Fuzzy Spatial Data Modeling for Integration of Heterogeneous Geospatial Information

Indira Mukherjee and S K Ghosh
School of Information Technology
Indian Institute of Technology, Kharagpur
Kharagpur-721302, West Bengal
Email: indira.mukherjee.23@gmail.com, skg@iitkgp.ac.in

Abstract—Geospatial data has become an integral part of many decision making processes specially related to socio-economic development. The geospatial information are often stored, maintained and processed by various diverse organizations in proprietary GIS systems. With increasing use in Internet and web services on line data sharing and ascension has increased. Thus the major bottlenecks of integrating these diverse datasets lie in the understanding of individual data structures or models, inter-relationships among the data and the accessibility of the repositories over web. In order to integrate these data, the structure of these diverse datasets are to be standardized. In addition, Uncertainty and fuzziness are inherent to the properties of spatial data. However, the current standards proposed for geospatial data modeling technique are not able to represent the fuzzy data. Thus most of the geospatial organization uses crisp data model to share and integrate the diverse datasets. However, when spatial phenomena is generalized in crisp form, then lot of quantitative informations are lost. This work presents fuzzy spatial data modeling technique, by incorporating fuzzy set theory concepts and fuzzy logic in UML class diagram. Integration of diverse, imprecise information based on the fuzzy spatial data model has also been shown using some spatial data sets.

I. INTRODUCTION

A geographic information system (GIS) or geographical information system captures, stores, analyzes, manages, and presents data that are linked to geographically referenced information. Geospatial data are special kind of data that must have some reference to locations on earth and should depict some geographic phenomenon (known as themes), like land use, population density, weather etc. related to those locations. The geospatial data are stored using a special data structure that captures information in multiple dimensions and distributions of information over space. The geospatial data [2] have some special characteristics: (i) geo-referencing, (ii) spatial and non-spatial attributes, (iii) spatial relationship - relationship may be crisp or fuzzy, (iv) temporal property.

Geospatial repository is an organized collection of geospatial information, often known as thematic layer (or layer) and usually refers some particular spatial feature. Thematic layers like, transport, land use, hydrology etc., provide the thematic information about a region. These layers are often stored and maintained independently by various organizations in different formats. Thus, for any spatial analysis, involving these heterogeneous repositories, there is a need for inter-operation and sharing of information among these diverse sources. The meta-data information or the data models corresponding to the individual repositories play a vital role in the integration process. The data model has to be harmonized for sharing geospatial data enterprise-wide.

Further, the real world is always complex and full of uncertainties. The properties of geospatial object and relationship among them may be ambiguous. The classical geospatial data models do not consider the fuzziness of the geographic objects. Thus to model uncertain objects and their properties, several effort have been made by using fuzzy set theory in classical data modeling concept [4] [11] [8] [9]. Fuzzy XML data modeling has been proposed by [4] [11] using "possibility distribution theory". Using fuzzy techniques fuzzy spatial objects and relationships have been represented [8] [9].The uncertainty in the spatial objects has been mapped in the UML class diagram using fuzzy class, fuzzy attributes and fuzzy association. The concept of fuzzy class, fuzzy association, and fuzzy attributes are incorporated in geospatial data model. The fuzzy classes are used to represent uncertain occurrence of geospatial objects, the fuzzy attributes that is used to represent the uncertain properties of the objects and fuzzy association is used to represent the uncertain relationship between spatial objects.

This paper proposes a fuzzy spatial data modeling technique, by incorporating fuzzy logic in UML class diagram. The major contribution lies in fuzzy data modeling and necessity of fuzzy data model for integrating and querying spatial data of heterogeneous repositories. The rest of the paper in organized as follows. Section II presents the overview about the fuzzy set theory. The fuzzy geospatial modeling concepts has been given in section III. Section IV presents an overview of Enterprise GIS (E-GIS) framework for integrating on geospatial services. Section V describes the integration of diverse fuzzy information through integration framework based on fuzzy spatial data model.

