Focus

Soft Computing 7 (2003) 402 – 411 © Springer-Verlag 2003
DOI 10.1007/s00500-002-0229-z

An agent system for managing uncertainty in the integration of spatio-environmental data

F. E. Petry, M. A. Cobb, M. Paprzycki, D. Ali

Abstract Recent applications in environmental systems have necessitated the integration of data from multiple, heterogeneous sources. The integration process involves challenges related to issues of uncertainty and imprecision associated with both the data and the process itself. While the handling of uncertainty in geographical information systems (GIS) has been a focal point of research in recent years, the additional challenges of dealing with multiple data sources and types, as well as specific fields of analysis, lead to much more complex situations. In this paper, we present a framework for the use of fuzzy mobile agents to address these additional challenges from the standpoint of large-scale environmental systems.

Keywords Fuzzy databases, Information retrieval, Large-scale systems, Geography

1 Introduction

A requirement of large-scale environmental systems is that a wide variety of data sources must be integrated, including especially data of geospatial format. For example, a study of water pollution from industrial and agricultural sources in a major river basin such as the Mississippi River involves information of tremendous diversity. This can include spatial hydrological and ecosystem descriptions, long term rain/snow fall, meteorological records, land use spatial coverage maps, etc.

A major problem with geospatial data, such as mentioned above, is the variety of uncertainty and imprecision that is associated with this form of data. The need to handle imprecise and uncertain information concerning spatial data has been widely recognized in recent years [4], particularly in the field of geographical information systems (GIS). GIS is a rather general term for a number of approaches to the management of cartographic and spatial information. At the heart of a GIS is a spatial database. The spatial information generally describes both the location and shape of geographic features in terms of points, lines and areas. The full use of many sources of spatio-temporal data is essential for comprehensive modeling of large-scale environmental problems.

A number of researchers in the GIS and spatial database area have explored models of spatial data using fuzzy set approaches. This can be seen, for example, in the recent edited book on the modeling of geographic objects with indeterminate boundaries [4] and the volume of *Fuzzy Sets and Systems* on the topic of uncertainty in GIS and spatial data [10]. Clearly, in order to address the integration of the various spatial data sources for large-scale environmental systems, approaches such as these must be considered.

Some early work by geographical scientists in the 1970s utilized fuzzy sets in topics such as behavioral geography and geographical decision-making [20, 28, 35]. However, the first consistent approach to the use of fuzzy set theory as it could be applied in GIS was developed by Robinson [38–40]. More recently, there have been a number of efforts utilizing fuzzy sets for spatial databases including: capturing spatial relationships [9, 10], querying spatial information [33, 49], and object-oriented modeling [11, 21, 34].

There are a number of specific spatial data topics that are relevant to environmental systems, including soils, land form classification, ecosystems, etc., for which uncertainty models have been developed. Issues of vagueness and inexactness in soil classification have been investigated by researchers [26, 32] and especially by Burrough [3, 5]. Uncertainties related to vegetation science description have been reported in [12, 19, 37]. The classification of landforms and land cover has also been the subject of considerable research [6, 15, 47]. Mackay and Robinson have developed an approach of combining the sub-models of larger, integrated ecosystem models by using fuzzy logic to combine conflicting results [31].

In a more general sense, the management of uncertainty for spatial information in environmental modeling [30] and decision support systems [29, 42] has also been extensively considered. Integrated assessment models of global climate change [36] have acknowledged the issue of handling uncertainty in such models at several levels [48]. Recently Yazici and Petry [53] have developed an approach to cultural theory using fuzzy logic that can be utilized for integrated assessment to provide quantitative representation of the social implications of decision

F. E. Petry (✉)
Naval Research Laboratory Mapping,
Charting & Geodesy Stennis Space Center,
MS 39529
e-mail: petry@eecs.tulane.edu

M. A. Cobb, M. Paprzycki, D. Ali
Department of Computer Science & Statistics,
University of Southern Mississippi,
Hattiesburg, MS 39406-5106

We would like to thank the National Imagery and Mapping Agency, the Marine Corps Warfighting Lab, PE 0603640M, and the Office of Naval Research, PE 0603238N, for sponsoring this research.

making in such large-scale environmental management systems.

