Spatial and Temporal Target Association through Semantic Analysis and GPS Data Mining

David Luper, Delroy Cameron, John A. Miller, Hamid R. Arabnia
Department of Computer Science
University of Georgia
Athens, GA, USA

Abstract *- We present an application for analyzing temporal and spatial interaction in an Association Network environment based on integrating Global Positioning Systems (GPS) Data, RDF Metadata and Data Mining. We argue that GPS data contains important information about relationships between people, through location and time, and can ultimately provide ideas about their association level and activities. We propose RDF Metadata and Data Mining can discover these Semantic Associations between entities, from which further insight into past, present and future associations and activities can be determined. We visualize our findings to illustrate effectiveness and also to persuade the user of the feasibility of our system called "GPODS" (**G**lobal **P**ositioning **O**ntological **D**ata **S**ystem). Ultimately, we argue that GPODS adopts a reasonable methodology which Association Network Analysis can use at this juncture of the GPS data "explosion" to bring association context to GPS data.*

Keywords: Global Positioning System, Association Networks, Data Mining, RDF Metadata, Discrete Mapping, GPS Binning

1. Introduction

The interaction of human beings often alludes to explicit as well as implicit social relationships among them based on physical proximity or common location. Association Network analysis can provide insight into their activities by observing and understanding these relationships. However, the ultimate benefit of analyzing human interaction lies in the ability to predict events and forecast Association Network evolution. Association Networks can be defined as ways people interact or different ways people and places can be associated with each other. For instance being in the same place at the same time, frequenting the same locations, interacting with the same group of people, being at the same types of locations (i.e. different branches of the same bank), etc. are all ways people could be in the same association network. In the past, various research techniques have been used to unravel Association Networks. For example, work in [7] presents semantic web techniques that use ontologies for discovering semantic associations among entities using them to speculate on interesting occurrences within the network, in [5] semantic techniques are used in web mining, and [8] presents analysis of a full blown geospatial ontology for finding association in geospatial data.

In this paper we propose that Global Positioning Systems Data can be used in conjunction with RDF Metadata and Data Mining to analyze Association Networks. It is our belief that spatial locations and physical proximity are reliable indicators of relationships and common activities among targets in an Association Network. We believe that the integration of GPS data and RDF Metadata can be a meaningful way of discovering, expressing and understanding such associations. Our motivation benefits from the realization that the upsurge of recent terrorist activity makes such prediction especially important in National Security. Moreover, other areas such as sex offender monitoring and both online and research community evolution could potentially be interesting areas to apply this methodology.

In our approach we extend our analysis across several domains. As such, we make some basic yet important assumptions. First, we assume that technologies exist for GPS data collection and monitoring (this will be discussed briefly in later sections). For this research we implement a simulation of a GPS data collection framework which also will be discussed later. At this point in our work, we are not taking into account privacy constraints that could potentially restrict GPS data gathering, but we focus on demonstrating the benefits and practicality of our methodology. With the recent upswing in applications utilizing the GPS system, i.e., automobiles, mobile phones and other communications devices, the usefulness of GPS data has become apparent and its usefulness will only continue to grow. This will without

a doubt shadow privacy concerns and a legal framework will have to be developed for handling this increasingly useful data.

2. GPS Data Acquisition

The term GPS, as noted above stands for Global Positioning System. One could think of the Global Positioning System as giving every centimeter on the earth a unique address (the accuracy of implementations of GPS receivers varies in range [9] from typical commercial accuracy of around 15 meters to high end technology that can be accurate to less than 5 centimeters). As discussed in [9] it consists of 24 satellites in orbit around Earth broadcasting their positions and time to various receiving stations. To calculate its position on the earth a receiver analyzes the spatial locality of GPS satellites within its range of visibility and calculates the latency of their broadcast transmission. If at least three satellites are visible a receiver can triangulate its coordinates (both latitude and longitude), and if enough satellites are visible, at least 4, the receiver's altitude can be obtained as well.