II. FUZZY SETS

Fuzzy set was introduced by Zadeh in 1963 [12]. Nowadays, fuzzy sets theory and fuzzy logics are widely used for capturing the inherent vagueness that exists in real life applications. In classical set theory an element either belong to a set or does not belong to a set, in fuzzy set theory containment of an elements a set is represented with a certain degree. Formally,
let $U$ be the collection of elements called universe of discourse. A crisp subset $A$ of $U$ is collection elements $U$ that is defined by a characteristic function $U_A(u)$ that assigns any element $u \in U$ to a value 1 if the element belongs to set $A$ and 0 if it does not belong to set $A$. On other hand, for a fuzzy subset $A$ of $U$ is defined by a membership function $\mu_A(u)$ for $\forall u \in U$. This membership function assigns any $u \in U$ to a value between 0 and 1 that represents degree to which and element of $U$ belong to subset $A$ of $U$. For example, a classical set of real numbers greater than 6 can be expressed as

$$A = \{u|u > 6\}$$

For similar condition fuzzy set can be defined as

$$A = \{u_A(u)/u_i|u_i \in U\}$$

Additionally fuzzy binary relation or an role between two crisp set $U$ and $V$ is a function $R : U \times V \rightarrow [0, 1]$. For example an object $o$ belong to class $C$ to a degree 0.6 by writing $belongTo(o, C) = 0.8$. Using the above idea, the operations defined on crisp set and relations, like the boolean operations (union, intersection, complement), are extended in fuzzy sets.

III. FUZZY SPATIAL DATA MODEL

So far spatial data modeling through classical UML class diagram has been investigated several literatures [3] [1]. In classical UML class the classes represents the objects and the relationships among the objects are represented by roles among the classes. Spatial data model using UML class diagram concept geospatial objects of similar group (river, road, building..) are represented by classes. The spatial relations between the geospatial objects (overlap, contain) are described by the association between classes. However, the classical data model fails to handle imprecise and uncertain information although such information exists in knowledge base and databases and has been investigated in many literatures [7] [10]. The concept of fuzziness in spatial data model facilitates structuring uncertain information that exist in real world and extensively used by real world applications. In this section fuzzy extension to of spatial data model using fuzzy UML class diagram has been discussed. A fuzzy UML diagram describes the types of fuzzy objects in the model and various kind of fuzzy relationship. A fuzzy class is used to describe a set fuzzy objects with similar structure, relationship and behavior. Three levels of fuzziness are defined in the UML classes.

- The first level, classes and attribute sets of a class may be fuzzy, i.e., existence of a class have possibility to the model and existence of attribute in class have a possibility.
- In order to model the first level of fuzziness, the attribute or class should be followed by a pair of words WITH mem DEGREE, where $0 \leq mem \leq 1$ is a measurement used to indicate the degree by which a class belongs to model or an attribute belong to a class. For example, class Employee in data model of students in universities may or may not need to be included in the data model. Another example, (refer figure 1) the attribute Annual Rainfall in class Town may or may not be included.
- The second level is related to the occurrence of the objects.
- To model second level of fuzziness the an additional attribute ($\mu$) is introduced to the class to represent the object membership degree to the class, with a attribute domain that is [0,1]. For example, (refer figure 1) instance of Town class is uncertain. A region may belong to Town class partially. That means it can fulfill properties of Town class with some degree. This kind of instances are denoted with membership value, the instance of the objects have some possibility degree between 0 to 1.
- Third level of fuzziness is related to the attributes that fuzzy values.
- For this a keyword FUZZY is placed in front of the attribute. The population attribute may take fuzzy values, namely, its domain is fuzzy (0.4/2300, 0.6/2500, 0.8/3000, 1/3100, 0.8/3400, 0.6/3700, 0.4/4000). The Medical_Facility domain of class Hospital is fuzzy it takes the values in terms of linguistic expression like good, medium, poor.