A major obstacle in large-scale environmental modeling systems is obtaining the required data from a large number of varied and distributed sources. We are developing an agent-based system to integrate such distributed heterogeneous environmental data, and are particularly concerned with the management of uncertainty and quality of data merger. This paper describes research developing autonomous updating methodologies to provide for the collection and integration of geospatial data from multiple sources, including web-based repositories, into a single database system for subsequent access and retrieval. Intelligent mobile agents are used as the primary mechanism for data identification and collection, integration (including conflation) and quality monitoring.

2
Spatial data integration approaches

Autonomous updating subsumes several research issues that must be resolved for a successful system implementation. Among these are integration of heterogeneous geospatial data types, resolution of multiple representations (conflation) and data validation. On the issues of data integration and conflation, we first define several of the most frequently used terms and their interrelationships within the general scope of GIS interoperability. These terms – *interoperability*, *integration*, *conflation* and *fusion* – are often used to convey very different ideas, or alternatively, used so loosely as to be somewhat interchangeable. Therefore, clarification of the use of these terms here will be beneficial. Table 1 shows the 3-tier hierarchy illustrating our use of these terms.

At the lowest level of the hierarchy is the concept of data integration. In keeping with the most widespread use of this term, e.g. [16], our use of data integration is intended to convey the idea of some process whereby incompatibilities among varying spatial data formats is resolved, allowing the various data types to be simultaneously analyzed/displayed/processed by a GIS. Data integration is therefore regarded as a low-level transformation procedure that requires no semantic knowledge of the various data. Integration of data types can be considered within the context of a single GIS, for example, the integration of vector and raster data for display purposes, or as part of a distributed system. The problem with integration techniques is that they tend to be ad hoc, resolving only specific formats for a particular application. Our goal for this approach is to develop a more general, robust integration methodology that is valid for the majority of existing geospatial data formats. The use of intelligent agents that incorporate fuzzy logic gives the ability to apply semantic knowledge to integration algorithms, enabling, for example, automatic schema extensions.

Conflation is a higher-level concept than integration, because it implies a deeper (semantic and "intelligent") knowledge about the data. Conflation results in a state of agreement among various data sources in which a single, "best" view of multiple data representations for similar data types is presented to the user. Thus, conflation logically can occur only if integration as defined earlier has already been resolved. Two basic approaches to conflation exist – a statistical-based approach [41] and a knowledge/rule-based approach [7]. Statistical techniques work well for regular data, such as road networks, while a rule-based approach is more profitable for irregular or less uniform data. While a few commercial GIS products employ a conflation component, it is still a highly manual task with little automatic support. Previously, [7, 8, 17, 18] we have developed knowledge-based conflation algorithms along with an object-oriented conflation model. This work is extended by incorporating these previously-developed concepts into a conflation agent, which will be able to act autonomously and intelligently to conflate data with multiple representations.

Beyond conflation, which is viewed as an issue only among similar types of geographic information, e.g., vector with vector, the concept of data fusion is the more generic idea of combining widely varying forms of data, e.g., multimedia, in a system that can effectively organize the information in a way that is of benefit to the user. This concept of the "omni-informational" GIS is discussed in [45].

Finally, "interoperability" is viewed as the ultimate goal, encompassing all aspects of representation and semantic integration and providing a truly seamless view of geographic data in all its many forms. A UCGIS white paper (available at http://www.ucgis.org/) notes several long-term goals related to interoperability, including machine-interpreted semantics of geographic data, improved semantic representation for the data, language support for communication of geographic information and the development of canonical data models of geographic information.

