

Ontology-Based Spatial Clustering Method: Case Study of Traffic Accidents

Julie Sungsoon Hwang
Department of Geography
State University of New York at Buffalo
105 Wilkeson Quad, Buffalo, NY 14260
E-mail: shwang5@buffalo.edu

Abstract. This paper explores the role of ontology in clustering, a key task in spatial data mining. Some spatial clustering algorithms could produce clusters inconsistent with fact without considering domain knowledge. The same data may have to be clustered in different ways depending on users' objective. Ontologies can play an important role in organizing the mechanism underlying the clustering phenomenon. It is argued that incorporating domain ontologies and task ontologies in spatial clustering algorithms can enhance the quality of clusters. Incorporating ontologies enables algorithms to reflect concepts implicit in domain, and to adapt to users view. In the case study, the constraint intrinsic in domain, and the scale implicit in task specification have been found to affect the quality of clusters. It suggests that it is promising to organize arguments required for algorithms within the framework of ontologies. Given this, ontologies imbue algorithms with semantics and viewpoint.

1. Introduction

Spatial clustering groups similar spatial objects into classes that are not pre-defined. Spatial clustering can be used to find the natural clusters (e.g. land use type extraction from the satellite imagery, merging regions with similar weather patterns), identify hot spots (e.g. epidemics, crime, traffic accidents), and partition the area based on utility (e.g. market area assignment by minimizing the distance to customers). Spatial clustering is a preliminary step that invokes further analysis by facilitating hypothesis formulation. In spite of the importance of spatial clustering, many clustering algorithms proposed so far have been vulnerable to criticisms. Criticisms arise from the fact that a clustering task is not dictated by the nature of the mechanisms that are thought to underlie the particular clustering phenomenon (Witten & Frank, 2000). There is a need for making spatial clustering driven by the mechanism underlying the clustering phenomenon – I argue that they are mainly organized into two parts: One is domain-

related (i.e. how the discourse of domain is organized) and the other is task-related (i.e. how the goal-achieving procedure is reasoned).

No single spatial clustering 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” (Guarino, 1998) in addition to “the explicit specification of conceptualization” (Gruber, 1993), can play an important role in organizing the mechanism underlying the clustering phenomenon. Spatial clustering can be thought of as an information system where different kinds of ontologies serve as active components. In this context, spatial clustering is driven by ontology.

This approach can be differentiated from other approaches in that it is concerned with endowing algorithms with semantics and adapting algorithms to users view rather than developing algorithms suited to a certain domain and problem. This paper is aimed at proposing a conceptual framework for spatial clustering system driven by formal ontology. The framework will clearly show how ontologies can be incorporated into spatial clustering algorithms. More specifically, the natural language given in metadata (domain) and user interface (task) is translated into corresponding ontologies. Relevant concepts are derived from ontologies such that they consist of arguments necessary for an algorithm. Results show that spatial clustering methods using ontology can take into account domain-specific concepts and users view that have not been handled well before. In short, resulting clusters are natural and usable.

Fatality Analysis Reporting System (FARS) (NHTSA, 1995) provides a good source for traffic accident analysis. FARS contains fatal traffic crashes that have occurred in the United States since 1975, and is composed of data in four levels - accident, person, driver, and vehicle. The accident level data have been georeferenced (Hwang & Thill, 2003), and these geospatial data are used to illustrate how an ontology-based method works.

The rest of the paper is organized as follows. It begins by discussing related works in the section 2. The section 3 describes the conceptual framework for ontology-based spatial clustering. In the section 4, the case study using the proposed method is illustrated, and the result analysis is given. I conclude by summarizing the project and remarking on further works.

2. Related Works

2.1 Spatial Clustering

Spatial clustering is a way of grouping spatial entities into the naturally homogeneous group without the need to specifying the predefined class – clustering is an unsupervised task in that sense. Spatial clustering can serve different purposes such as identifying natural classes (e.g.

satellite image processing), detecting hot spot (e.g. epidemiology, crime analysis), partitioning similar areas (e.g. demographic market segmentation), summarization (e.g. grouping similar weather pattern), and detecting outliers. The clustering is achieved by maximizing the intra-cluster similarity, and the inter-cluster dissimilarity. Spatial clustering typically deals with one-dimensional numeric attributes, which is the Euclidean distance between spatial objects.

In data mining, clustering methods are divided into three types (Witten & Frank, 2000). Spatial clustering follows this convention (Han, Kamber, and Tung, 2000). The first one is a partitioning method. Classic k -means (or k -medoid) methods belong to this type. The partitioning method creates k -clusters from n -instances by iteratively assigning instances to their closest cluster center according to the ordinary Euclidean distance function until they stabilize. The second one is the hierarchical method. This method groups data objects into t tree of clusters by merging or splitting cluster(s) incrementally. In contrast to partitioning methods, data objects can be decomposed into a multi-level nested tree like a dendrogram. The third one is the statistical method. Wherein instances are assigned to classes probabilistically, not deterministically based on the mixture model of different probability distributions, unlike the other two methods.

