

Spatial Data Mining and Geo-spatial Interoperability

Shashi Shekhar

Computer Science Department, University of Minnesota
200 Union Street #4192, Minneapolis, MN 55455.

shekhar@cs.umn.edu

www.cs.umn.edu/~shekhar, www.cs.umn.edu/research/shashi-group

EXTENDED ABSTRACT:

Spatial data mining systems (SDMS) [1] extracts previously unknown, interesting, and useful, spatial patterns and relationships within spatial datasets. SDMS interoperability refers to the ability of diverse SDM systems to cooperate towards global goals by communicating and exchanging spatial data, patterns, relationships and other intermediate results.

SDMS interoperability is important for the following reasons. Many spatial datasets are extremely large and distributed over multiple sites. It is often not possible to bring the entire spatial dataset to one location for mining interesting global patterns due to concerns like privacy, security, communication cost, etc. SDMS interoperability may be helpful in overcoming these barriers. For example, one may run individual SDM system at each site to extract local patterns from the subset of spatial data stored at each site. The local SDMS may take advantage of interoperability to interchange local patterns and intermediate results to extract global patterns.

SDMS interoperability poses many difficult challenges at multiple levels. At the highest level, semantic interoperability level, it needs a specification of agreement about content descriptions of spatial data, patterns and relationships. Ideally, a common vocabulary of concepts and a common ontology are shared among all systems. Alternatively, well-defined translations are available to map concepts used by the sender to those used by the receiver. The middle level may focus on structural interoperability to provides means for specifying semantic schemas (or meta data) for sharing. At the lowest level, syntactic interoperability specifies common message formats (e.g. tags and marking) to interchange spatial data, patterns and relationships.

Recent development in spatial web standards (e.g. Web Mapping Service (WMS), Web Coverage Service (WCS), Web Feature Service (WFS), Web Terrain Service (WTS), Geographic Markup Language (GML), etc.) have addressed many issues in syntactic and structural interoperability related to spatial data. However, these standards have not directly addressed exchange of spatial patterns and relationships. One may encode spatial patterns as spatial data to establish basic communication among cooperating SDMS. However, it will still leave a fair amount of work towards creating common data mining concepts, ontologies etc.

Another option is to use a combination of spatial web standards and web based data mining standards (e.g. Predictive Model Markup Language (PMML)). PMML supports syntactic and structural interoperability for classical data mining by providing facilities like data dictionary, mining schema, transformation dictionary, model statistics and model parameters. However, they focus on classical data mining patterns, e.g. regression, which assumes that learning data samples are independent from each other. This assumption rarely holds on spatial datasets, which exhibit high auto-correlation. SDMS has developed new concepts (e.g. colocation [3] , spatial outliers [4], location prediction[2]) and models (e.g. spatial auto-regression, join-based colocation [3]) to address these limitations of classical data mining.

Thus, it is important to extend spatial web standards and web based data mining standards (e.g. PMML) to support novel SDMS concepts and models. At semantic level, this requires development of a consensus among SDMS researcher, users, and software developers towards a common set of concepts and ontologies. At structural level, it requires development of a consensus representation for exchanging schema. At syntactic level, it requires development of common message formats (e.g. XML tags).

Other key challenges include development of distributed algorithms for mining spatial patterns (e.g. colocations, spatial outliers, parameters of spatial auto-regression models) to compute global patterns from local patterns without copying local datasets to a common site. For example, our recent work [4] showed that almost all tests for spatial outliers belong to a special subclass of statistical functions, namely algebraic functions, which can be decomposed easily. This result provides a foundation for developing distributed algorithms for detecting spatial outliers by exchanging a few intermediate aggregate results without requiring exchange of the large datasets. We are now working on similar algorithms for mining other spatial patterns and relationships.

REFERENCES

1. S. Shekhar and S. Chawla, Spatial Databases: A Tour, Prentice Hall, 2003 (isbn 013-017480-7).
2. S. Shekhar, P. Schrater, W. Vatsavai, S. Chawla, W. Wu, Spatial Contextual Classification and Prediction Models for Mining Geospatial Data, IEEE Transactions on Multimedia, 2(4), June 2002.
3. S. Shekhar, Y. Huang, Discovering Spatial Colocation Patterns, Intl. Symposium on Spatial and Temporal Databases, 2001. (Extended version to appear in IEEE Transactions on Knowledge and Data Engineering).
4. S. Shekhar, C. T. Lu, P. Zhang, A Unified Approach to Spatial Outlier Detection, GeoInformatica: An Intl. Jr. on Adv. in Computer Science for GIS, 7(2), 2003. (A summary of results appeared in ACM SIGKDD Intl. Conf. on Data Mining and Knowledge Discovery, 2001).