2.1 GPS Format

There are many different formats for displaying GPS coordinates. We adopt the floating point number system, otherwise referred to as the Decimal Format System, which ranges from -90 to 90 latitude and -180 to 180 longitude.

Figure 1 – Decimal System for GPS Coordinates

2.2 Data Collection

Obtaining a constant stream of GPS data requires a transmitter and a receiver. A transmitter is a locating device placed on a target communicating signals to a receiver. Receivers are often part of a database system, developed primarily for data logging. Research in [10] discusses architecture for data acquisition using common communications devices such as mobile phones fitted with integrated hardware tracking devices. The mobile phones act as GPS receivers that calculate coordinates and also transmit those coordinates to a logging system. Opponents to the embedded devices in personal gadgets paradigm will argue that such an approach constitutes invasion of privacy. Our counter argument is that confidentiality agreements subject only to government issued warrants for data release can allay privacy violation concerns, especially given that various telecommunications giants currently engage in the practice of record logging on subscriber activities. Some contend that GPS data harvesting is no more invasive.

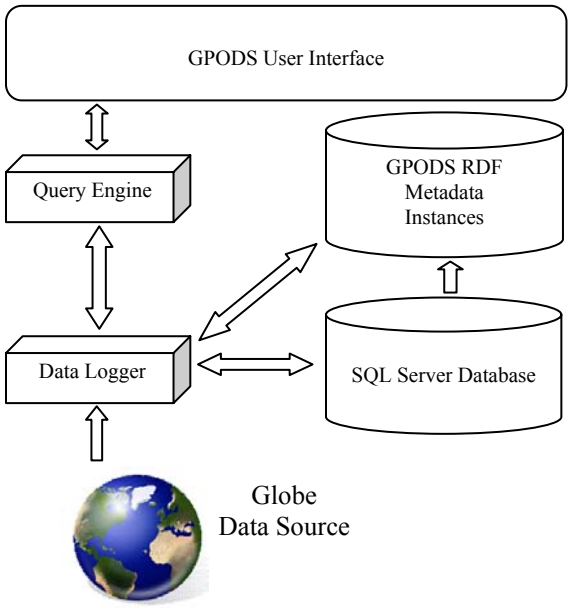

Figure 2 - GPODS System Architecture

3. System Architecture

Figure 2 shows the GPODS System Architecture. At the core of our system is a database implemented with SQL Server and RDF Metadata queried at runtime. Our Globe Data Source is in fact an intelligent, random generator that feeds GPS data into a simulated Data Logger which is a virtual wrapper between the Global Data Source and the database. The design of this simulated data source was given much consideration. It followed one hundred and thirty three targets throughout a time period of 60 days logging points for each of the targets at ten second intervals. Thirty three of the targets were considered terrorist targets while one hundred of the targets were simply good people with no affiliation to terrorism. Each of the targets were generated a home around the city of Athens, GA (USA) and every target started the 60 day time period from their home. Each day was simulated as a real day with day time and night time and all of the movement for each person in our simulator was generated with random

probability. At the begging of each day a number of meetings for the day and their corresponding times were randomly generated. Each day there was between 25 and 75 meetings of between 2 and 20 people. Through each day the targets moved around and when a meeting time came up somewhere between 2 and 20 targets were chosen randomly to meet at some location (destinations could either be a special predetermined region or not). All movement for the simulation, for every target, was kept inside the city of Athens, GA. As a final note it is important to mention that the one hundred good people and the thirty three terrorist were generated in separate runs of the system. This means that they shared the same special predetermined regions (again these are potential randomly selected destinations for a target) but the random meetings are different for the good people and the terrorists.

Moving on with the system architecture, the Query Engine directs all queries to the Data Logger, which may trigger the ontology (always derived from the database) or the database directly, depending on the nature of the query. Ultimately, query results are returned to the user through the Data Logger, then the Query Engine.