A partitioning algorithm is relatively efficient, but often terminates at a local optimum and is not suitable to discover non-convex clusters. In a hierarchical algorithm, it is hard to choose merge/split points even though merging/splitting decisions are critical. Moreover, it is sensitive to order of input records. Statistical methods use probability measurements to determine the cluster or concept. However they assume attributes are independent of each other and all are distribution-based even though most of distributions are unknown.

Other clustering methods have been developed based on the notion of density (Ester, et al, 1996). Connectivity between clusters (e.g. epidemic spread path, traffic accident along the road segments) may also have to be taken into account. Imposing constraints (e.g. river, road, topography) may yield totally different results from those without constraints (Tung, et al, 2001). Some domains are more likely to undergo spatial autocorrelation (e.g. epidemic spread) than others (e.g. traffic accidents). That is, characterization of clusters needs to be considered prior to choosing algorithms. Some algorithms produce pre-specified number of clusters, while others automatically calculate the natural number of clusters (Estivill-Castro & Lee, 2001, Kang, et al, 1998).

Factors to be considered for choosing a right clustering algorithm are two-faceted. One is the domain in hand, and the other is the task in hand.

Different similarity functions can be used depending on the domain. For example, the similarity function used to cluster traffic accidents can be based on the notion of distance while taking into account the connectivity between accidents or spatial constraints – accidents occur along the road segments unless specified otherwise (i.e. by default). Likewise, the similarity function for clustering epidemic occurrences may be defined based on the closeness where

temporal sequences of each occurrence have to be considered. Spatial dependency between close occurrences may have to be considered as well. Even if different domains are represented as the same feature type (e.g. point), they all involve different conceptualizations. Traffic accident is considered *event* whereas epidemic or weather pattern is considered *process*. The conceptualization of domain provides different notion of similarity.

Clustering method depends on the task. Detecting crime hot spots requires a certain notion of density (e.g. more than median) in addition to the closeness whereas allocating a certain number of ATM to potential customer sites does not. Outliers handling policy is also different depending on the task. Hot spot detection algorithms treat objects below a cut-off value, as outliers while partitioning algorithms do not allow for outliers. For the former problem, the number of clusters, k is derived from the data distribution (e.g. mean, standard deviation). In the latter problem, k is a resource constraint (e.g. the number of ATM). So the necessity to specify k depends on the nature of task in hand. Moreover, a resolution or scale is an important notion to be considered (e.g. merge/split point) in the task specification because how much detail users want will significantly affect the results. For example, in small-scale map, smoothing techniques such as kernel density estimation may be useful, but in large scale map smoothing would mask the detail since smoothing degrades the resolution, thereby leading to the loss of information.

2.2 Ontologies

The term, ontology has been borrowed from the field of philosophy studying the existence. In the context of artificial intelligence, existence means what can be represented. The set of objects that can be represented is called the universe of discourse. Ontology can be described by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (i.e. classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Thus, the problem of implicit and hidden knowledge can be overcome by making the conceptualization explicit (Visser, 2002). In sum, ontology can be defined as the explicit specification of conceptualization (Gruber, 1993).

Ontology belongs to the knowledge level specification. The knowledge level, according to (Newell, 1982), is a level of description above the symbol level. He advocates the existence of knowledge independent of its symbol level description. The symbol level provides a means to 'mechanize' a behavior. The symbol level does not make a claim about the real nature of the agent. The symbol level is system oriented, whereas the knowledge level is world oriented (Van de Velde, 1993). Since different knowledgeable agents will often have different symbol level representations, it is convenient to formulate ontologies at the knowledge level (Van Heijst, 1997). Ontologies try to capture the intended (or generic) meaning of terms in a way that they

serve as a mediator between the world and their representation.

