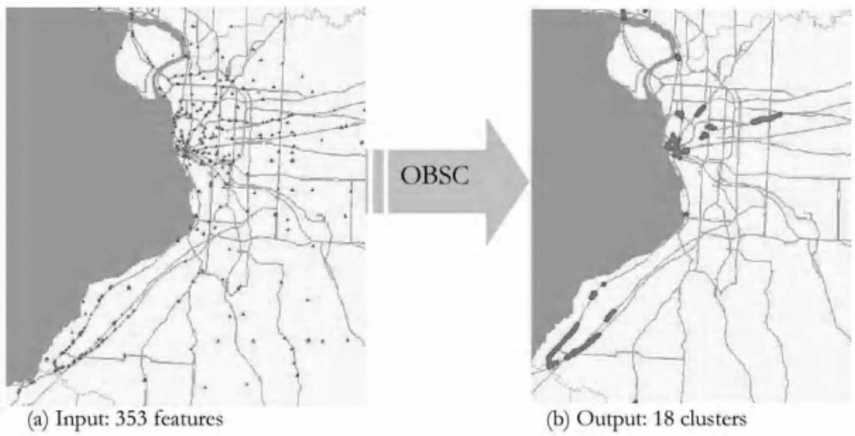
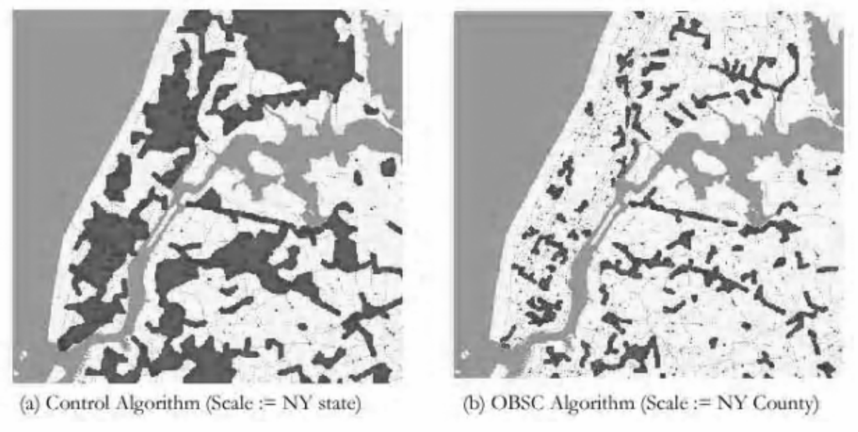


Fig. 4. The Result of Ontology-Based Spatial Clustering Algorithm

To evaluate the benefit of using ontologies, we compare a control algorithm (i.e. without using ontologies) [18] with a test algorithm (OBSC) in terms of Geographic-Scale and Spatial-Constraint with other factors controlled. Scale is implicit in task specification, and constraint is given in domain ontologies.

3.2 Analyses of Results

Effect of Scale (Task Ontologies). To illustrate the point, suppose a user wants to pinpoint the spot where traffic accidents occur with higher frequency in Manhattan, not other localities. In Fig.5, map (a) results from a control algorithm, and map (b) results from an ontology-based algorithm. Two algorithms are the same except that an ontology-based algorithm prompts a user to choose geographic scale of his interest. Two maps show clusters in different level of detail. With a control algorithm, traffic accident cases are lumped into three large clusters, which mask the detail. It is mainly due to the averaging effect of fixed scale. On the other hand, an ontology-based algorithm discovers hot spots of traffic accidents in the desired level of detail because the necessary information is conveyed to task ontologies through a user interface. To recapitulate, the result of an ontology-based algorithm reflects spatial distribution specific to the scale of users' interest, thereby resulting in *usable* clusters.

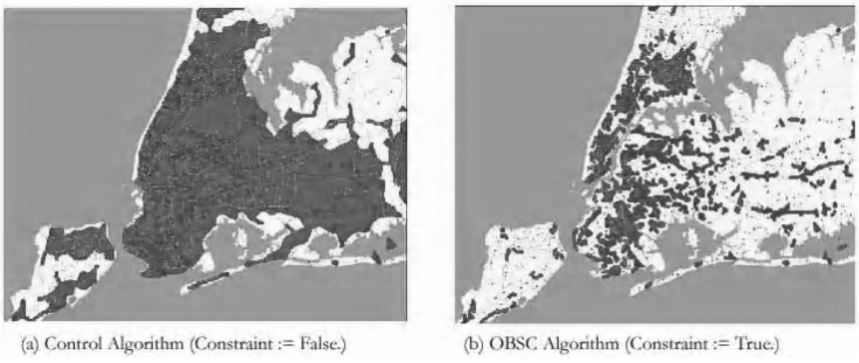


(a) Control Algorithm (Scale := NY state)

(b) OBSC Algorithm (Scale := NY County)

Fig. 5. OBSC clusters reflect spatial distribution specific to the scale of users' interest

Effect of Constraint (Domain Ontologies). In Fig.6 (a), no consideration of spatial constraint (i.e. the occurrence of accidents is spatially constrained on the road) produces a large cluster spanning both sides of New York harbor. The control algorithm overlooks the existence of a body of water between Manhattan and Brooklyn. On the other hand, an ontology-based algorithm separates clusters because domain ontologies inform the algorithm that an accident cannot be on the body of water. It shows that domain ontologies enable clustering algorithms to embed domain knowledge, thereby resulting in *natural* clusters.