There are several types of fuzzy relationship between the fuzzy classes:

1) Fuzzy Generalization: Fuzzy generalization is when a new class is produced from another class called super class, by inheriting all of its attribute and methods. Let A and B be the (fuzzy) classes with first level of fuzziness A WITH $a_{deg}$ DEGREE and B WITH $b_{deg}$ DEGREE, $\beta$ be the given threshold, and the degree that B is the subclass A be $\mu(A, B)$. Then B is subclass of A

$$(\forall e)(\beta \leq \mu_B(e) \leq \mu_A(e)) \land (\beta \leq b_{deg} \leq a_{deg})$$

Here $e$ is the fuzzy objects of class A class B and, $\mu_A(e)$ $\mu_B(e)$ are the membership degree of $e$ to class A and class B respectively. $a_{deg}$ and $b_{deg}$ are the scalar membership degree, i.e., the degree to which the class belongs to the model.

2) Fuzzy Association: Two levels of fuzziness in association between classes can be identified. The first level is association fuzzily exists between the associated classes. Again, if the instance of the classes are uncertain then the association between the classes are fuzzy.

The second level of fuzziness in the association occurs due to uncertainty in object instance of the classes. The association $ass(A, B)$ between two fuzzy classes is one without
The first level of fuzziness in association between classes with first and second level of fuzziness can be given by

$$
\mu_{ass}(e,f) = \min(\mu_A(e), \mu_B(f), a_{deg}, b_{deg}, deg_{ass})
$$

Here e,f are the instance of class A and B respectively, $\mu_A(e)$ and $\mu_B(f)$ are object instance membership of class A and B respectively. $a_{deg}$ and $b_{deg}$ are membership degree to which the class belongs to the model and $deg_{ass}$ is the association degree.

The association between the class represents the relationship between the objects. For example (refer figure 3) An association Connectedby exists between Hospital class and Road class. The association exist with possibility degree($deg_{Connect}$) let it be 0.8 degree the object membership degree ($\mu_{Road}(e)$) of object e of Road class is 1.0 degree since Road class is without any fuzziness at the three levels. The object membership degree ($\mu_{Hospital}(f)$) of object f of Hospital class is also 1.0 degree. Thus, the fuzziness exists in the association Connectedby itself.

IV. ENTERPRISE GIS FRAMEWORK FOR GEOSPATIAL WEB SERVICES

Geospatial repository is an organized collection of Thematic Layers on several perspectives, e.g. Landuse, Hydrology, Vegetation etc. Each of these layers provides the corresponding thematic information over the region. Information about each of these layers is usually collected and maintained by several federal agencies in a distributed fashion. This necessitates the need for a flexible approach that will facilitate sharing of information which can ensure interoperability. The service-oriented architecture (SOA) provide the functionality of discoverable and dynamically bound, self contained and modular, stress interoperability, loosely coupled, network-addressable interface, location transparent, composite modules, comprised of components. Services are self-contained, stateless, and do not depend on the state of other services. The proposed Enterprise GIS (E-GIS) framework aims at integrating the independent geospatial datasets and service providers distributed over the web. The service-oriented GIS model uses OGC complaint service standards, namely, Web Feature Services (WFS), Web Map Services (WMS), Web Registries Services (WRS) etc. The specifications allow seamless access to geospatial data in a distributed environment [6] using spatial web service technology. The overall service base integration framework of the E-GIS system is shown in Fig 2.

As discussed earlier, the geospatial information of a region is categorized into different themes (or thematic layers) like land-use, settlements, transport etc. For any decision making process, these themes act as information sources, which are stored and collected independently by various organizations. For queries involving one or more themes, the data need to be integrated and represented in overlaid form by the E-GIS system. Further, for achieving seamless data sharing across organizations, there is a need to address both the proprietary data models and syntactic interoperability. Hence, the data models corresponding to the decision support system (or the base model) and the individual data repositories are to be generated and advertised using web services. The appropriate data models or schema’s make the backbone of the data sharing and interoperability framework.