Currently there are no available capabilities for automatically and intelligently: (1) determining available network-based digital geospatial data resources, (2) integrating the various geospatial data formats into a single database schema, (3) validating data quality, and (4) conflating multiple representations. Because the nature of the problem of integration and interoperability is naturally

Table 1. Hierachy and examples of terminology

2.1.1.1 Hierarchy of terms		2.1.1.2 Examples
Interoperability	1. Fusion	Image + Text + Video + Vector + Raster + ···
	2. Conflation	Bridge representation 1 + Bridge representation 2 → "Best" bridge representation
	3. Interchange	Proprietary vector format

distributed, mobile agent technology is a prime candidate for implementing a solution. Agents can loosely be defined as any software object operating on the behalf of a person or business entity. Agents may be permanently fixed on a given host (stationary agents), or they may be capable of moving from host system to host system (mobile agents) in order to perform a given task or series of tasks. The autonomous, and perhaps even disconnected, nature of agents enables operation to a designated task or series of tasks. The following section addresses the ways in which our use of agents will resolve the problems above related to spatial data integration.

3
Agent-based system approach

Source data for comprehensive environmental systems are likely to include vector feature data, image data, terrain data and possibly other 3D or 4D (temporal) data. To account for these widely-varying forms of complex spatial data, an object-oriented (OO) approach to design and implementation issues is taken. The use of an OO paradigm is beneficial in several ways relative to implementation issues: (1) an OO approach is the generally accepted one for handling complex data; (2) web-based technologies, including Java and communication protocols such as CORBA, are object-oriented, and (3) the encapsulation properties of an OO approach are compatible with the self-contained nature of mobile agents. Specialized agent classes are used for performing continuous, automatic updates to the database, conflating the integrated data, and monitoring the quality of data added to the database. Following is a general description of the use of agents in this methodology.

3.1
Updating and integration

Updating of a database is a topic that must be considered with respect to two key items, automatic updates and conflation. The utilization of intelligent mobile agents by extending and providing them with memory capability is one avenue that can help resolve these issues. Intelligent mobile agents can make decisions based on similar conditions encountered in the past and the corrective actions that were taken.

Mobile agents offer several unique advantages over their stationary counterparts [25]. First, the ability to move to the source of activity, i.e. a database server, reduces the additional network overhead involved in remote communication. For instance, by executing a series of database queries locally, intermediate results are not required to be transmitted to the remote system; rather, transmission is delayed until the final query has been executed, thus reducing overall execution time and bandwidth. In conjunction with mobility is the ability to easily deploy software to a remote site. This merely involves instructing the agent to move to the remote site and begin execution. In this manner, a server's functionality may easily and unobtrusively be extended. Another advantage is the autonomous nature of agents. Once an agent has moved off of the client machine to another host, the client machine may safely shut down. Upon restarting, the client machine needs merely to re-establish contact to the mobile agent to regain control. Alternatively, the agent may be programmed to periodically attempt to return to its originating site.

Agents can be used to update the database (by constantly searching for new and/or updated information in previously identified repositories) and to systematically check for any discrepancy in the data collected from different sources. Upon detection of potential data sources, the agent analyzes the validity of the data and its use in filling a need for the database. Regarding performance considerations, it is extremely likely that various components of the database system will not be used with the same intensity at all times. Thus, the mobile agents will be running constantly, utilizing all available "free cycles" on all available computers to perform their tasks. In this manner, the proposed system should be capable of achieving almost 100% resource utilization.

An integration agent is used to merge new data into the existing database schema. The primary responsibility of an integration agent is to analyze the data format, and decide the best manner of incorporating the new data into the main database repository. This obviously entails issues of both data interchange and conflation components of interoperability discussed earlier.