3.1 Data Logger

Commercial Data Loggers exist for small scale use. However, in the context of the ideas we present in this research, a GPS Data Logger could potentially contain large amounts of data and would need to be structured accordingly. We express the relationship between the quantity of data, the number of targets in a dataset, and the time resolution of data transmission in the following way. The amount of data collected in a GPS Database is directly proportional to the frequency of data transmission multiplied by the number of transmitters within the global search space. A measurement of the size of the database per month of data can be calculated as:

$$Size = packet\ size * packets\ per\ day * number\ monitored * 30 \quad (1)$$

The actual database size per month for 100 people in the GPODS system is approximately 700 megabytes.

3.2 Path Points Schema

The Path Point schema used in the GPODS dataset holds paths of monitored targets with TargetID and DateTimeStamp as the primary Keys. In this implementation of GPODS there is two months worth of data for each target. To store path points for each target, we exploit the assumption that "a person can only be in one place at one time". Connected data points for a target yield paths reflecting target movement over certain regions. It quickly becomes obvious that some means of data management for effective querying is necessary, since data volume can become excessively large. Indexing is used to solve this problem. Due to the fact that indexing has time/space trade-offs we suggest using only necessary indexes. A list of relevant indexes is as follows.

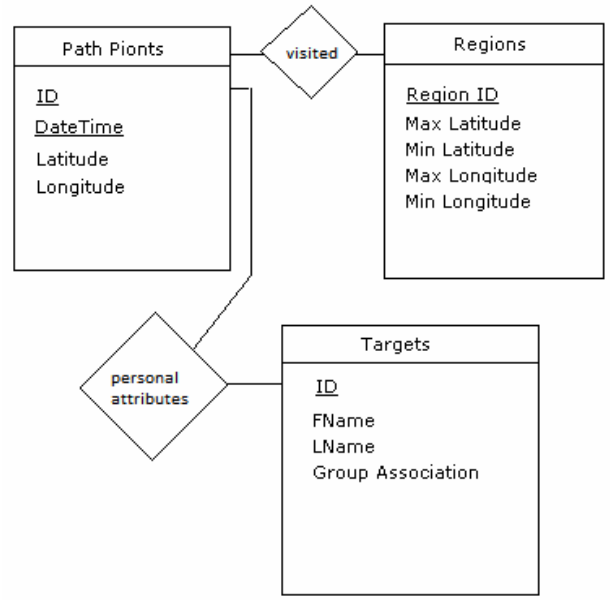

Figure 3 - GPODS ER Diagram

- [TargetID], [DateTimeStamp]
  The primary key allowing for retrieval of a particular person's coordinates at a specified slice of time.
- [Latitude], [Longitude]
  Allows for querying the entire database concerning who has ever been at a specific location.
- [DateTimeStamp]
  Allows for fast retrieval of where everyone in the database was at a specific time.
- [DateTimeStamp], [Latitude], [Longitude]
  Allows for retrieval of who was at a specified location at a specified time.
- [ID], [Latitude], [Longitude]
  Allows for querying all the times a person was at a specific location.

3.3 Targets Schema

This table holds information about targets in the database. The most useful piece of information in this

table is the Group Associations field. This field indicates organizations or groups to which targets are known or suspected to be associated with. It is this information that we rely on for discovering "hidden relationships" among targets using RDF metadata.

3.4 Regions Schema

The Regions table allows for storage of special monitored or "critical regions." These regions are well defined and have special interest in our database. These could be anything from government building to banks, or whatever would be necessary. The idea is that these regions being well defined, allows them to be explicitly monitored as an important region in GPODS.

4. GPODS Data Mining

The areas of data mining that are of most use to GPODS are Binning and Association Rules. The process of analyzing an Association Network consists of a two phase process. First Data Mining techniques are used to find interesting targets and places associated with those targets. Once interesting targets and places are found the RDF Metadata is used to find semantic associations for these interesting targets and places.

