TOWARD SEMANTIC WEB INFRASTRUCTURE FOR SPATIAL FEATURES' INFORMATION

R. Arabsheibani [a*,] S. Ariannamazi [b,] F. Hakimpour [a]

[a]Department of Surveying and Geomatics engineering, College of engineering, University of Tehran, Tehran, Iran
[b]Department of Geographic Information Systems, Faculty of Civil Engineering, Kerman Graduate University of Technology, Kerman , Iran
(rasheibani, sobhan.namazi,fhakimpour)@ut.ac.ir

KEY WORDS: Semantic Web solution, Dynamic Schema, OSM, storage and retrieval, GeoSPARQL

ABSTRACT:

The Web and its capabilities can be employed as a tool for data and information integration if comprehensive datasets and appropriate technologies and standards enable the web with interpretation and easy alignment of data and information. Semantic Web along with the spatial functionalities enable the web to deal with the huge amount of data and information. The present study investigate the advantages and limitations of the Spatial Semantic Web and compare its capabilities with relational models in order to build a spatial data infrastructure. An architecture is proposed and a set of criteria is defined for the efficiency evaluation. The result demonstrate that when using the data with special characteristics such as schema dynamicity, sparse data or available relations between the features, the spatial semantic web and graph databases with spatial operations are preferable.

1. INTRODUCTION

Modelling of the geospatial data in the Semantic Web structure offers advantages for ease of computational and analytical procedures, although it confronts impediments and challenges. In particular circumstances in which annotations and spatial information do not comply with a unit schema or prototype, or schemas are dynamic and gradually change due to the information applications, the conventional models like relational databases cannot efficiently accomplish the process of storage and retrieval of every aspect of the data sets such as information relations. In this condition the new approach of Semantic Web can benefit the information modelling. Constructing a spatial data infrastructure must follow a specifically designed architecture. The aforementioned architecture should be able to deal with the data heterogeneity, sporadic data source, and variety of data types and information schemas. Hence the outputs must be in a way that a calculation machine in a processing procedure could utilize the outputs, i.e. rather than human perception, computers can understand the outputs. Considering the availability of different aspects of the data and the relations among the geo-referenced objects in the implemented infrastructure, "data interpretation" can be more comfortable for users. In the present study we encounter a data set that does not follow conventional data schemas, so our modelling requires a theoretical background that supports this kind of datasets. We have developed an architecture by which the Open Street Map (OSM) data, the epitome of Volunteer Geographic Information (VGI), can be used in a spatial data infrastructure. A modelling can merge dynamic attributes to geometry objects. This research will follow the goal of building a Spatial Semantic Web model for a set of point of interest's information and try to evaluate and compare the spatial operation results with other conventional models such as relational models by combining the mentioned components including:
- spatial and aspatial data that can be shared and changed in type
- modelling that support theoretical background for the data infrastructure

- GeoSPARQL and SPARQL as a unique storage and retrieval language for spatial and aspatial data respectively
- graph database management systems (GDBMS) that support spatial operations

We test some operations like injecting the database with huge amount of data and extended datasets and spatial indexing in spatial-semantic Web database –i.e. graph databases that support spatial prototypes- to evaluate whether or not this kind of modelling can handle the operations in an acceptable time. Besides a set of spatial or spatial-related information query are run in modelled data over graph database in parallel with relational databases. The result showed us that if one chose an efficient semantic technology, the data conversion, data storage and information retrieval procedures would be done in an acceptable period in semantic web databases. Moreover it will be proved that using OSM POIs' information that does not have any special schema and can be changed by any user in graph data bases offer the benefit that make data interpretation more comfortable. Considering the lack of tools that relational databases provide for object relation modelling – such as JOIN command for the information tables- the spatial semantic web model offers an appealing possibility of spatial objects' relation modelling that can be used in a variety of applications such as spatial-social network analysis and spatial-semantic similarity evaluations.