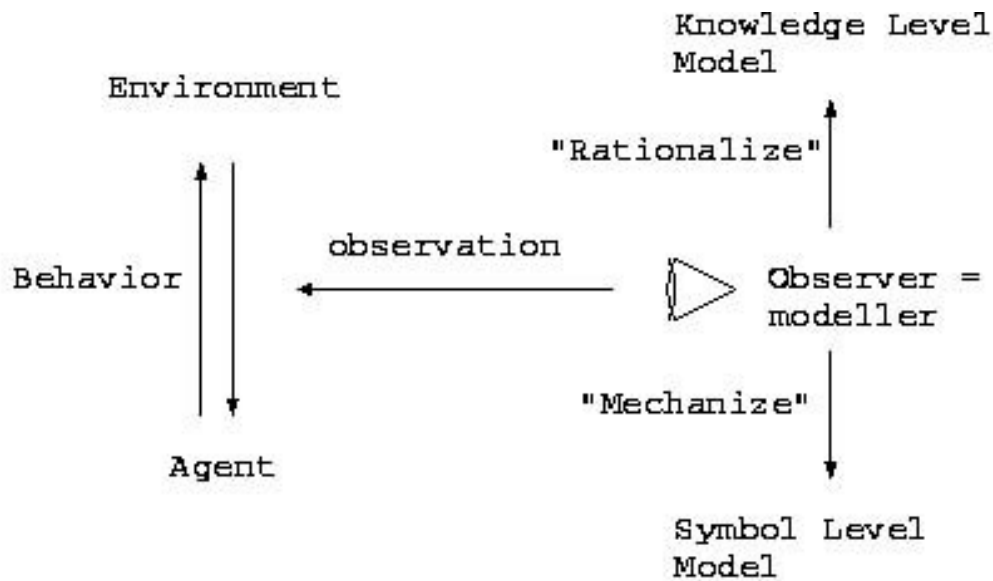


Figure 1. Newell's Knowledge Level Hypothesis (Van de Velde, 1993)

The knowledge level provides the means to 'rationalize' the behavior of a system from the standpoint of an external observer. The principle of rationality can be decomposed into two steps; not only apply knowledge, but also configure the knowledge into the problem. It reflects the fact that adequate behavior is both rational and practical. The second step allows the agent to reach a limited range of goals in a limited range of situations. Figure 2 shows that knowledge and goals are respectively organized in domain models and task models. A domain model is a means to talk about domain knowledge in a precise and systematic way. It expresses what one assumes collections of statements about the world to mean. A task model is a means to talk about goals in a precise and systematic way. It expresses what it means to achieve a goal and how the goals are interrelated. A problem solving method is a means to relate task and domain models in order to accomplish goals. It describes roles that task and domain models have to play in order to achieve goals (Van de Velde, 1993).

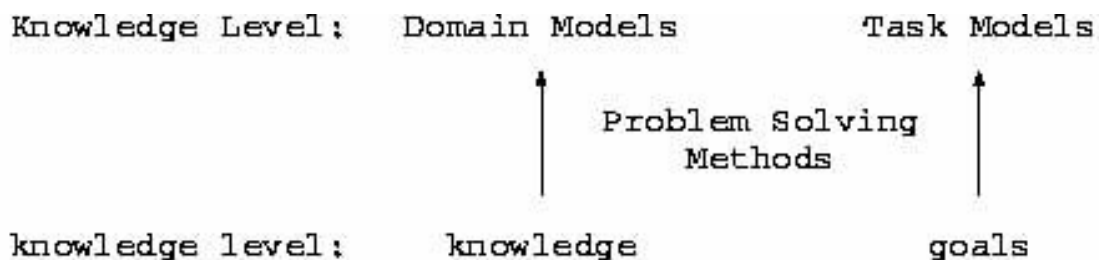


Figure 2. Domain Models and Task Models (Van de Velde, 1993)

Different kinds of ontologies can be defined as a way of conceptualizing their dependence and hierarchy. Figure 3 shows the kinds of ontologies. Top-level ontologies describe very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain. Domain ontologies and task ontologies describe, respectively, the vocabulary related to a generic domain (like accident, or medicine) or a generic task or activity (like accident analysis, or diagnosing), by specializing the terms introduced in the top-level ontology. Application ontologies describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies.

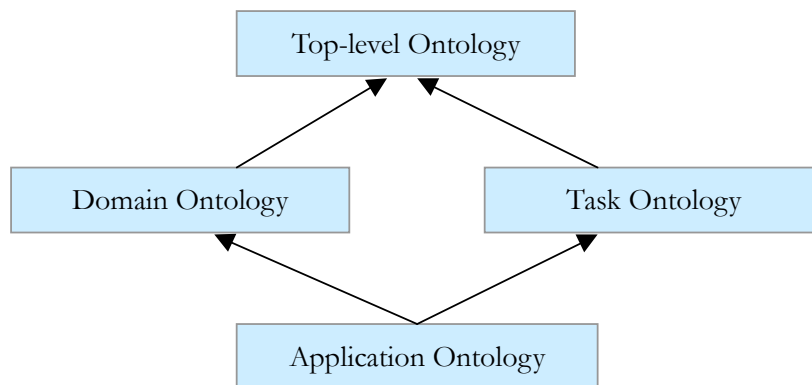


Figure 3. Kinds of Ontologies according to their level of dependence on a particular task or point of view vertically, and their subject horizontally. The arrows represent specialization relationships. (Guarino, 1998)

Ontologies have been used in disciplines like knowledge sharing/reuse, the development of knowledge-based system, and semantic web. In the real world applications, ontologies can be classified according to the type of structure of the conceptualization (Van Heijst, 1997). Terminological ontologies specify the terms like a thesaurus. For example, the information on place names can be retrieved using these types of ontologies. Information ontologies specify the record structure of databases like database schemata. Metadata can serve this purpose. Knowledge modeling ontologies specify conceptualization of the knowledge. For example, “traffic accidents” tagged by <theme> in XML can be conceptualized as a subclass of event whereas “stores” are conceptualized as a kind of physical object.