(a) Control Algorithm (Constraint := False.)

(b) OBSC Algorithm (Constraint := True.)

Fig. 6. OBSC clusters identify the physical barrier due to concept implicit in domain

4 Conclusion

This study demonstrates that it is worthwhile using ontology in a spatial data mining task in several respects: (a) Ontology provides the systematic way of organizing various features that consist of mechanism underlying the data mining phenomenon. (b) Ontology-based methods produce more intuitive results. (c) The need to specify input parameters arbitrarily is reduced. (d) Ontology provides the semantically plausible way to re-use existing algorithms.

Findings can be summarized as follows: First, in ontology-based method data mining mechanisms are dictated by concepts implicit in domain. For instance, the resulting clusters of traffic accidents are concentrated along road network because a spatial constraint is a priori implicit in domain. Second, ontology-based method is responsive to users view. The user-supplied task requirements make a cut-off value depend on the distribution specific to the scale of users' interest. To sum up, ontology-based methods make the result of spatial data mining natural and usable.

This study can advance the field of geographic knowledge discovery by introducing a novel approach to data mining methods based on ontology. This study is important because the attempt has been made to present the mechanism that ontologies are incorporated in spatial data mining algorithms in a way that algorithms can be built at a semantic level. Formalizing knowledge (i.e. ontology construction) is not a focus of this study, but the semantic linkage between ontologies and algorithms through the parameterization process has been emphasized.

References

1 Gruber, T. R., 1993, Formal Principles for the Design of Ontologies Used for Knowledge Sharing, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academic Publishers

2 Guarino, N., 1998, Formal Ontology and Information Systems, *Proceedings of FOIS'98*, Trento, Italy, 6-8 June 1998, Amsterdam, IOS Press, pp. 3-15

3 Fayyad, U. M, Piatetsky-Shapiro, G., and Smyth, P., 1996, From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp.1-34

4 van Heijst, G., Schreiber, A., and Wielinga, B., 1997. Using explicit ontologies for KBS development, *International Journal of Human-Computer Studies*, 46(2/3): 183--292.

5 Visser, U., Stuckenschmidt, H., Schuster, G., and Vogele, T., 2002, Ontologies for Geographic Information Processing, *Computers & Geosciences* 28: 103-117

6 Koperski, K, Adhikary, J., and Han, J., 1996, Knowledge Discovery in Spatial Databases: Progress and Challenges, *Proceedings of Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 55-70, Montreal, QB, June 1996.

7 Witten, I. H. and Frank, E., 2000, *Data Mining: practical machine learning tools and techniques with java implementations*, Morgan Kaufmann

8 Newell, A., 1982, The Knowledge Level. *Artificial Intelligence*, 18, pp. 87-127

9 van de Velde, W., 1993, Issues in Knowledge Level Modeling, in David, J-M., Krivine, J-P., and Simmons, R. (Eds.) *Second Generation Expert Systems*, pp.211-231. Springer Verlag, Berlin

10 Han, J., Kamber, M., and Tung, A.K H., 2001, Spatial Clustering Methods in Data Mining: A Survey, in H. Miller, J. Han, (Eds.) *Geographic Data Mining and Knowledge Discovery*, Research Monographs in Geographic Information Systems, Taylor and Francis

11 Russell, S., and Norvig, P., 1995, *Artificial Intelligence: A Modern Approach*, Prentice Hall

12 Fonseca, F. T., Egenhofer, M. J., Agouris, P, and Gamara, G, 2002, Using Ontologies for Integrated Geographic Information Systems, *Transactions in GIS*, 6(3): 231-57

13 Smith, B. and Mark, D. M., 2001, Geographic Categories: an Ontological Investigation, *International Journal of Geographical Information Science*, 15(7): 591-612

14 Tung, A K H, Hou, J., and Han, J., 2001, Spatial Clustering in the Presence of Obstacles, *17th International Conference on Data Engineering* April 02 - 06, 2001 Heidelberg, Germany

15 Estivill-Castro, V and Lee, I., 2001, AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles, in Roddick, J. F., and Hornsby K. (Eds.) *Temporal, Spatial, and Spatio-Temporal Data Mining*, Springer-Verlag pp.133-146

16 NHTSA, 1995, FARS 1996 Coding and Validation Manual, National Center for Statistics and Analysis, National Highway Traffic Safety Administration, Department of Transportation, Washington, D.C.

17 Hwang, S., and Thill, J-C, 2003, Georeferencing FARS accident data: Preliminary Report, NCGIA and Department of Geography, State University of New York at Buffalo, Unpublished document

18 Kang, I., Kim, T., and Li, K., 1998, A Spatial Data Mining Method by Delaunay Triangulation, *Proceedings of the 6th ACM GIS Symposium*, pp.157-158, November 1998