4.1 Binning

For discovering implicit associations (associations based on how targets actually interact) in GPODS we implement a Discreet Map data structure that effectively bins the path points of the different targets. Implicit associations are discovered through this Discreet Map and then potentially submitted to the RDF Metadata for further analysis and discovery of relational paths. The Discreet Map is a three dimensional array of Map Entries. Map Entries consist of a target ID, GPS coordinates, and the arrival and departure time for the target to and from a specific cell in the Discreet Map. The Discreet Map represents a two dimensional area from a real map only it maps the real latitude and longitude values to discreet space. The third dimension, z index, of the Discreet Map consists of a list of entries at each x and y location representing people that have visited a specific cell of the Discreet Map during the monitored time period. Once the Discreet Map is initialized, paths from desired targets can be placed into it by converting each point in the respective target's path from latitude and longitude values to x and y coordinates in the discreet space. Each map cell can hold many entries and is a lookup table keyed on Target ID. If all of the GPODS targets are placed into the map for a specified time region then it is possible to mine every meeting between two or more targets. As the entries from the targets' paths are inserted into the Discreet Map a global lookup table of traversed indexes is maintained. This allows for faster analysis of the Discreet Map because the visited map cells are all indexed. Once the desired paths are inserted into the Discreet Map, the Map cells can be indexed into another lookup table that proves very useful. This lookup table is a table of lookup tables of lists of entries and contains all of the traversed Discreet Map cells. The main (or outer) lookup table is indexed on the number of targets that visited each cell. If two targets visited a certain cell then that cell gets put into the main lookup table under the index two. This allows us to see all locations that more than one person visited. These points are of interest because if more than one person visited them they could be of greater importance or potentially be a meeting location. The inner lookup table consists of lists of integers representing Target IDs and is indexed on x and y locations. With this structure it is possible to obtain a list of all locations frequented by more than one person and then iterate through those locations finding out who visited each of them. It becomes possible to obtain the time each target visited a specific x and y location and thus provides a way to discover meetings as well. It is also possible to build this Discreet Map into a pyramid which effectively expands the latitude and longitude values covered by each individual x and y cell in the Discreet Map. This structure helps build the RDF metadata that will be talked about later which associates people to places for hierarchical association tracing.

4.2 Target Groups

For GPODS to be meaningful we identify specific types of groups and relationships being sought. The groups table consists of three different terror cell groups, each with a local leader, and one global terrorist leader. The complexity of terrorist networks makes current data mining techniques alone insufficient for truly understanding and disrupting their plans. Placing contextual meaning, including examining defined, prior known relationships, will provide more substantial means of detecting suspicious human interaction because it will allow us to analyze an association network related to specific interesting data rather than just the interesting data alone. This idea of association context allows us to see interaction between two people as something of concern. For this level of granularity we use RDF Metadata and Semantic Web techniques.

4.3 GPODS RDF Metadata

According to [3] an Ontology is "a formal explicit description of concepts in a domain of discourse." From

[4] an Ontology is defined as a set of logical axioms designed to account for the intended meaning of a vocabulary. We draw from these two definitions, and generate RDF Metadata as a way to denote objects and manage their relationships in a graph like structure.

The GPODS RDF Metadata thus contains three classes and various attributes. Targets and Regions are replicated from the database schema and a third class is created for maintaining Target Groups. The basic idea is that people can be associated with other people through prior known associations, places can be associated with places through prior know associations, and finally people can be associated with other people and or places through their actual physical interaction. Figure 4, shows this RDF Metadata schema. The GPODS RDF Metadata provides only one function not provided by our database, which is giving relationships between pieces of data. It is through this RDF Metadata that hierarchal levels of association can effectively be gathered about associated targets and places. It is for this reason that the RDF Metadata is critical to the success of GPODS, without it data can be found through mining but this mining is not effective at producing contextual relationships. Through contextual relationships the whole picture of Association Networks can be unraveled and a person can view this picture, allowing that person to make a decision as to whether it is potentially alarming.