A variety of applications and studies such as linked information analysis researches have used Spatial Semantic Web for the retrieval and storage of the spatial data. We will overlook the researches from two different aspects. Firstly, we will review the literature from the perspective of the state of the Spatial Semantic Web. Simon Scheider (2008) has modelled the characteristics of the roads and streets such as intersections and limitations as POI's information on a Spatial Semantic data base. Although the model is designed considering the semantic criteria, the schema and the developed system is not web-based and can be used on desktop applications. Using the simple web standard WFS, Fonts et al. (2010) has modelled the storage and the retrieval of the spatial information. Basically, WFS cannot offer an appealing performance for semantic data because the standard's design doesn't have the purpose of spatial semantic modelling. Hence, there is a challenge that has made developers

to do a considerable endeavour to cover the shortage of the WFS in semantic applications. As a following research, Lin (2011) proposed a framework for the storage and extraction of the volunteer geographic information (VGI) to ease the process the knowledge transfer. Ye (2011) had also developed a system being able to categorize and classify POI's annotations in Semantic Web structure. Two latter studies have utilized the Semantic Web although they do not comply with a specific standard and forced to design a proprietary schema. The present study follows the codes of the GeoSPARQL standard defined and supported by the OGC. In another research Baglatzi (2012) has proposed a complicated modelling for using the semantic information as a middle layer that makes the semantic queries possible. Also Michele Ruta (2012) proposed a model which the POI's information is injected in model with other features to find the most desirable path considering semantic and spatial criteria. These two researches only respect some part of the GeoSPARQL standard. Robert Battle (2012) in his research stood up for the definition of the GeoSPARQL and have mentioned the capabilities of the standard that enable semantic researches. A crew of Parliament (a spatial semantic database) developers have introduced and supported GeoSPARQL in the process of designing the database.

In the literature and from the data types perspective that is used in the research procedure, Fernández (2008) have studied social information and economic data as annotations and geo-referenced to deal with the problem of the spatial semantic storage and retrieval of the data. Caro and Varanka (2011) have developed a semantic framework on the topographic points and topography related annotations (such as peaks, valleys, contours and so on) and the relation between the points to enhance the base maps. Ballatore (2014) have extracted OSM schema and evaluated the quality of relatedness between them and measured the similarity between the content of the OSM extracted schema. The difference between the present study and the mentioned research is that we have analysed OSM from the perspective of data contents rather than OSM data schema.

2. METHODOLOLGY

There is a variety of spatial semantic modelling and corresponding knowledge extraction systems. The type of the system is defined and classified based on the interactions between the introduced system and the environment that system is about. Using the spatial semantic modelling for knowledge extraction enhances the capability of the systems to deal with the multisource data and increase the power of the analysing in incomplete information (Kuipers 2000). Each category in the datasets could have some specific and distinctive characteristics that leads to special annotations. As an example consider a restaurant that offers home-made drinks and midday restrooms that is not available as a service on other restaurants. If one is about to model this annotations in relational models, there will be lots of columns that most of them contains meaningless information or null values. This condition may add inconsistency, inhomogeneity, complexity and more cost to the model. Spatial Semantic Web has the capability to manage data with dynamic schema.

There is required a set of data component and probably their relationship according upon the purpose of the semantic spatial modelling of the information in an application. The architecture of the modelling must contain different layers with a specific functionality and a set of assigned operations in each of which. The components of the modelling relate to each other through the Web and networks and the costumers can find and bind to the facilities that the model offers. The model developed by the present study has (Figure 1) includes four distinct layer namely data acquisition layer, data pre-processing layer, data storage layer, and application and representation layer. If a physical sensor is supposed to be included in the design of the system, another layer, namely physical layer, will be added to the proposed architecture.
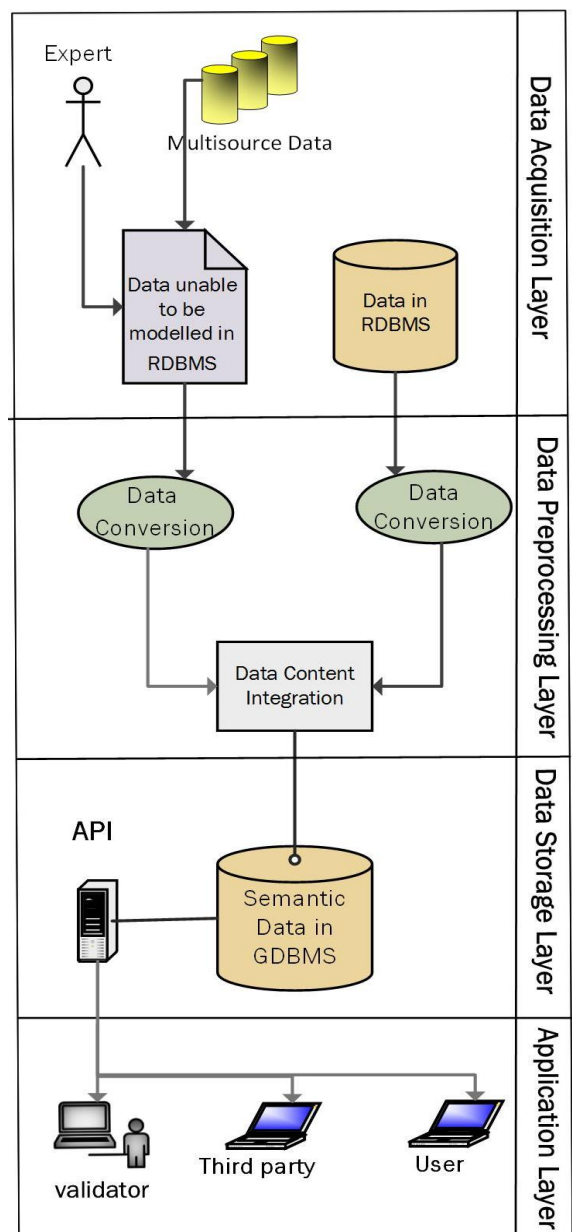

