

Data Processing in Space, Time and Semantics Dimensions

Farshad Hakimpour, Boanerges Aleman-Meza, Matthew Perry, Amit Sheth

LSDIS Lab, University of Georgia, Athens, GA 30602, USA
farshad@hakimpour.com, boanerg@cs.uga.edu, mperry@cs.uga.edu, amit@cs.uga.edu

Abstract. This work presents an experimental system for data processing in space, time and semantics dimensions using current Semantic Web technologies. The paper describes how we obtain geographic and event data from Internet sources and also how we integrate them into an RDF store. We briefly introduce a set of functionalities in space, time and semantics dimensions. These functionalities are implemented based on our existing technology for main-memory based RDF data processing developed in the LSDIS Lab. A number of these functionalities are exposed as REST Web services. We present two sample client side applications that are developed using a combination of our services with Google map service.

Keywords: Semantics, RDF, Spatiotemporal, spatiotemporal thematic (STT) functions and proximity, GIS, Event

1 Introduction

With popularity of spatial data on the Web and increasing adoption of Semantic Web technologies, the idea of Geospatial Semantic Web is introduced [4]. Adding temporal dimension alongside spatial and semantic dimensions [9,12] increases our analytical capabilities and requires addressing new data integration challenges. This paper describes our experience in integrating spatial information with event data (i.e., temporal and thematic data) and performing semantic, spatial and temporal analysis on the results. Using spatial and temporal data where available can increase accuracy and efficiency of processes such as disambiguation (as we show in section 2.3).

The technical contributions of this paper are in three areas:

- We represent spatial data using Semantic Web technology (RDF) and enhance this information with spatial relations. We experimented with a geographic dataset of the state of Georgia for which we generated RDF metadata representing major geographic features and their topological relations.
- We enrich the event data by relating them to associated spatial data. Specifically, we added geographic positions to event descriptions (by geo-coding the address of the venues). We also relate address information (street, zip code, state) to the spatial data described above.
- We introduce a set of processes on spatial, temporal and semantic dimension of events and show applications built using these processes. Using a set of semantic

analytic and event query processing tools, we show how the generated data can be used to build applications.

This paper is organized as follows. Section 2 gives an overview of our data acquisition and preparation including integration issues and disambiguation. In Section 3, we present a set of operations for querying space time and semantics. Section 4 presents our experimental systems using the data and operations introduced previously. We discuss the related work in Section 5, and Section 6 provides conclusions.

2 Data Preparation

We prepared two types of data: first, geographic data from Census Bureau and second, entertainment events from several sources on the Web. The resulting datasets are publicly available at <http://lsdis.cs.uga.edu/projects/semdis/spatiotemporal/RDFdata.html>.

2.1 Geographic Data

We prepared RDF metadata from four different data sets of counties, urban areas, roads and water bodies. The source of the datasets is publicly available geographic information provided by the U.S. Census Bureau for the states of Georgia and Florida. We enhanced the RDF dataset by adding the topological relations between entities. Figure 1 illustrates the model in which the data is represented. The main components of this model are as follows:

- *Geographic Feature Class* is the super-class of the main geographic entity classes. These entities are transformed to RDF with their corresponding attributes.
- *Geometry* class is foreseen in the model to keep position and shape of geographic features and complies with the OGC Simple Feature Specification [10] (Figure 2).