V. INTEGRATION OF IMPRECISE INFORMATION USING FUZZY SPATIAL DATA MODEL

In order to realize the integration of heterogeneous datasets and actual transfer of data, a spatial data infrastructure build on the fuzzy schema need to be established. An enterprise GIS (E-GIS) framework aims at integrating the independent geospatial data providers through spatial web services. E-GIS is built on OGC standard [5] so that it can be easily interfaced with various geospatial service providers and consumers.

In this section, the query resolution through geospatial data integration and fuzzy spatial data model has been presented with help of an example. Some test data sets have been used to manifest the work flow. The data sets are categorized into following themes.

- Region
- Hospital
- Road

Generally, the datasets of these themes are stored by different organizations. Each organization maintains its own standard for data encoding and access mechanism. The data
are needed to be integrated for operating on spatial data. In the given example whole system is categorized into three different layers:

- **Layer 1** contains the raw datasets.
- **Layer 2** represents the E-GIS system that integrates to distributed datasets.
- **Layer 3** represents the end-user which accesses the service from E-GIS system.

A fuzzy geospatial data model has been developed (refer figure 3) that defines the structure of the datasets Region, Hospital, Road. A Region is defined with attribute Name and Shape. Hospital is has attribute Category, defines the category of hospital (e.g. Government, private, semi-government), No.Beds, No.Doctors, Medical_Facility and Shape. The class Road has two attributes Road_Name, Road_Type and Shape. The Medical_Facility attribute of class Hospital is fuzzy attribute. It stores the value in linguistic terms good, medium, poor. The value depends on ratio of number of beds is to number of doctors (number of beds/ number of doctors) in hospital. The optimum ratio is decided by domain expert, in this case study the well suitable ratio is considered as 7.0. i.e. if there are 70 beds then number of doctors in the hospital should be 10. The domain of the ratio is from 1 to 20. The GML schema 4 is generated based on the fuzzy spatial data model, based on which data are shared and integrated.

![Fuzzy Spatial Data Model](image1)

The fact that Medical_Facility of hospital is good if the ratio (R) is 1 \(\leq R \leq 7.0\) and number of beds should be \(\geq 50\). In this case study second condition is true for all hospitals. Now, in real world application, the one may get a ratio like 7.1 or 7.2 or 8 as well, classifying Medical_Facility with \(R=7.0\) as good and \(R=7.1\) as medium is intuitively unreasonable. Thus a fuzzy set theory has been applied to capture the fuzzy values. The ratio \(R\) is characterized by membership function \(\mu_{MF}(R)\) (refer figure 5) that gives fuzzy sets flexibility in modeling common linguistic expression.

The membership function \(\mu_{MF}(R)\) of Medical facility with respect to ratio of beds/doctor is defined by left open trapezoid membership function. The trapezoid membership functions are extensively used in real time implementations due to their simple formula. For same reason the trapezoid membership function has been used here. Describing \(\mu_{MF}(R)\) function of hospital with respect to \(R\), universe of discourse of \(R\) are 1 to 20. The value of good Medical_Facility reaches the peak (membership value 1.0) when ratio \(R\) is within 0 to 7 and gradually falls i.e between 0 to 1 for \(R\) between 7 to 20. In similar manner other membership functions of Medical_Facility medium and poor are described.

The Medical_Facility=good of hospital can be given as follows.

\[
\mu_{MF_{\text{good}}}(R) = \begin{cases} 
1 & \text{if } 1 \leq R \leq 7 \\
\frac{(19 - R)}{12} & \text{if } 7 \leq R \leq 20, \\
0 & \text{if } R \geq 20
\end{cases}
\]

The Medical_Facility=medium of hospital can be given as
The Medical Facility=poor of hospital can be given as follows.

\[
\mu_{MF_{\text{poor}}}(R) = \begin{cases} 
0 & \text{if } 1 \leq R \leq 7 \\
((R - 7)/7) & \text{if } 7 \leq R \leq 14, \\
1 & \text{if } 14 \leq R \leq 20 
\end{cases} 
\]

Suppose data provider \( P_1 \) have data of Region and \( P_2 \) have data of Hospital, \( P_3 \) has data of Road. An user wants to retrieve the hospital of a region \( R_X \), connected by road and have good Medical Facility both with threshold 0.6. Now, suppose there 8 hospitals, the beds is to doctors ratio is given by the data provider or it is calculated by data consumer from the available information of beds and doctors. The membership value \( \mu_{MF}(R) \) of good Medical Facility is calculated based on equation 1. In table I the detail of hospitals according to the fuzzy data model is given.