Figure 1. Spatial Semantic Web Multi-Layer Architecture

Data acquisition: The information must be converted to the triples format in order to inject the triples to the graph databases that are the foundation of every semantic web system. A part of the information that is used in this study is extracted from the free and open source information which is stored on the relational models. For the adaptation of the information for the system that is the purpose of this research, the information is converted from relational formats to graph-based formats and triples. If some data is not available, in relational databases a specific amount of memory will be assigned with null content however, in graph databases corresponding arc or predicate will be omitted from the main graph. Generally, two types of information are in hand, the first one is the information which is

modelled on the relational databases. The other type of information is from the multiple sources that doesn't comply from a static schema and the schema may change spontaneously by adding the new data. To acquire the unity in order to build a seamless system, each type of information must be converted to the triples. The process of the conversion includes assigning a specific URI based on a standard to each unit of the information and each part of the triples including subjects, predicates and objects. In this study we extracted OSM point of interests and convert the coordinates and other annotations based on GeoSPARQL specifications. The relations and other attributes are linked to the modelled POIs.

Data Pre-processing: In this layer there is the procedure in order to unify and inject the converted information into a graph database. The relational model have more supporting technology and formats in comparison with the graph databases, so the conversion process should be inclusive enough to handle the data inhomogeneity because the data in question does not follow a static schema. A specific standard must be considered in the process of data conversion in order to achieve more interoperability.

Data Storage: Undoubtedly there it is essential to use a specific standard in order to perform the queries. Therefore, GeoSPARQL query specification is considered in the process of data conversion and information query. The main element of data storage layer is a graph database supportive for spatial operations that could handle the process of data storage and retrieval.

Application: in order to perform the information retrieval it is required to design an interface between the database and the users. The retrieved information either may be required directly by the users or a third party software may use the information as the inputs or a part of the elements in an overhead program. A temporary supervision may be required to check the quality and validity of the retrieved data.

Due to explained layers in the designed architecture, the proposed model actually can be used as a conceptual model for the data schema and content conversion and storage and the retrieval.

3. RESULTS AND DISCUSSION

A set of points of interest extracted from the OSM such as shops, military spots, airports, banks and cinemas are considered in the process of spatial semantic modelling. The utilized points are the subclasses of geo:Geometry and geo:Feature. These classes are stored as triples and relate to other annotations such as coordinates, name, ID, categories (if available) other attributes. The study area is in Tehran-Iran and the extracted data from the OSM site about the study area is in XML format and with *.OSM extension. Although the main OSM interface provide desirable outputs data based on the spatial intervals assigned by the users, there is some APIs with the ability of query exchange (such as OverpassAPI & geofabrik) but still these APIs' functionality are available for a country or include a limited area. The OSM XML includes a block of the nodes, a block of the arcs and streets and a block consisting of the relationships between the features in the nodes and ways.