5. Query Processing

There are several queries that can yield interesting results from GPODS. Depending on the nature of the situation various approaches can be taken based on the GPS data alone. However, we focus our attention in this section to queries based on spatial and temporal data events within the system. It should be mentioned that the example queries are structured in such a way that they are reacting to some event after the event happened. It is easier to explain the GPODS process using these examples; however GPODS encompasses other ideas in addition to reacting to events such as discovery of occurrences leading up to events, in the hopes that they can be intercepted before they happen.

5.1 Temporal Locality Query

Consider the scenario in which a terrorist bombing event (E_A) occurs on a building for which there is only one suspect (P_A). Due to the extent of the blast, authorities speculate that the suspect did not act independently. Much to their chagrin, however, the suspect refuses to volunteer any information. We outline the GPODS approach to this plight:

- *Use the GPS database, to obtain information regarding all targets that interacted with suspect through locations within a certain time frame*
- *Find all semantic associations between the suspect and each target using the ontology*
  a. *Query all temporal associations from the database involving our target (people who met with our target).*
  b. *Query all spatial associations from the database involving our target (people who visited the same places as our target) .*
- *Cluster high ranking persons for questioning*

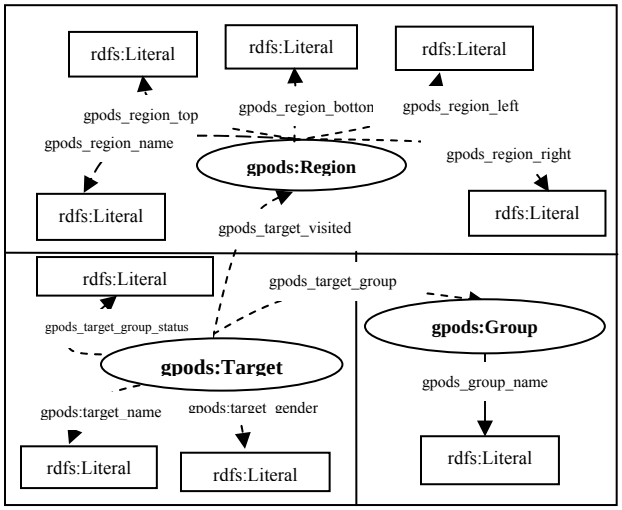

Figure 4 - GPODS RDF Metadata Schema

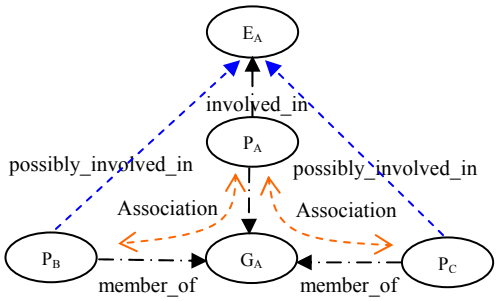

Figure 5 - Semantic Association

It is obvious that targets belonging to similar groups (G_A), with whom there was recent contact, have a higher probability of involvement in this crime, as shown in figure 5. It can also be deduced that any semantic relation with heavy overlap, i.e., another target who visited all the same locations in the specified time frame, should be flagged as an interesting target in need of further analyzing. In the past, using ontologies to find Semantic Associations alone, would indeed uncover

implicit relationships pointing to various other targets. However, GPS data refines this mechanism by identifying clusters of targets, with a higher probability of involvement by supplementing Semantic Associations with timestamps and physical proximity. The real power of the system can be shown here. This example shows the power of discovering an Association Network, not only do we have the potential group of involved persons, we have specific locations and networking paths from which to ascertain how an investigation could potentially be furthered. It is the three dimensional relationship between GPS, RDF Metadata and Binning, that demonstrates the ultimate contribution of the GPODS triad architecture.