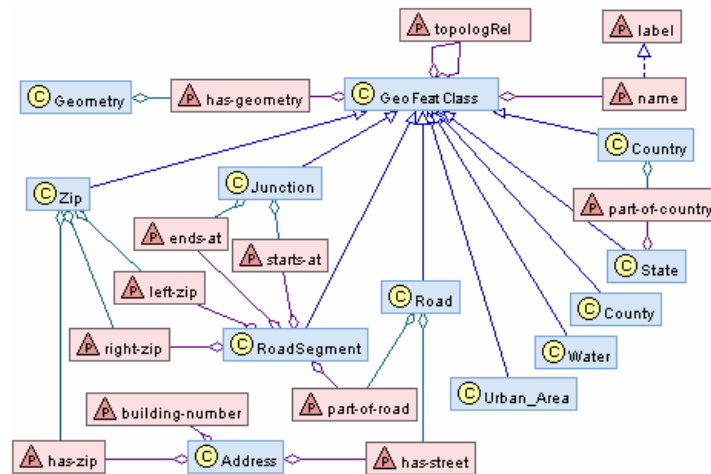


Figure 1. The RDF schema for geographic features.

However, we did not populate our RDF data sets by the geometry of the objects. In fact, one of our objectives in this work has been that of performing semantic analysis on the spatial objects while relying on the existing spatial processing engines (as presented in section 4.1).

- *Topological Relations* are added values obtained by the Oracle Spatial engine (e.g., relations between zip and state, county and state, road and county, etc.)
- *Address* is a placeholder that can be used in any other data set to relate other objects (e.g., venue in Figure 3) to spatial entities, such as zip, road and state.

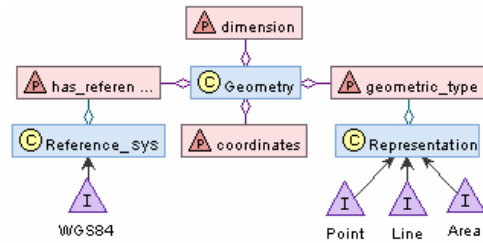


Figure 2. OGC Geometry model on RDF

2.2 Event Data

The event data presented here are extracted from three different Web sites: evenful.com, atlanta.creativeloafing.com and ticketmaster.com. For scraping we used NekoHtml Java library (<http://java-source.net/open-source/html-parsers/nekohtml>). Data items obtained and modeled for every event include (Figure 3):

- Event time: It could be a time point or a time interval. In most of the cases we have only the starting time.
- Event location: It is the venue where the event takes place.
- Geometry: Keeps the geographic position of the venue or the event (Figure 2). The geometry information is obtained from the Yahoo geo-coding service (<http://developer.yahoo.com/maps>). Events are also related to geometry class for special cases where an event occurs in a position without a venue, such as an accident.
- Address: This class relates venues to the spatial data (see event model in Figure 1).

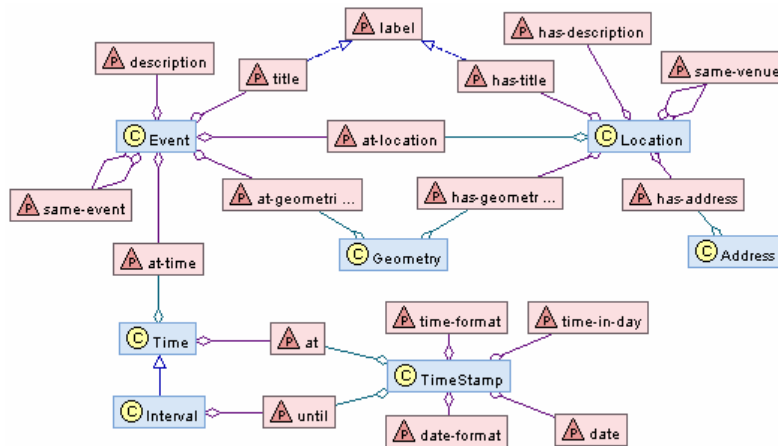


Figure 3. RDF schema for events and their time and venue.

2.3 Data Integration and Disambiguation

Schematic and semantic integration of the data sets obtained from several sources is the next step [13]. The schematic integration has not been a major challenge considering flexibilities provided by RDF. Semantic integration however presented significant challenges. Due to the use of several data sources for events and venues, obtaining different event (or venue) resources referencing the same real world entity is inevitable. This problem is known as the reference reconciliation or entity disambiguation problem [3, 14]. Furthermore, various forms of objects may be incompatibilities or conflict [8]. Such ambiguities are resolved during our integration process.