One of the characteristics of the OSM is that every user can edit and add his/her own desirable annotation to the features that can be considered either as an advantage or disadvantage for these VGI source contributors want to encounter with minimum proof and formal certification to be able to edit the map comfortably that these degree of freedom imposes semantic complexity and redundancy. Considering these fact, Tehran OSM POIs dataset includes approximately 1.5 million annotations and records. Generally there are 252 different attributes in the dataset, so if one wants to model the data with the relational databases the result table would consist of 252 distinct columns. As it was estimated not only none of the features has whole 252 attributes but the most available attributes for a feature is 161. Illustrated in Figure 2. is the histogram of the attributes for the POIs. If relational databases were the framework of the modelling, about 90% of the tables' cells would include meaningless and null values (Figure 3.).

For the conversion of OSM XML to prevailing semantic formats such as RDF/XML, Turtle, Notation3, N-triples and GeoJason two different solutions are possible. The first one is to convert data to shapefile format and in another conversion process convert the result shapefiles two semantic formats. Two points should be considered if this solution is chosen; each shapfile contain only one kind of spatial features such as points, lines, polygons etc. and it is not possible two model the relationships between the features in shapefiles. The second solution is using convertors which can convert the data directly from the OSM outputs to the semantic formats. Some of the convertors are able to convert spatial and aspatial attributes simultaneously while others can convert only one of the spatial or aspatial annotations. In the present study we have used Datalift program in order to convert the data which is able to convert spatial and aspatial annotations. Among different
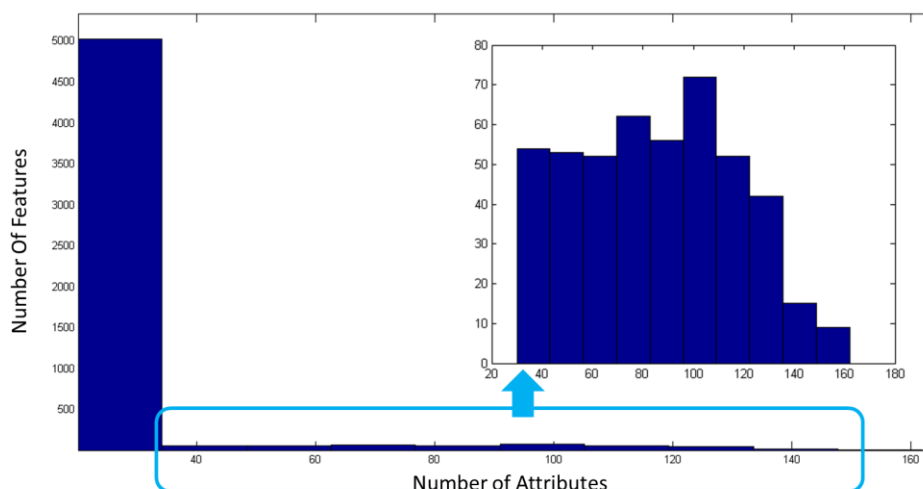

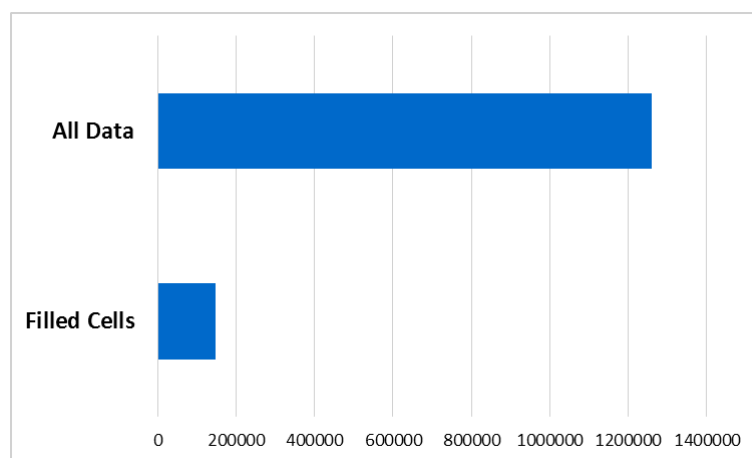Figure 2 Histogram of the Meaningful Features Attribute

Figure 3 Amount of the Meaningful Data In Comparison With All Possible Data

technologies the Parliament triple store is considered for the storage and retrieval procedure because the Parliament comply fully from the GeoSPARQL specifications and support all the geometrical and topological function and operation like spatial and temporal indexing that is based on the R- tree indexing method.