In this paper, domain/task ontologies at the level of knowledge modeling are applied to spatial clustering applications. I adopt the view that domain knowledge cannot be represented independently of assumptions of how it will be used in reasoning (also known as interaction problem) for a practical reason. It may be undesirable for ontologies reuse, but it helps select relevant items in domain ontologies dependent on the task instead of specifying complete items.

3. A Conceptual Framework for Ontology-Based Spatial Clustering

3.1 Data Mining and Ontologies

In the context of data mining, knowledge is discovered. In the context of ontologies construction, knowledge is acquired. Wherein different notion of knowledge can be noted; the knowledge discovered is data-specific whereas the knowledge acquired is data-independent. This fact arises from the different approaches taken. Data mining is a bottom-up (i.e. data-drive) approach whereas ontologies construction is a top-down approach. However, the gap is only due to the different level of abstraction. Generic ontologies are located in the very end of generic knowledge which data mining ultimately pursues. The knowledge from data mining and that from ontologies have different uses. The former applies to a limited range of domain in more detail while the latter applies to generic domain in less detail. The former has small coverage (or scope) with high resolution while the latter has large coverage with low resolution.

	Data Mining	Ontologies Construction
Process	Discovered	Acquired
Approach	Bottom-up	Top-down
Data	Data-specific	Data-independent
Coverage	Small	Large
Resolution	High	Low

Figure 4. Comparison of the notion of knowledge in Data mining vs. Ontologies

Figure 4 clearly shows that two kinds of knowledge are at the continuum that is only manipulated by the level of abstraction. The role of ontologies in data mining is to provide the context in which the knowledge discovered is interpreted and evaluated in the short term. Conversely, the knowledge confirmed in the knowledge discovery process can be seen as candidates for ontologies in the long term. The interaction (or interference) between induction (knowledge discovery) and deduction (ontologies construction) process can proliferate our knowledge base in a general context (not in the context of expert system).

Along this line, exploring the role of ontologies in data mining can be significant in enlightening the short-term interaction between knowledge discovery and ontologies construction. I narrow down the scope to the following research question: “What can ontologies do for spatial clustering?” or “Can ontologies really enhance spatial clustering?”

3.2 Rationale of Using Ontologies for Spatial Clustering

Users are often prompted to specify input parameters in using spatial clustering tools even

though sometimes they do not understand their meaning. For example, classic k -means clustering algorithms require users to specify the number of cluster. But the best number of clusters cannot be chosen arbitrarily. Rather, the number of clusters should be given as a result of learning the data or underlying phenomenon. Likewise, the desired level in the case of hierarchical clustering (i.e. cutting the link in dendrogram) should not be chosen arbitrarily, but rather be specified in a way that the level fits the users view.

In some cases, input parameters are implicitly given without a need for users to specify them. For example, the fact that traffic accidents occur along road segments is already given in the domain. Moreover, spatial representation of accident is typically a point, which enables clustering algorithms to involve simpler consideration of topology. Furthermore, metadata that contain the information on data quality and size can be used to optimize algorithms.

Best clustering 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 work. Ontology can provide the systematic way of organizing those factors so that they can be contributed to the results. The need for users to specify the input parameters will be reduced if we are able to utilize available ontologies in a generic level (i.e. top-level ontology of space and time, domain ontologies) and to construct ontologies in a specific level (i.e. application ontologies such as FARS analysis). The input parameters necessary for running algorithms can be derived in a well-grounded manner by letting different ontologies communicate to each other. Therefore, ontology-based spatial clustering can overcome the shortcomings of existing spatial clustering methods.

3.3 Conceptual Framework for Ontology-Based Spatial Clustering

The component of ontology-based spatial clustering systems can be divided into three parts: Input, Ontology-based spatial clustering method, and Output. 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 specify a goal to be achieved and related parameters if necessary. Ontology-based spatial clustering 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 contain domain-specific concepts, relation, function, and properties. The generic characteristics inherit from the top-level ontologies. Task ontologies interact with domain ontologies to filter the relevant information to enable the operations defined in the method. In a spatial clustering algorithm builder, algorithms are dynamically built from the items (method, concept inherent in domain) derived from the ontologies. Output component is the geographic visualization tool for displaying the resulting

clusters.