Existing disambiguation approaches typically rely on either text matching such as [7] or object attribute matching [3, 14]. Our approach extends traditional methods by incorporating spatial and temporal attributes. We used a combination of two stages of position matching and then title matching for resolving ambiguity of the identity of venues. For events, the disambiguation process is performed in three steps Time Matching, Venue Matching and finally Title Matching. Figure 4 illustrates an example disambiguation process for event E1 by matching it against other events.

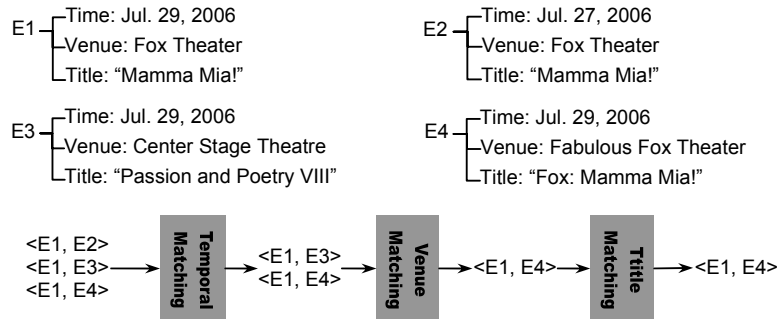


Figure 4. Illustration of an example for event disambiguation.

During the extraction process, we obtained events that we were not immediately able to classify due to lack of information. However, we are able to improve the event classification by the knowledge acquired from their venues. First, our system assigns usage tags to venues specifying the type of events taking place in a venue. Second, for every unclassified event, the system classifies the event based on the usage tags assigned to its venue. Finally, we created required relations between the address of venues and the geographic features such as roads and zip codes in our geographic dataset.

3 Spatial Temporal and Semantics Analysis

In this section we introduce a set of spatial, temporal and thematic (or semantic) operations we provide on our event dataset. These operations are used in our STT (spatial, temporal and thematic) disambiguation process, also used by the sample

application described in Section 4.2. The main focus of these operations is finding STT proximity in these three dimensions.

We measure proximity in space based on a distance function. Finding nearest neighbor for a position is a known operator in the spatial domain. We define this functionality by the following operation:

(1) $nearestEvent(type, pos, n)$

where $type$ is the type of event of interest, pos defines the position for the neighborhood function, and n defines the number of events in the result list. The result list is sorted by the distance from pos . An example of such proximity query is “finding the closest musical play near my office”:

$nearestEvent(<musical_play>, <33.946, -83.374>, 1)$

We extend the above proximity operation in time as measured through the following two functions:

(2) $nearestEventBefore(type, t, n)$
 $nearestEventAfter(type, t, n)$

where $type$ is the event type of interest, t specifies the time for the neighborhood measure, and n defines number of events in the sorted result list. The result of $nearestEventBefore$ is descending and that of $nearestEventAfter$ is ascending. An example of such a query is a request to “find 10 speeches right after the working hour on July 22”:

$nearestEventAfter(<class>, <July 22, 2006, 17:30>, 10)$

We use the association ranking developed at LSDIS and introduced in [1,2] as a measure for semantic proximity:

(3) $associatedEvent(type, resource, n)$

where $type$ is again the event type of interest and $resource$ determine an instance in the RDF graph. This function finds an event that is associated to the resource through a path in the RDF graph and returns the ones ranked highest. An example of such request is a query to find a performance involving a particular favorite artist or an event organized by a specific charity organization:

$associatedEvent(<comedy_play>, <Reed Martin>, 1)$

The proximity operators shown above operate on each of the dimensions. However, one may look for a nearest musical show in both temporal and spatial dimension. In such cases the nearest neighbor in temporal and spatial dimensions often are not necessarily the same events. For example, an event $e1$ is the nearest event in temporal vicinity (one hour) of our requested time and spatial vicinity of 20 miles while event $e2$ is the nearest event in spatial vicinity of our requested location (3 miles) but takes place four hours after our preferred time.