3.2
Conflation and agent management

Conflation is the area of cartography that involves the combining of two or more data sources representing the same geographical location into a single representation of the area. Traditionally, this has been done manually; however, efforts begun in the 1980s have led to automated techniques suitable for certain types of digital spatial data. Recent efforts have enhanced the original statistical techniques used with a rule-based reasoning system capable of more sophisticated feature-matching and feature deconfliction (removing conflicting data) capabilities. Data conflicts from multi-source data are inevitable, and this presents a challenge for users who must sort out the various representations without any help from the system. Ideally, multiple representations should be detected and resolved without the user's being aware that conflicts exist. For this purpose, conflation in our methodology is based on agents that continuously or cyclically check for conflation issues as updates to the database are performed. The integration of a conflation component for the source data is a vital step in ensuring that only the best quality data are available for user access.

Conflation of data necessarily implies a component of uncertainty. That is, how do we know that the new data is a representation of something that already exists in the database? And, given that the previous issue is resolved, how do we determine which representation is "better"? The issue of data quality as it applies to web-based data sources must be evaluated on the basis of available information to aid in the conflation process. A rule-based system for reasoning under uncertainty has been utilized for resolving these issues [7].

The database agent protects the data stored in the database management system by automatically monitoring and managing its databases and applications as an integral part of the organization's computing environment.

intelligent agents can monitor the performance of key database management system (DBMS) resources and, if customizable thresholds are exceeded, send alerts to a system monitor. The system monitor can then correlate these events with other system, network, and application information to determine the cause of a problem, and correct the problem and/or notify the database administrator. This ability to anticipate problems before they impact performance helps to ensure the reliability and availability of mission-critical applications. The database agent engages in overall management activities, including activating agents for updating, integration and conflation when this is deemed necessary or desirable.

In summary, we believe that the ideal situation is a database management system that is self-adapting. More precisely, to address the above-described processes, we design a system framework in which:

a) data are collected and integrated constantly by the mobile agents and constantly is being worked on by the internal automatic conflation agents, and
b) monitoring systems will search for old data and request that new data will be searched for if the specified data becomes older than a given limit.

This system of agents enables the automatic updating and management of large databases with a degree of efficiency and accuracy that is not possible with the use of existing techniques.

4
Agent design

This framework incorporates teams of cooperative intelligent agents, all of which are capable of mobility, to provide automatic, even offline, updating of geospatial information through conflation of data collected from various source databases. Optimally, all agents are targeted at a specific functionality and kept as lightweight as possible, thus enhancing agent mobility. Each specific database format has a specially designed agent to access the repository and collect and convey data in a common, neutral format. This affords a tremendous advance in the key goal of seamless interoperability among the various geospatial data formats. A certain amount of overhead will be incurred in providing the required uncertainty reasoning; however, by leveraging object-oriented technology, most of the common features can be abstracted into ancestor classes and preloaded onto each gateway node. To promote a wider range of uses, all agents are configurable, allowing the end users to optimize on performance versus advanced feature resolution.

4.1
Fuzzy components in spatial agents

In the context of large, environmental problems we consider two important aspects of the design of our spatial agents that can utilize fuzzy components. It is clear in this problem domain that we must be able to: (1) locate the needed spatial data, and (2) be able to integrate the data that has been located. Each of these tasks can be approached by a separate fuzzy component that is part of the fuzzy agent design.

The first component is a fuzzy search or fuzzy querying component and as we have discussed, the nature of spatial data and its complexity dictates that our search must utilize fuzzy matching techniques. The approach we utilize in this component of our fuzzy agent is based on the searching techniques that have been used successfully by the researchers in fuzzy information retrieval [1]. In particular we have modeled our query formulation on that of Bordogna and Pasi [2]. They provide an approach to an extended Boolean query model by using linguistic query weights. This lets a query utilize linguistic variables and the term "important" as linguistic hedges treated as query weights. The evaluation of the relevance of some given data to the query is based on the evaluation of the function

$$\mu_{\text{important}}: [0,1] \rightarrow [0,1] \ ,$$

representing the importance of an attribute value in a spatial object that might be retrieved.