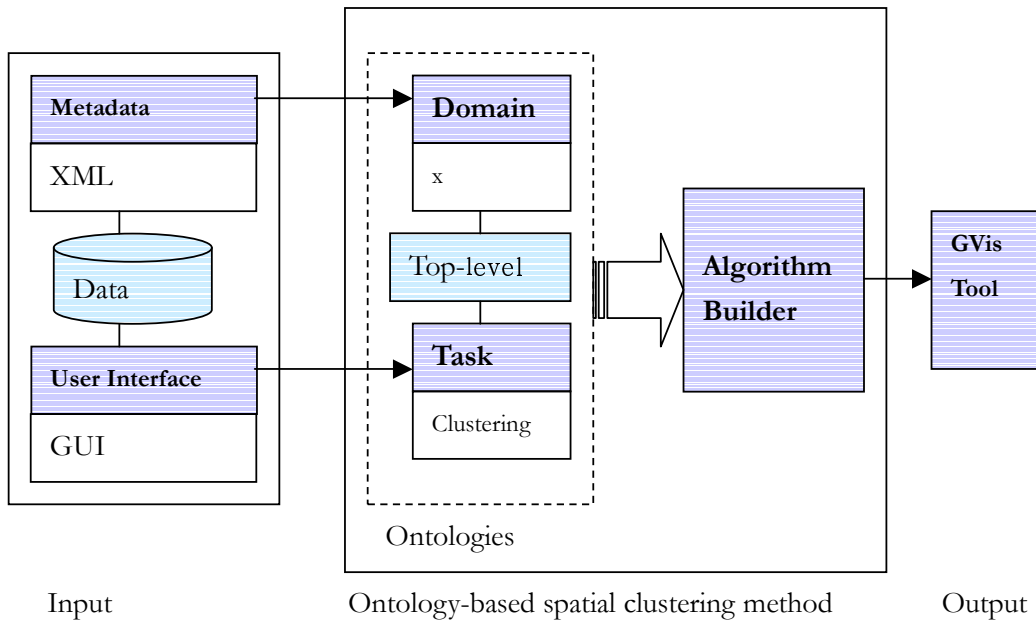


Figure 5. Conceptual Framework for Ontology-based Spatial Clustering System

Metadata: Metadata is a summary document providing content, quality, type, creation, and spatial information about a data set. Figure 6 shows the part of metadata stored in XML in conformance to FGDC standard, FGDC-STD-001-1998. “Keywords” tag contains three elements – theme, place, and temporal. Utilizing tag structure, a parser program informs domain ontologies of the semantics of data.

User Interface: User interface allows users to effectively perceive and express information. Task-centered user interface prompts users to select a goal - find hot spots, group similar patterns, and partition into k -cluster. Users can specify the geographic area of interest, and the level of detail.

Domain Ontologies: Terms given in 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). Properties of class inherit from upper-level ontologies. For example, Named-Roadway is a subclass of Path-For-Wheeled-Vehicles from which the characteristics inherit.

```

- <keywords>
  - <theme>
    <themekt>Road Accident</themekt>
    <themekey>Traffic Accident</themekey>
  </theme>
  - <theme>
    <themekt>Auto Crash</themekt>
    <themekey>Motor Vehicle Crash</themekey>
  </theme>
  - <place>
    <placekey>New York State</placekey>
  </place>
  - <temporal>
    <tempkey>1996-2001</tempkey>
  </temporal>
</keywords>

```

Figure 6. Metadata in XML format facilitates accessing the semantics of data

Class Named-Roadway

- **Defined in Ontology:** Hpkb-upper-level-kernel-latest
- **Source code:** hpkb-upper-level-kernel-latest.lisp

Documentation:
 The collection of named parts (stretches or segments) of roadways. (Not roadway the stuff.) Instances are named pieces of roadways (or highways or streets) which have names and lengths.
 Value-Type: *String*

Instance-Of:
 Existing-Object-Type, Primitive, *Class*, *Collection*, *Intangible*, *Mathematical-Or-Computational-Thing*, *Object-Type*, *Set*, *Stuff-Type*, *Temporal-Stuff-Type*
 Value-Type: *Class*

Name-In-Cyc: #*\$*NamedRoadway

Subclass-Of:
Street-Generic, *Ecological-Region*, *Geographical-Region*, *Human-Residence-Area*, *Humanly-Occupied-Spatial-Object*, *Individual*, *Outdoor-Location*, *Partially-Tangible*, *Path-For-Wheeled-Vehicles*, *Something-Existing*, *Spatial-Thing*, *Temporal-Thing*, *Urban-Area*

Template Slots:

Borders-On:
 Value-Type: *Geographical-Region*

Geographical-Sub-Regions:
 Value-Type: *Geographical-Region*

Axioms for Named-Roadway:
 (Nth-Argument-Name Named-Roadway 1 ?X)

Figure 7. Domain ontologies specify definition, class, and properties (Ontolingua Server)

Task Ontologies: The goal specified by users in the user interface is translated into task

ontologies where methods suitable for the goal are explicitly specified. The method, requirement, and constraint adequate to the user-supplied goal are specified in task ontologies. A certain class requires the interaction with domain ontologies: For example, in figure 8 Spatial Objects that play a role of Constraint are linked to domain ontologies.

Theory SPATIAL-CLUSTERING-TASK

Documentation:

This theory defines a task ontology for the spatial clustering task. The spatial clustering task, which is a class of clustering task, is a problem of grouping similar spatial objects into classes.