### TABLE I

#### MEDICAL_FACILITY VALUE CALCULATION

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Beds</th>
<th>Doctors</th>
<th>Ratio</th>
<th>( \mu_{MF}(R) ) of good Medical_Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h_1 )</td>
<td>100</td>
<td>10</td>
<td>10.0</td>
<td>0.75</td>
</tr>
<tr>
<td>( h_2 )</td>
<td>80</td>
<td>5</td>
<td>16.0</td>
<td>0.25</td>
</tr>
<tr>
<td>( h_3 )</td>
<td>110</td>
<td>15</td>
<td>7.33</td>
<td>0.97</td>
</tr>
<tr>
<td>( h_4 )</td>
<td>80</td>
<td>10</td>
<td>8.0</td>
<td>0.91</td>
</tr>
<tr>
<td>( h_5 )</td>
<td>120</td>
<td>20</td>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>( h_6 )</td>
<td>90</td>
<td>12</td>
<td>7.5</td>
<td>0.95</td>
</tr>
<tr>
<td>( h_7 )</td>
<td>120</td>
<td>10</td>
<td>12</td>
<td>0.58</td>
</tr>
<tr>
<td>( h_8 )</td>
<td>60</td>
<td>7</td>
<td>8.5</td>
<td>0.87</td>
</tr>
</tbody>
</table>

The hospitals that satisfies the query is given as \((h_1, h_3, h_4, h_5, h_6, h_8)\). The next thing that has to calculated is the roads that connects the hospitals with threshold above 0.6. The spatial relationship of the each hospital with road has been calculated using spatial operations. The concept of connectivity is also defined by domain expert, in this case study the connectivity of hospital and road is defined as road within 150 radius of hospital taking hospital as the center. In table II the membership value of Hospital-Road connectivity is given for considered datasets. The membership value of connectivity is calculated based on 4.

The Hospital-Road connectivity \( \mu_{C}(d) \) membership function with respect to Road Distance \( d \) is given by

\[
\mu_{C}(d) = \begin{cases} 
0 & \text{if } 0 \leq d \leq 150 \\
((800 - d)/650) & \text{if } 150 \leq d \leq 800, \\
0 & \text{if } d \geq 800 
\end{cases} 
\]

From the above calculation given in table II the hospital connected by road with threshold membership value \( \geq 0.6 \) are \( h_1, h_3, h_4, h_5 \). However, in case of crisp data modeling and integration process on hospital \( h_1 \) would have qualified. However, though other hospitals disqualifies the given property of good Medical Facility and connectivity to road by little difference but crisp data model is unable to capture it. In other hand, fuzzy spatial data model is able to represent those information which are close to the given condition. THus using fuzzy spatial data model a realistic and flexible result is obtained, which may be important for real world decision support system. The integrated data based on fuzzy data model and crisp data model is given in figure 7 and 8. To resolve the structural heterogeneity issue the data are shred in form of GML. The GML generated based on fuzzy data model is given in figure 9.
VI. CONCLUSION

This paper presents a fuzzy spatial data modeling approach for designing and structuring imprecise information. In the proposed framework, the heterogeneous, imprecise spatial data are integrated through E-GIS based on fuzzy spatial data model. An extension to UML class diagram has been discussed in the paper based on which the uncertain spatial data are modeled. Based on fuzzy set and fuzzy set theory operation, the uncertainty in the information is managed efficiently. The integration and query processing based on fuzzy spatial data model gives more realistic information that are used in various decision making processes.

REFERENCES