5.2 Spatial Locality Query

Consider a second scenario involving the assassination of a public official. Pure GPS data, although useful, would be insufficient for addressing this problem, for two reasons. First, since several targets with the same timestamp could appear in close proximity to the event, temporal data alone could be insufficient. Second, it is unlikely that useful relationships exist between the victim and the target because the event attracts many unrelated targets. With spatial locality we note that it is worthwhile to retroactively obtain a list of places visited by various targets and use the ontology to obtain clusters of targets based on these locations to be analyzed separately. For example isolating targets who visited previous locations as the victim, but have criminal records would be of value. Alternatively, studying targets that canvassed the area prior to the event would also be of potential value.

6. Evaluation

We ran several queries against our simulated system in order to find general temporal and spatial association levels between targets and other people. The queries involved an assumed event in our simulated data involving one of the terrorists. We took the time between the event and 5:00 a.m. the day before and queried the database getting all one hundred and thirty three of our people (thirty three of which are terrorists) and put them into a discreet map. We did not tell the system who was a terrorist and who was not because we wanted to see who our system would associate with our target based only on the interaction of the people in our simulated world. From the discreet map we outputted any place that two or more people visited (the places were approximately 50 feet by 50 feet in resolution) to the RDF Metadata along with a list of meetings that took place and the meeting participants. We built the RDF Metadata and queried it for paths to our target of length two edges. These queries produced many paths that were ranked with a frequency count and their semantic association was scored. In order to further understand temporal and spatial association we used association rule mining to produce a probability representing the likelihood of someone being at a meeting if our target was at a meeting and a probability representing the likelihood of someone visiting a place if our target visited that place. We took each person's association probability and multiplied it by the semantic score to produce a final temporal and a final spatial association, respectively. The more interaction spatially and temporally between our target and another person the higher that person scored. Our results, sorted by temporal association level, were structured like so:

ID	Name	Temporal Score	Spatial Score
31	Jamy Puetty	Temporal=36.13	Spatial=3628.34
79	Inirey Locez	Temporal=35.43	Spatial=7518.07
11	Joe Quimnley	Temporal=21	Spatial=2585.02
64	Tyler Herpndon	Temporal=17.71	Spatial=122.99
58	Lamar Kiurte	Temporal=2.5	Spatial=26513.56
3	Amber Whamptley	Temporal=2	Spatial=4978.6
57	David Gringter	Temporal=2	Spatial=76.9
29	Erin Herlndon	Temporal=2	Spatial=147.3
25	Taylor Tannpser	Temporal=1.5	Spatial=1633.48
33	Taylor Crowgrey	Temporal=1.5	Spatial=48.1
5	Rachael Lomnez	Temporal=1	Spatial=198.86
38	Jessica Mannvly	Temporal=1	Spatial=170.5
42	Hannah Herpgenrader	Temporal=1	Spatial=52.3
34	Michelle Grintger	Temporal=1	Spatial=175.88
37	Chris Jorhddan	Temporal=1	Spatial=6048.56
26	Delroy Bondstys	Temporal=1	Spatial=3655.81
48	Chad Costanzkla	Temporal=0.5	Spatial=217.76
35	John Hernjdon	Temporal=0.5	Spatial=3658.65
18	T.J. Frenkder	Temporal=0.5	Spatial=47.6
46	Michelle Granskaata	Temporal=0.5	Spatial=4333.55
45	Chris Trekeear	Temporal=0.5	Spatial=2911.33
131	Random Person 131	Temporal=0	Spatial=21206.64