The development of a consistent ontology is always a problem for both the matching and integration of spatial data [14]. Certain governmental agencies have developed controlled representation and terminology for the large sets of spatial data they have available. One example is the development of the Vector Product Format (VPF) products by the National Mapping and Imagery Agency [13]. We have developed an approach to the matching of spatial data for conflation in VPF format [7] that can be generalized and used in our fuzzy agent design. Thus, we consider the representation of the spatial features to have the general form of attribute-value pairs from the defined classes of VPF attributes for the specific features allowed, such as bridges, lakes, railroads, etc.

4.1.1
Fuzzy matching

The matching technique developed is able to accommodate cases where values for corresponding attributes differ, as well as cases where the attribute sets themselves differ. For implementation, each feature is considered as a set of attribute-value pairs:

$$((a_{11}, v_{11}), (a_{12}, v_{12}), \ldots, (a_{1n}, v_{1n}))$$
$$((a_{21}, v_{21}), (a_{22}, v_{22}), \ldots, (a_{2m}, v_{2m}))$$
$$\ldots$$

From this representation, a degree of matching similarity is determined. A different approach is used for different attribute value domains. For numeric domains, a membership matching function is used, while a similarity table is used for linguistic domains. Our approach to linguistic attribute matching is to establish a similarity value s (in the range [0, 1]) for each pair of elements of the domain. This value is determined from the semantics of the domain and the linguistic terms. The characteristics of the similarity function s are:

$$s_A(x, y) = s_A(y, x) \quad \text{symmetric}$$
$$s_A(x, y) = 1 \quad \text{reflexive}$$

for all values $x, y \in$ domain of attribute A. For well-defined values of the domain (e.g. not "unknown" or "other")

where x is a well-defined value.

As an example, we consider the **Railroad** feature's attribute RRA (Railroad Power Source), which is restricted to the values allowed in VPF (0 – Unknown, 1 – Electrified track, 3 – Overhead electrified, 4 – Non-electrified, 999 – Other). The similarity table for RRA is shown in Table 2. Since linguistic similarity is symmetric, we need only show the lower triangular values in the table. Note that since 1, 3, and 4 are well-defined linguistic terms they are shown with the reflexive value of 1 on the diagonal, e.g. $s_{RRA}(3,3) = 1$. However, since 0 and 999 are non-specific categorical values for this particular domain, their diagonal values were determined to be less than 1, reflecting the potential lack of exact matching for such linguistic terms.

Because most features have more than one attribute, we must also consider semantic interrelationships among the attributes in determining matching features. These are represented as rules in an expert system that return associated weights. These weights are used to either add credence to the hypothesis of matching features, or to weaken it. As an example, consider the **Railroad** attributes LTN (Number of tracks) and RRC (Railroad category). The rule for the relationship between the two attributes is expressed as:

IF((RR1.ltn = 3 and RR2.ltn = 2) and
(RR1.rrc = 16 and RR2.rrc = 16))
THEN $w_{rra} \leftarrow 1.0$
$w_{ltn} \leftarrow 0.5$,

where RR1 represents the first **Railroad** object and RR2 represents the second. This rule illustrates a conflict in the values of the length attribute of the two features. We see from this example that the resulting weight for LTN is weakened, giving it less influence than that of RRA in the combined matching score.

A composite matching score is then computed from the combination of the expert system weights and the similarity table values. This score is given as:

$$MS_{i,j} = \left(\sum_{k=1}^{N} [simA_k(F_i, F_j) \times ESW_{A_k}] \right) \Big/ N$$

where $A_k = k$th attribute in both F_i and F_j, where $0 \le k \le N$, N = number of attributes that describe both F_i and F_j, ESW_{A_k} = weight associated with A_k computed by the expert system.

4.1.2 Fuzzy integration

Once relevant spatial data has been located, we are faced with the problem of appropriately integrating it into the overall primary data. We take here a data fusion approach that has been developed by Yager and Petry in a somewhat broader context [52]. For data integration or fusion one must evaluate whether the currently available spatial data is adequate for the particular problem or application at hand. If the available primary information is satisfactory then we can avoid the costly operations of having to both find and merge additional information. In case of doubt about the credibility of our primary information, or if there is some inconsistency, we can try to resolve these problems by possibly using the supplementary information that might be located.