Super classes: Clustering

Subclasses:

Sub goal:

“Find hot spots”

“Group similar patterns”

“Partition into k -clusters”

Requirement:

Assignment-Object

Source: Spatial Objects

Target: Clusters

Geographic-Scale

Detail-Level

Constraint:

Spatial Objects

Operational Constraints

Methods:

Partitioning Method

Hierarchical Method

Statistical Method

Figure 8. Task Ontologies specify the generic structure of a class of problems

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.

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.

4. Case Study

4.1 Illustrated Examples

When data is created, metadata is produced as well. Parser program associates data with appropriate domain ontologies by the token of <theme> in metadata stored in XML. A good thing about using domain ontologies is that it allows for utilizing the inheritance feature given by taxonomic structure. Building Traffic Accident domain ontologies only requires specifying properties specific to the domain since generic properties are already defined in the upper-level ontologies, and it is sufficient to have a link to the upper or lower level ontologies. Figure 9 reproduces domain ontologies of Traffic Accident in part using first-order logic. This domain inherits generic properties from the upper-level ontologies such as Accident, Vehicle, and Roadway. For instance, Accident is a subclass of Event. Event has interval and duration as a slot (i.e. properties specific to the class). Thus Traffic Accident also has interval properties as shown in figure 9. Domain ontologies describe their properties in an explicit way using knowledge representation language.

Dimension	Knowledge Representation Language	Natural Language Translation
Space	$\text{Point}(x) \wedge \text{On}(x, y) \wedge \text{Roadway}(y)$ $\text{Line}(y) \wedge \text{In}(y, z) \wedge \text{Geographic-Region}(z)$	Traffic Accident is a point, and occurs on Roadway. Roadway is in Geographic-Region.
Time	$\text{Point}(x) \wedge \text{At}(x, y) \wedge \text{Time}(y)$ $\text{Event}(x) \Leftrightarrow \text{Occurrence}(x) \vee$ $\text{Notification}(x) \vee \text{Response}(x) \vee \text{Arrival}(x)$ $\text{Before}(\text{Occurrence}(x), \text{Notification}(x))$	Traffic Accident occurs at a point of time. Accident event is composed of sub events such as Occurrence, Notification, Response, and Arrival Notification follows Occurrence
Attribute	$\text{Accident } x \wedge \text{RelatedTo}(x, y) \wedge \text{Vehicle}(y)$	Traffic Accident is a kind of Accident, and vehicle-related.

Figure 9. Domain ontologies of Traffic Accidents

User interface enables users to specify the goal and geographic scale. Available goals are listed in figure 8. Geographic scale can be selected among different jurisdiction levels such as State, County, and Town. Spatial Clustering task ontologies specify the methods suited to goals. The notion of cluster is quite different depending on the goal. The goal - “Find hot spots” requires a cut-off value to extract dense clusters from *some* spatial objects whereas the goal - “Partition into k -cluster” requires k to assign *all* spatial objects to k -clusters. Different goal setting leads to a different set of arguments. User interface allows users to customize algorithms suited to goal and geographic location of interest. Without these components (user interface and task ontologies), the algorithm would not be able to capture users view.

Arguments required for algorithms are derived from input components (metadata, user

interface) and ontologies (domain, task). Suppose users specify the goal (“Find hot spots”) and Geographic-Scales (Erie County) in the user interface. Task ontologies tune algorithms to the user-supplied arguments: (a) Hierarchical method is suited to a specific goal. (b) Operations require a cut-off value to find denser clusters. (c) Clusters below a cut-off value are treated as outliers. A cut-off value is arithmetically derived from the distribution specific to the Geographic-Scales. To find out if the operation has Spatial Constraints, task ontologies interact with domain ontologies associated with data in hand. In response to the query of task ontologies, Traffic Accident domain ontologies inform task ontologies that Traffic Accident has spatial constraint that is roadway. That way, a resulting algorithm takes into account goal, geographic-scale, and constraint through the interaction between components of ontology-based spatial clustering system.

The resulting clusters are visualized using geographic visualization tools. Same clusters can be interpreted in a different way depending on how they are visualized. Optimal visualization method can be seen as a function of goal, geographic scale, and data characteristics. Suppose users’ goal is to partition into k -clusters. Each resulting cluster is displayed as a different color to indicate their unique identity. If the goal is to detect hot spots in large area, the resulting clusters are displayed as smoothed grid cells with varying shades to indicate different magnitudes in each cluster.

4.2 Result Analysis

I used 7413 cases that have been reported to occur in New York State from year 1996 to 2001. Figure 10 shows the input data and output clusters in Erie County, where output clusters are created by ontology-based spatial clustering (OBSC) method.