There is a need for a compromise or prioritization to identify a more suitable events in such cases. Using cost coefficient we define a spatiotemporal nearest neighborhood position as follows:

(4) $nearestEventBefore(type, t, pos, tCost, dCost)$
 $nearestEventAfter(type, t, pos, tCost, dCost)$

where *type* is the event type of interest, *t* and *pos* declare the point of interest in time and space dimensions, *tCost* is the cost of time difference per hour, and *dCost* is the cost of the distance per mile. The above function returns those events that minimize the following cost function:

$$(5) \quad cost(e) = (tCost * timeDiff(time(e), t)) + (dCost * dist(position(e), pos))$$

and returns a list of events sorted by the cost function. Finally, adding a parameter to the query in (6) for finding an event associated to an entity can satisfy major proximity queries:

$$(6) \quad nearestEvent(type, t, pos, res, tCost, dCost, rank)$$

An example of such a query would be finding a theater play starring a particular actor and taking place close to my office after working hour on 22nd July. However, if the venue is close to the office, I am willing to wait a day or two, rather than traveling a long way to the neighboring town and join the event right away:

$$nearestEvent(<theater_play>, <July 22, 2006, 17:30>, <33.946, -83.374>, <Reed Martin>, 6, 1, 0.2)$$

By setting *tcost* = 6 and *dcost* = 1, we express the fact that for the cost of traveling 1km we would wait 6 hours. Finally by setting *rank* to 0.2, in fact, we accept most of events that have any association with ‘Reed Martin.’ Alternatively, an application may wish to bias this cost function to favor time (e.g., it may be preferable to drive 20 miles than to go to an event that impinges on the dinner time so far as the event is on the preferred day).

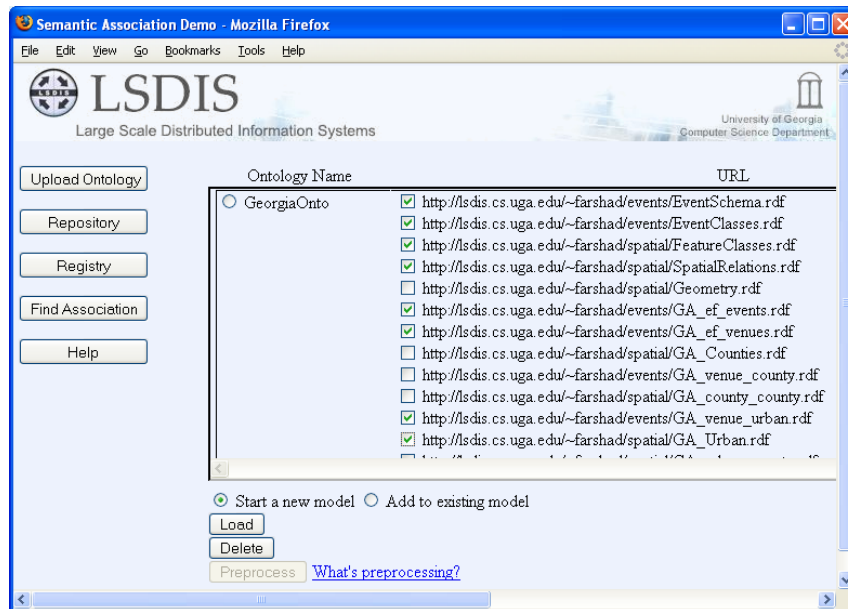


Figure 5. Loading RDF metadata sets to find semantic associations.

4 Sample Applications

This section introduces two applications that work with our datasets. One application is based on a generic semantic analytic tool that finds and ranks semantic associations in an RDF graph. With the addition of spatial knowledge to our dataset, this tool can associate events in spatial dimension. The second application uses the proximity functions introduced in the previous section to find suitable entertainment events. The analysis is performed using the integration of constraint in space, time and semantics.