Table 2. Similarity table for attribute RRA

RRA	0	1	3	4	999
0	0.2				
1	0.2	1			
3	0.2	0.6	1		
4	0.2	0.1	0.1	1	
999	0.2	0.2	0.2	0.2	0.2

A general approach to the fusion of information can be based upon the use of fuzzy modeling technology [51]. Within this approach one can express rules guiding the fusion process in a natural language-like manner and then use this knowledge to obtain a formal model for implementing the fusion. Our focus here is on the application of this methodology to the problem of fusing our primary information with supplementary information about related objects that we have searched for and located.

We will assume our primary information is associated with some spatial objects and that their attribute value's uncertainty can be represented by a possibility distribution. Specifically, let V be a variable corresponding to an attribute value of some object taking its value in the set $X = \{x_1, \ldots, x_n\}$ where our concern is the determination of the value of V. Using the notation of Zadeh's theory of approximate reasoning [55] we can express knowledge about the value of a variable V as

V is A ,

where A is a constraining value. For example if we are considering the mixing of industrial pollutant outflow in river systems, a critical variable is the depth of rivers and streams. So we may have an expression for the Mississippi River basin such as:
Depth of the Ohio River at St. Louis is about 40 feet.
A statement of this type influences our belief about the possibility of an element in the domain X being the actual value of V. Recalling that many types of linguistic information can be represented as fuzzy subsets, a formalization of the association of linguistic constraints with variables can be implemented using fuzzy subsets [56]. Let A be a linguistic expression describing a constraint on V and associate with it a fuzzy subset A of X. So the knowledge that V is constrained by A, V is A, affects the possible value for V and induces a possibility distribution Π on X such that $\Pi(x) = A(x)$ indicates the possibility that x is the value of V.

For example, recall the statement: Depth is about 40 feet. Here the value "about 40 feet" could be represented as in Fig. 1.

So the induced possibility is:

$$\Pi(x) = \begin{cases} 0 & x < 30,\ x > 50 \\ (x-30)/8 & 30 \le x \le 38 \\ 1 & 38 \le x \le 42 \\ 1-(x-50)/8 & 42 \le x \le 50 \end{cases}$$

One measure of the uncertainty associated with a possibility distribution introduced by Yager [50] is the

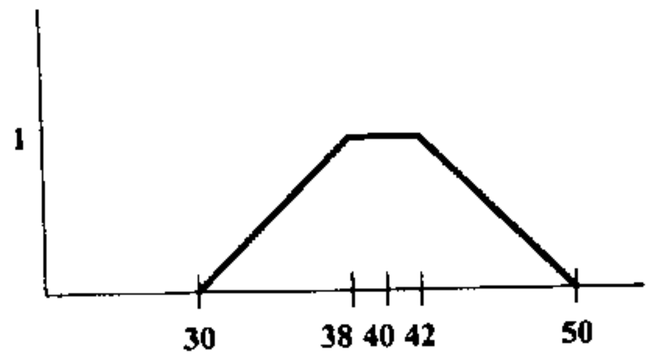

Fig. 1. Fuzzy Subset "about forty"

specificity of the distribution. Let Π be a possibility distribution defined over the set $X = \{x_1, \ldots, x_n\}$. The measure of specificity associated with Π, $\mathrm{Sp}(\Pi)$ is defined as

$$\mathrm{Sp}(\Pi) = \Pi(x_q) - (n-1)^{-1} \sum_{j=1, j \neq q}^{n} \Pi(x_j)$$

where x_q is the element in X with the largest membership grade. It can easily be seen that $\mathrm{Sp}(\Pi) \in [0, 1]$.