After the conversion of the data and injecting to the Parliament database In order to evaluate the efficiency of the developed model spatial queries and operations are accomplished on the stored POIs. The queries and operations are designed to test the general characteristics of the model such as injection and indexing operations, general spatial queries, spatial queries within a specified interval, spatial join and vicinity and neighbourhood queries (Table 1.).

Table 1. Spatial semantic operation and Queries

Name	Query or Operation Performed On the Spatial Semantic Model
o1	Injection of the spatial semantic data of Tehran to the Parliament database
o2	Spatial indexing of the stored data
q3	Retrieving the name of the POI by having its coordinates
q4	Retrieving the coordinates of the POI by having its name or other annotations
q5	Retrieving POIs within a defined polygon
q6	Calculation of a buffer around a specific POI
q7	Finding the three nearest POI to a specific feature

In order to perform the operations o1 and o2 a similar hardware are used and no task are ran additional to necessary operating system task. The system was not connected to the Web and similar data are used for the injection procedure. The consumed time for the injection the data and spatial indexing are compared in two different kinds of databases such as PostgreSQL V8.1 +PostGIS 2.0 and Parliament V2.7.9. Although we expect the

processes to be very time expensive in graph database because of the nature of the triples, it turned out with inconsiderable difference and nearly the same (Figure 3 and Figure 4). Query q3 does not include any specific spatial operation and retrieves objects that have a determined name. Moreover, Query q1 may retrieve more than one object from the database. Shown in figure 5 is the GeoSPARQL syntax of q1. Queries q4-q7 demonstrated in GeoSPARQL in Table 2. The consumed time for the queries was almost the same in graph and relational database.

In addition to the comparison of the relational and graph databases, it is observed that performing of the spatial indexing operation before the process of the retrieval enhances the time efficiency of the developed system. There is demonstrated in Table 3. the graphical retrieved information on satellite imagery of the Tehran on Google Earth.
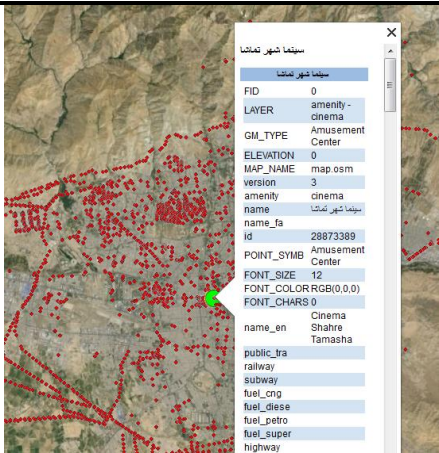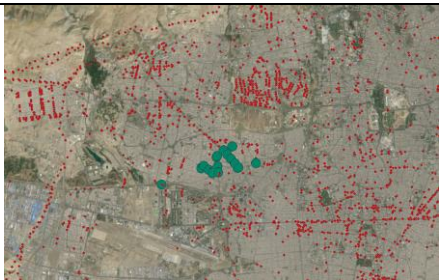
4. SUMMARY AND CONCLUSION

As previously mentioned, Spatial Semantic Web enables the users to utilize the spatial functionalities such as spatial indexing and spatial queries on a huge amount of spatial semantic RDF data. There is a variety of standards like st-sparql and GeoSPARQL and also technologies like the Parliament triple store that we have used in order to store and retrieve semantic data. Consumed time period for the spatial queries and database operation like injection of the huge data and spatial indexing doesn't have any considerable discrepancy and generally it can be judged as an acceptable period for the time for both relational and graph databases. When using the data with special characteristics such as schema dynamicity, sparse data or using more than one identification for a spatial feature, the spatial semantic web and graph databases with spatial operations are much more preferred. Also, changing and editing the model is more convenient in cases that overall schema could change. Moreover a spectrum of spatial and topological functions is available in the GeoSPARQL specifications enhancing the capability of the standard to respond to the requirements of developers and system.