4.1 Adding Spatial Information to Semantic Analysis

First, we show how spatial relations can enrich semantic associations. In short, a semantic association is a sequence of resources and properties in an RDF graph in a way that from each resource there is one property to the succeeding resource. There can be a very large number of semantic associations between two resources – often much larger than the number of documents that a search engine can find in response to keywords. This makes the issue of ranking semantic associations very important as well as challenging. Several approaches for finding and ranking these associations are discussed in [1, 2]. By means of adding spatial information to entities in the RDF ontologies, spatial objects and their topological relations take part in identifying and ranking the semantic associations.

Figure 5 shows how different RDF ontologies can be selected and loaded into the system for finding semantic associations. The ontologies are organized in modules to avoid loading unnecessary data into the memory. For example, if urban areas are of our interest we do not load the spatial information about counties.

In the next step we run one of our semantic association ranking algorithms and also add an ability to visualize these associations. A query to find associations between

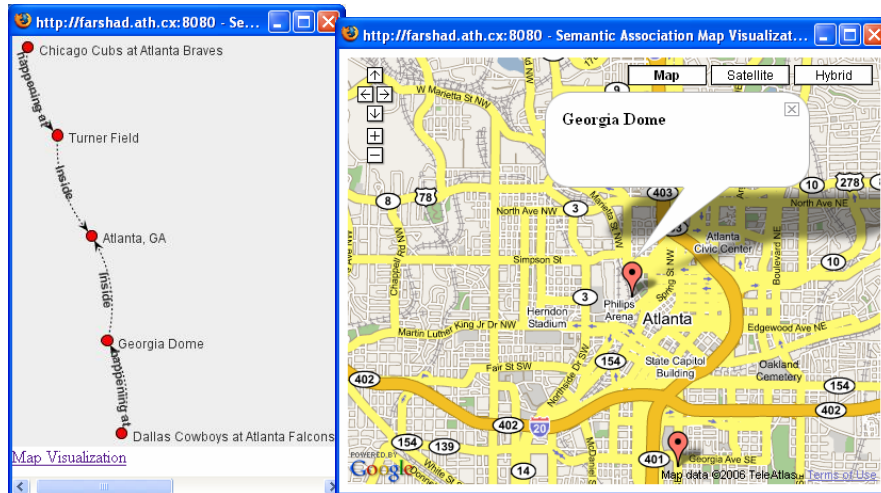


Figure 6. A semantic association involving spatial relations on the left. Geographic entities in the association are illustrated on the right.

“Dallas Cowboys” and “Chicago Cubs” results in a number of associations. An association that contains spatial relations is illustrated in Figure 6 (left). The association shows that both teams have matches scheduled at venues in Atlanta. As two of the resources in the association are venues and related to geographic positions, we are able to illustrate them on a map. The visualization of the venues in our example path (Georgia Dome and Turner Field), using Google map API is shown in Figure 6 (right).

4.2 Semantics as a Dimension alongside Space and Time

In this section, we show how an application using the functionalities introduced in Section 3 is able to find suitable events. As the first step, a set of REST Web services based on the functionalities in section 3 are exposed to the Web. These services are available at <http://lstdis.cs.uga.edu:8080/SemDisServices>. We provide a client side application that allows a user to specify a set of request parameters. These parameters are used to invoke our REST services as follows:

1. Time: date and time of day (default: current browser time)
2. Space: location by specifying an address or by clicking on the map. In the case of entering an address, the client geo-codes the address using Google geo-coding service on the client-side and then sends the position.
3. Semantics: semantics of events can be constrained in two ways. First, by specifying an event type, the user can narrow down the type of events. Second, by