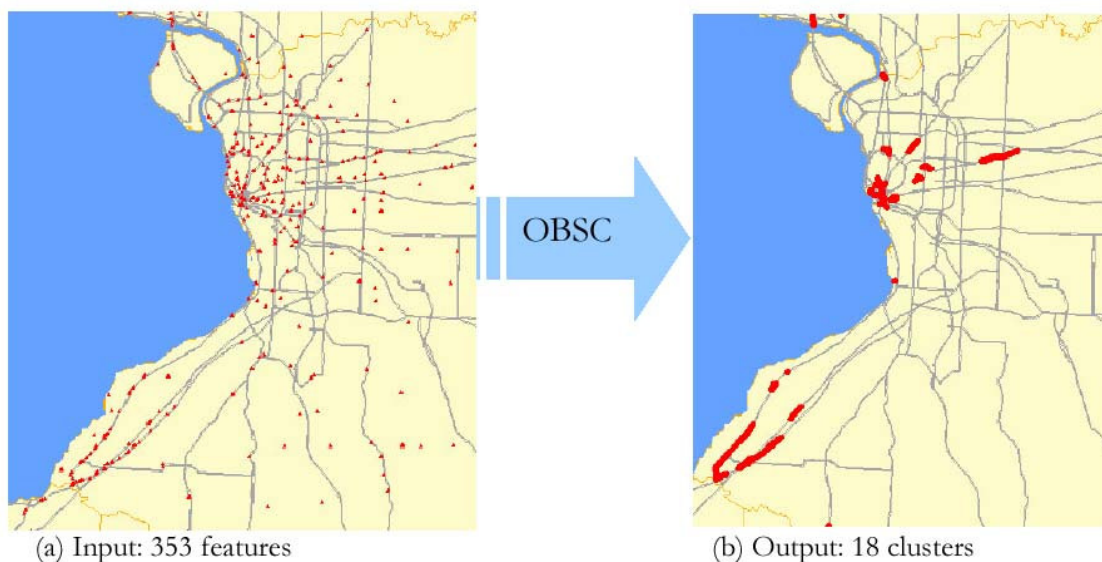


Figure 10. The Result of Ontology-Based Spatial Clustering Algorithm

To evaluate the benefit of using ontologies, I compare a control algorithm (i.e. with no use of ontologies) with a test algorithm (OBSC) in terms of Geographic-Scale and Spatial Constraint while other factors being controlled. Scale is implicit in task specification, and constraint is given in domain ontologies.

4.2.1 An 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, and he is not interested in other areas. In figure 11, 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. Map (a) is characterized by a highly generalized spatial pattern whereas map (b) shows more details. An appropriate level of detail can only be accounted for by users' objective. Map (a) does not meet the purpose of pinpointing hot spots in local scale due to the averaging effect of fixed scale. On the other hand, map (b) meets the requirement of discovering unique pattern of traffic accidents in a relevant level of detail. To recapitulate, a spatial clustering algorithm that incorporates task ontologies allows scale to be flexible by making algorithms reflect spatial distribution specific to the scale of users' interest.

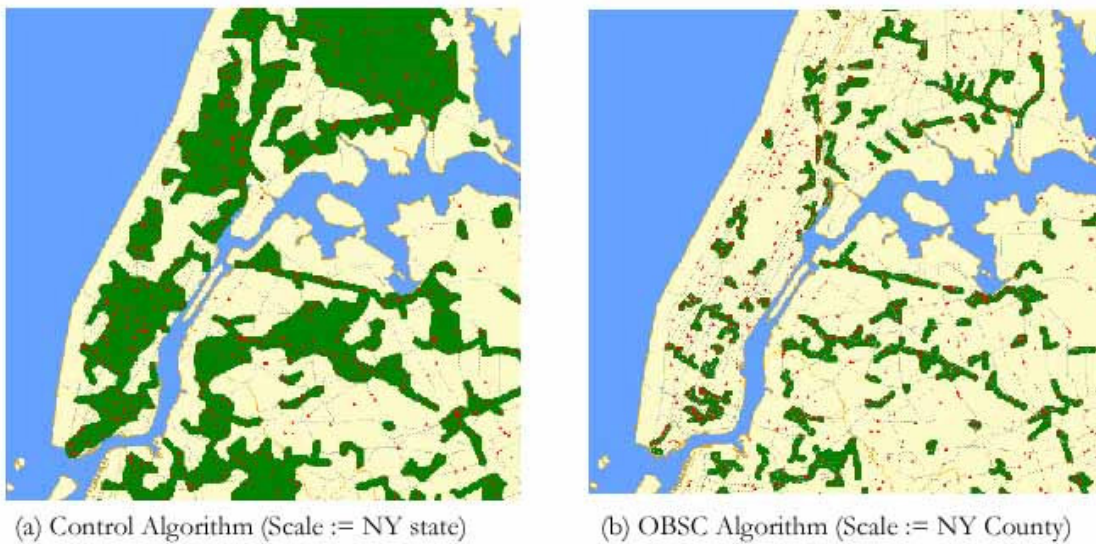


Figure 11. OBSC clusters reflect spatial distribution specific to the scale of users' interest

4.2.2 An Effect of Constraint (Domain Ontologies)

In figure 12 (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 OBSC 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, and result in natural clusters.