Restricting ourselves to the case of a normal possibility distribution, at least one element has possibility one. If we let x_1 have possibility one, then

$$\mathrm{Sp}(\Pi) = 1 - (n-1)^{-1} \sum_{j=2}^{n} \Pi(x_j)$$

Thus in this case where x_1 is fully possible, any increase in possibility associated with the other elements results in a decrease in $\mathrm{Sp}(\Pi)$, a decrease in certainty. This is, again, in accord with our intuition of the measure of specificity as a measure of certainty, for by increasing the possibility of some element we have increased our uncertainty. We can observe that if Π and Π' are two normal possibility distributions such that $\Pi(x_j) \leq \Pi'(x_j)$ for all x_j, then $\mathrm{Sp}(\Pi) \geq \mathrm{Sp}(\Pi')$.

Now, assume additionally that we have found supplementary data that can also be represented in the form of a possibility distribution Π_s on X associated with the value of $V = x$. Our goal is now to fuse these two pieces of information to obtain information specifically about the variable V. Our fused information should induce a constraint on the possible values V can assume; therefore we want our fused information to also be in terms of a possibility distribution over V.

The secondary information should be considered only when the primary information is not "good enough," that is, not of sufficient quality, conflicting, not credible or too imprecise. This sets up a kind of priority between the two types of knowledge. This observed relationship between the two types of information can be captured with the following two simple rules:

R1: If the quality of the primary information is "good" then use it.
R2: If the quality of the primary information is not "good" then use the secondary information.

With the aid of fuzzy systems modeling methodology [51], this knowledge about when to use the two types of information can be used to obtain a formal fusion rule. In order to use the fuzzy systems modeling method we must express our knowledge base, R1 and R2, in the form of fuzzy if-then rules. Let the variable Q stand for the quality of the primary information and use the specificity of the distribution, $\mathrm{Sp}(\Pi)$ to measure it. Since the specificity of a possibility distribution takes its value in the unit interval then the linguistic term "Good" can be represented as a fuzzy subset on the unit interval. Many choices exist for the selection of G satisfying usual conditions. Generally the actual selection of G will be subjective and context dependent. Each selection of G will result in a different instantiation of the fusion rule. One very natural and neutral choice for G we will use for illustration in the following is a simple linear form, $G(r) = r$.

Next, consider the consequent portion of the rules. Let Π_f indicate the possibility distribution resulting from our fusion process. Then our fusion model is expressible as the two fuzzy if-then rules:

if Q is G then $\Pi_f(x) = \Pi(x)$,

If Q is G' then $\Pi_f(x) = \Pi_s(x)$.

This model is an example of a TSK type fuzzy model [54], and using the reasoning mechanism of fuzzy systems modeling we can obtain an analytic formulation for $\Pi_f(x)$:

$$\Pi_f(x) = G(\mathrm{Sp}(\Pi))\Pi(x) + (1 - G(\mathrm{Sp}(\Pi))\Pi_s(x))$$

If we choose as our definition "Good information," $G(r) = r$ we finally get the following fusion rule we have been able to easily use:

$$\Pi_f(x) = (\mathrm{Sp}(\Pi))\Pi(x) + (1 - (\mathrm{Sp}(\Pi))\Pi_s(x))$$

4.2 Agent classes

The basic agent framework design incorporates the use of four distinct classes of agents. These are the updating agent (UA), the integration agent (IA), the conflation agent (CA), and the manager agent (MA). The actual number of instances of agents of each class is not limited, and determination of optimal numbers of each for a given system configuration should be derived as part of the performance enhancements after implementation. The various agents work both in isolation and in cooperative efforts with other agents. Members of the UA class are responsible for actually traversing the network in search for geospatial data relevant to given constraints. These agents are comparable to the "databots/infobots" employed by web-based search engines. Much work has been done in the realm of configuring agents to selectively filter vast amounts of information based on selected parameters, and many existing implementations perform adequately for this purpose [27]. The design of the UA class essentially consists of refining general-purpose databots to selectively seek out geospatial information based on, for example, domain names or keywords. A close example of an existing system is the GeoSpecific Search Engine at http://search.geocomm.com/. An increasing level of intelligence can be added to the UAs, such that, initially, a static list of sites is provided (as for the previously identified search engine); then, as the agents learn about geospatial data properties,

they are "turned loose" to identify potential sites on their own.