Looking at the temporal association scores it can be deduced that the four people who were associated temporally the most to our target were Jamy Puetty, Inirey Locez, Joe Quimnley, and Tyler Herpndon (names were purposely obfuscated to reduce probability of a real person's name being used). The temporal scores effectively separated the terrorists from the "good people" which is what was expected, however there was a couple of queries where a good person was scored high temporally because they randomly walked into a terrorist meeting. In the future the duration that someone stays at a meeting in combination with whether a person has any prior association with anyone in the meeting could be taken into account to more effectively manage this. The spatial scores made sense in the fact that associations were found because people had spatial overlap. The simulated data presumably had too much overlap and the in further research the terrorists and the "good people" should select from different special areas with less probability of overlap in order to more effectively test spatial scoring. As it stands with this data there is too much overlap between the terrorists and

the "good people" to consider it a realistic model. That being said it did effectively produce a valid score, the overlap occurred and the system detected it; the way the data was simulated was the problem. We ran multiple queries and the temporal results were consistent in scoring terrorist that were in the targets association network higher than the "good people". Furthermore, it clearly defined who among the terrorist group was more associated with our target over the time range of the query, and their relative degree of association based on the value of their score. In future work these general association scores could be used to pose more specific queries on the subset of people scoring high association levels

7. Future Research

Current topics being explored for further research include the following. Different methodologies to extract patterns of behavior borrowing from the computational intelligence field are one area of focus. Time sequence neural networks are of particular interest. Computational intelligence or the AI field in general yields some interesting properties that could help mining through the large amount of data produce better results. Another potential area for this research to head deals with approximate association rule mining mentioned in [6]. The idea behind this deals with association rule mining between participants in an Association Network to determine association strength between the targets. A closer proximity could potentially account for a stronger association. This idea could potentially be benefited by the use of fuzzy logic as well. Finally another area that is being considered is path prediction. Humans by nature are creatures of habit and in relation to this it would be possible to generate potential future destinations for targets based on prior training data. This is particularly interesting as it could help detect potential events in the future that could be investigated before they occurred and could also have substantial impact on commercial applications geared towards consumers.

8. Conclusion

In conclusion this paper presented a system for analyzing an Association Network using GPS data, RDF Metadata and Data Mining. Although there may be existing techniques for Association Network Analysis isolated in GPS databases, RDF Metadata and Data Mining separately, we adopt an interesting approach by pointing to the benefits of a cohesive amalgamation of these three areas. Our general idea of finding interesting paths based on spatial and temporal data, proves to be a reasonable approach. We contextualize the GPS data by giving meaning to various relationships by finding Semantic Associations from the GPODS RDF Metadata. In the process, we identify clusters of targets and/or locations of greater interest and relevance to our problem. We see path prediction as the ultimate functionality of an Association Network system, but do not yet focus effort on designing such a system.

References

[1] Anyanwu Kemafor, Sheth Amit P.: The ρOperator: Discovering and Ranking Associations on the Semantic Web, ACM SIGMOD Record, v.31 n.4, 1–6, Dec. 2002

[2] Ester Martin, Kriegel Hans-Peter, Sander Jörg.: Algorithms and Applications for Spatial Data Mining, Geographic Data Mining and Knowledge Discovery, 160–185, 2001

[3] Gruber Thomas R.: A Translation Approach to Portable Ontology Specifications, Journal of Knowledge Acquisition, 1 - 27, 1993

[4] Guarino Nicola.: Formal Ontology and Information Systems, International Conference on Formal Ontology in Information Systems, Trento, Italy, 1-10, June 1998.

[5] Lim Ee-Peng, Sun Aixin.: Web Mining - The Ontology Approach, International Advanced Digital Library Conference, Nagoya Japan, 1-8, 2005

[6] Nayak Jyothsna R., Cook Diane J.: Approximate Association Rule Mining, The Florida Artificial Intelligence Research Society Conference, Key West, 1-6, Florida, 2001

[7] Noy, Natalya and McGuinness, Deborah L.: Ontology Development 101: A Guide to Creating Your First Ontology Stanford KSL Technical Report KSL-01-05, 1-25, 2001

[8] Budak Arpinar, Amit Sheth, Cartic Ramakrishnan, E Lynn Usery, Molly Azami, Mei-Po Kwan, Geospatial Ontology Development and Semantic Analytics, Transactions in GIS 10 (4), 551–575, 2006

[9] Geospatial Resource Portal - gisdevelopment.net

[10] ZDNet - news.zdnet.com