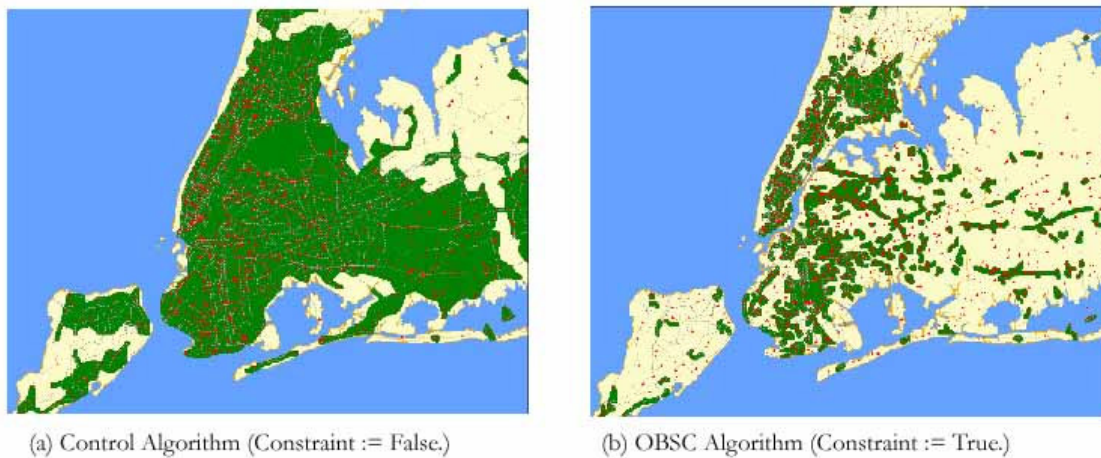


Figure 12. OBSC clusters identify the physical barrier due to concept implicit in domain

5. Conclusion

5.1 Summary

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

Findings can be summarized as follows: First, in ontology-based method clustering mechanisms are dictated by concepts implicit in domain. The resulting clusters of traffic accidents are concentrated along road network because a spatial constraint is priori implicit in domain. Second, OBSC 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. In a nutshell, ontology-based methods make clusters 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 clustering 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.

5.2 Further Works

Synthesis of spatial data mining and spatial statistics: There is a lack of cooperation between data mining and spatial statistics mainly due to different traditions and insights. I believe it may produce synergetic results when we can synthesize both two fields rather than deal with them separately.

Formalization of spatial/temporal knowledge: Even though spatial and temporal knowledge are ubiquitous, a lack of formalizing them prevents researchers from constructing and reusing ontologies. Consequently, it can facilitate spatio-temporal analysis including spatial clustering, thereby enhancing our understanding of spatio-temporal phenomenon.

Application ontology: It may be costly to build application ontology such as traffic accident analysis in a short term. However, it can significantly change the way of analyzing traffic accident data in a long term.

Human-computer interaction in GIS: I did not pay much attention to input/output components of the spatial clustering systems proposed here, but well-designed user interfaces and geographic visualization tools can make a huge difference as a facilitator of identifying patterns discovered in data mining tools.

Acknowledgements

This research was supported in part by the Center for Transportation Injury Research (CenTIR) in Buffalo, NY with funding from Federal Highway Administration (FHWA) and technical oversight by National Highway Traffic Safety Administration (NHTSA).

References

Abraham, T., 1999, Knowledge Discovery in Spatio-Temporal Databases, PhD Thesis, University of South Australia

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

Ester, M., Kriegel, H. P., Sander, J., and Xu, X., 1996, A Density-based Algorithm for Discovering Clusters in Large Spatial Databases, *KDD '96*

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

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

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

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

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

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

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

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.

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

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.

- Rainsford, C. P., 1999, Accommodating Temporal Semantics in Data Mining and Knowledge Discovery, PhD Thesis, University of South Australia
- Roddick, J.F. and Lees, B.G., 2001, Paradigms for Spatial and Spatio-Temporal Data Mining, in H. Miller, J. Han, (Eds.) *Geographic Data Mining and Knowledge Discovery*, Research Monographs in Geographic Information Systems, pp.33-49, Taylor and Francis
- Russell, S., and Norvig, P., 1995, *Artificial Intelligence: A Modern Approach*, Prentice Hall
- Smith, B. and Mark, D. M., 2001, Geographic Categories: an Ontological Investigation, *International Journal of Geographical Information Science*, 15(7): 591-612
- 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
- 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
- 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.
- Visser, U., Stuckenschmidt, H., Schuster, G., and Voge, T., 2002, Ontologies for Geographic Information Processing, *Computers & Geosciences* 28: 103-117
- Witten, I. H. and Frank, E., 2000, *Data Mining: practical machine learning tools and techniques with java implementations*, Morgan Kaufmann.