Table 2. Spatial semantic Operation and Queries

Name	Query in GeoSPARQLSyntax
| 4 | ```
1  SELECT ?s ?fName
2  WHERE {
3  ?s geo:hasGeometry ?p .
4  ?p geo:asWKT ?fWKT .
5  ?s poi:name ?fName .
6  FILTER (geof:sfContains(?fWKT,
7         "POINT(51.3427 35.7334)"^^geo:wktLiteral))
8  } LIMIT 1000
``` |
| 5 | ```
1  SELECT DISTINCT ?o WHERE {
2    ?s data:geometry ?p .
3    ?p geo:asWKT ?o
4     FILTER (geof:sfWithin(?o,"POLYGON ((51.3427 35.7334,
5         51.3558 35.7262,  51.3046 35.7168,
6         51.3427 35.7334))"^^geo:wktLiteral))
7  } LIMIT 1000
``` |
| 6 | ```
1  SELECT (geof:buffer(
2         ?pWKT, 100, uom:metre) AS ?buf)
3  WHERE
4  {
5  ?p POI:name "Cinema Shahr Tamasha" .
6  ?p geo:hasGeometry ?pGeom .
7  ?pGeom geo:asWKT ?pWKT .
8  }
``` |
| 7 | ```
1  SELECT ?POIID ?POIC
2  WHERE {
3  ?s geo:hasGeometry ?p .
4  ?p geo:asWKT ?fWKT .
5  ?s poi:name "Cinema Shahr Tamasha" .
6  ?r poi:label "Cinema" .
7  ?r poi:id ?PointID .
8  ?p geo:hasGeometry ?pGeom .
9  ?pGeom geo:asWKT ?pWKT .
10 ?p poi:id ?POIID .
11 BIND (geof:distance(?pWKT, ?POIC ,uom:metre) AS ?dist)
12 }
13 ORDER BY ASC (?dist)
14 LIMIT 3
``` |

Table 3. Geographical representation of retrieved data

Name	Retrieved data representation	Name	Retrieved data representation
q3 and q4		q6	
q5		q7	

REFERENCES

Alkyoni Baglatzi, Margarita Kokla, Marinos Kavouras. (2012). Semantifying OpenStreetMap. Paper presented at the Proceedings of the 5th International Terra Cognita Workshop 2012 - Foundations, Technologies and Applications of the Geospatial Web, CEUR Workshop Proceedings.

Ballatore, Andrea, Bertolotto, Michela, & Wilson, David C. (2014). An evaluative baseline for geo-semantic relatedness and similarity. GeoInformatica, 18(4), 747-767.

Caro, Holly K, & Varanka, Dalia E. (2011). An analysis of spatial relation predicates in US Geological Survey feature definitions: US Geological Survey.

Fernández, Tatiana Delgado, & Fernández, José Luis Capote. (2008). Towards Semantic Spatial Data Infrastructures: A framework for sustainable development. Paper presented at the Proc. GSDI 10th Conference.

Kuipers, Benjamin. (2000). The spatial semantic hierarchy. Artificial intelligence, 119(1), 191-233.

Lin, Yu-Wei. (2011). A Qualitative Enquiry intoOpenStreetMap Making. New Review of Hypermedia and Multimedia, 53-71.

Michele Ruta, Floriano Scioscia, Saverio Ieva,Giuseppe Loseto, Eugenio Di Sciascio. (2012). Semantic Annotation of OpenStreetMap Points of Interest for Mobile Discovery and Navigation.

O. Fonts, J. Huerta, L. Díaz,C. Granell. (2010). OpenSearch-geo: The simple standard for geographic web search engines IV JORNADAS DE SIG LIBRE

Robert Battle, Dave Kolas. (2012). Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. Semantic Web journal, 3(4), 355-370. doi: 10.3233/SW-2012-0065

Simon Scheider, Werner Kuhn. (2008). affordance-Based Categorization of Road Network Data Using a Grounded Theory of Channel Networks. Journal of Geographical Information Science, 1-19.

Ye, M., Shou, D., Lee, W.C., Yin, P., Janowicz, K. (2011). On the semantic annotation of places in location-based social networks. Paper presented at the The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, New York, NY, USA.

Smith, J., 1989. *Space Data from Earth Sciences*. Elsevier, Amsterdam, pp. 321-332.

Smith, J., 2000. Remote sensing to predict volcano outbursts. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Kyoto, Japan, Vol. XXVII, Part B1, pp. 456-469.