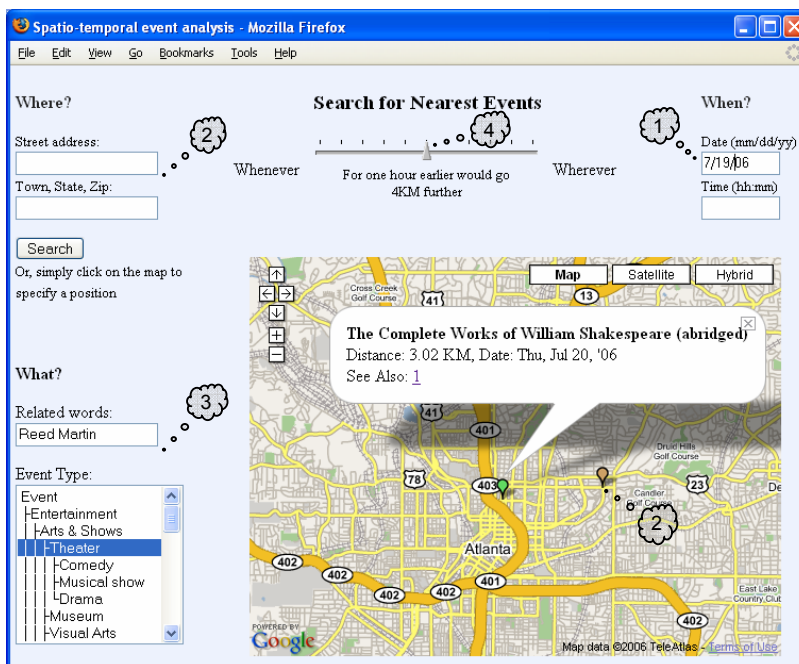


Figure 7. A sample application based on our proximity functionality.

providing keywords that we relate to the resources in our RDF graph and then associate with the events in our dataset.

4. Costs: cost ratio of time and space. The client provides a slider that helps the user to specify the importance of the temporal constraint as related to the spatial constraint. The cost ratio is translated to a verbal sentence describing the preference expressed by the ratio. For example, how much one would be willing to travel to join an event that takes place an hour earlier; or, how long one would wait to travel one kilometer less.

Finally, the result of the service invocation is displayed on the map. A snapshot of the client side user interface is presented in Figure 7.

5 Related Work

Our work is related to literature in different domains, namely, data acquisition, spatial data modeling in RDF, disambiguation, and finally event modeling and processing. We used tailored Java code (using NekoHtml library) for web scraping, because of the flexibility in generating output RDF datasets and in scheduling of extractors. However, as Semantic Web technologies are gaining popularity, more extraction tools [5] and specifications [6] are becoming available with enhanced capabilities. We believe that in mid-term future, there will be more RDF metadata available as well as better alternative tools for RDF data extraction.

On modeling of spatial information, activities of the RDF community are limited to modeling latitude and longitude of points (see www.w3.org/2003/01/geo/). We used a more expressive model by adopting Open GIS Consortium specification in [10]. Another alternative in this area would be adopting GML [11]. GML is a more complex specification, and we believe such level of complexity is not needed for lightweight spatial processing needed by the Semantic applications of the types discussed in this paper. However, enterprise-centric and scientific semantic applications may benefit from more complex specifications.

Work on disambiguation can be divided into two categories: disambiguation of objects in text as in [7] and disambiguation of objects from different datasets as in [3] and [14]. Our work is similar to [3] and [14] in the sense that they are also concerned with object disambiguation based on object attributes. However, we take advantage of temporal and spatial attributes of venues and events.

Part of this work is about event modeling and processing. There is a good body of work on spatiotemporal data processing; however, this paper is aiming at modeling and processing in semantics, space and time. A similar work in this domain that pays reasonable attention to the STT dimensions is presented in [15]. It presents an event-based system for a different domain of application, multimedia information management, and a vision of emerging event-based applications.

6 Conclusion

The focus of this paper is presenting our experience in integrating semantics, space and time. As information related to events is increasing in these three dimensions, we explored the integration of such information from different sources. The paper also

presents query operators that allow integrating constraint on proximity in these dimensions.

The paper presents a description of steps for data preparation and integration. We introduce a subset of proximity operators developed at LSDIS for querying event data. Finally, we discuss two systems working with semantic, spatial and temporal data.

Acknowledgments. This work is partially funded by NSF-ITR Award#0325464 titled ‘SemDIS: Discovering Complex Relationships in the Semantic Web.’

References

1. B. Aleman-Meza, C. Halaschek-Wiener, I. Budak Arpinar, C. Ramakrishnan, and A. Sheth, Ranking Complex Relationships on the Semantic Web, *IEEE Internet Computing*, Vol. 9 No. 3, pp. 37-44, May/June, 2005.
2. K. Anyanwu and A. Sheth, “The r-Operator: Discovering and Ranking Associations on the Semantic Web”, in *Proc. of the 12th Int’l World Wide Web Conf.*, 2003, pp. 690-699.
3. X. Dong, A. Halevy and J. Madhavan, “Reference Reconciliation in Complex Information Spaces”, in *Proc. of the 24th Int’l Conf. on Management of Data, ACM SIGMOD*, pp. 85–96, 2005.
4. M. J. Egenhofer, “Toward the Semantic Geospatial Web”, *Proceedings of ACM-GIS 2002*, McLean, VI A. Voisard and S.-C. Chen (eds.), pp. 1-4, Nov. 2002.
5. B. Hammond, A. Sheth, and K. Kochut. “Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content”, in *Real World Semantic Web Applications*, V. Kashyap and L. Shklar, Eds., IOS Press, pp. 29-49, 2002.
6. D. Hazaël-Massieux, D. Connolly, “Gleaning Resource Descriptions from Dialects of Languages (GRDDL)”, W3C draft, www.w3.org/2004/01/rdxh/spec, 2006.
7. X. Li, P. Morie and Dan Roth, “Semantic Integration in Text: From Ambiguous Names to Identifiable Entities”, in *AI Magazine: Special Issue on Semantic Integration*, pp. 45-68, 2005.
8. V. Kashyap and A. Sheth, “Schematic and Semantic Similarities between Database Objects: A Context-based Approach,” *Very Large Data Bases (VLDB) Journal*, 5(4), Oct. 1996, pp. 276-304.
9. J.L. Mennis, D. J. Peuquet and L. Qian, “A conceptual framework for incorporating cognitive principles into geographical database representation”, *International Journal of Geographical Information Science*, vol. 14, No. 6, Sep. 2000, pp. 501 – 520.
10. Open GIS Consortium, Inc., “OpenGIS Simple Features Specification for SQL”, http://portal.opengeospatial.org/files/?artifact_id=829, May 1999.
11. Open GIS Consortium, Inc., “GML simple features profile”, http://portal.opengeospatial.org/files/?artifact_id=11266, 2005.
12. M. Perry, F. Hakimpour, A. Sheth, “Analyzing Theme, Space, and Time: An Ontology-based Approach”, to appear in the *proc. of the 14th Int’l Symp. on Advances in Geographic Information Systems ACM-GIS’06*, Nov. 2006.
13. A. Sheth, “Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics”, in *Interoperating Geographic Information Systems*. M. F. Goodchild, et al. (eds.), Kluwer, Academic Publishers, 1999, pp. 5-30.
14. S. Tejada, C. Knoblock, and S. Minton, “Learning Object Identification Rules for Information Integration”, *Information Systems* Vol. 26, No. 8, pp. 607-633, 2001.
15. U. Westermann and R. Jain, “Events in Multimedia Electronic Chronicles (E-Chronicles)”, in *International Journal on Semantic Web and Information Systems*, Vol. 2, No. 2, pp. 1-23, Apr. 2006.