After the UAs have identified and retrieved a new set of data, the IAs begin work. These agents are responsible for intelligently analyzing the format of the new data and integrating it into the existing object-oriented database schema. Three possibilities exist for the integration procedure: (1) the format of the data exactly matches the qualifications of an existing class in the database schema, (2) the data is in a format that can easily be converted to that of an existing class, or (3) the data is in a new, currently incompatible format. For the first two possibilities, the data is fairly easily integrated into the existing database by using the current schema. The third possibility represents a situation that must be more thoroughly analyzed. In this case, the IA must decide whether to expand the database schema to include the new format, or to reject the data. The schema can be automatically (programmatically) expanded by the IA if, for example, the new data is perceived to be of tremendous value (perhaps the only-available data for a particular area of interest (AOI), or if the IA anticipates (perhaps after conferring with the UA) that a substantial amount of data of the same format is expected to be included in the future. This can be particularly difficult for spatial data involving, for example, fuzzy boundaries in which the class matching is not exact. We are currently classifying spatial data uncertainties to be able to extend the schema integration agents.

The CAs incorporate a body of knowledge pertaining to conflation, including information about data quality parameters and topological and geometric analysis capabilities such as those described in [7]. Invocation of CAs can take place in two ways. The first occurs when a user requests data from the database for a particular area-of-interest (AOI). The query manager may request the CAs to identify any possible matching features and to resolve the conflicts before presenting query results to the user. In this process, we adapt the previously developed fuzzy matching approach to the agent environment. Another possibility is that the user requests updated information for an AOI. In this case, the UAs must be dispatched and conflation performed "on the fly."

The following scenario illustrates this incidence of intelligent updating and the resulting conflation process. This type of updating is performed when a determination has been made that better (possibly meaning more current, or more accurate) data is required for a given AOI. First, multiple intelligent mobile agents are simultaneously dispatched to search through all of the potential source maps to identify and collect data on candidate matching features. Each agent is tailored to focus on a single feature category, enabling a team of agents, one for each feature type, to simultaneously analyze an individual AOI for a single data source. Moreover, all potential data sources may be analyzed by agent teams in parallel. Identified features may be collected by the agents and returned to the originating server, or optionally, the agents may remain on the remote server and communicate with the originating process via remote messaging.

Using the collected information, the agents coordinate through a single conflation manager agent developed for the specific role of feature resolution, and containing fuzzy components introduced in the preceding section. The manager agent may be resident at one of the gateway nodes, or may reside on an independent central node, possibly even a client machine. This process may extend over multiple iterations until a decision is obtained. Possible outcomes include the following: data of sufficient quality was obtained to justify the continuation of the conflation process, a decision to abort the conflation due to lack of quality data, or the necessity of human intervention due to a neutral decision. It is our goal to keep the latter to a minimum, if not eliminate this outcome entirely. In any event, the result is the generation of map data equal to or superior to any of the original map sources. Figure 2 shows our three-layer approach to system design, while Fig. 3 shows a breakdown of the mediation layer, in which our work is primarily concentrated.

Intelligent mobile agent technology enables the possibility of seamlessly extending current conflation technologies to include distributed parallel processing. This not only efficiently utilizes existing resources, but also provides for autonomous, perhaps offline, operation.

To summarize the design thus far, the system contains:

updating agents that can be trained to locate reliable, network-resident geospatial data sources,
integration agents that can intelligently analyze and compare geospatial data formats and programmatically extend the object-oriented database schema to include new formats,
conflation agents that can detect multiple feature representations and implement conflict resolution strategies, and,
one or more *manager agents* that coordinate and facilitate collaborative efforts among the working agents.

The design for integration agents is based on previously-discussed work in database schema integration, and the conflation agent design draws largely from the rule-based approach developed in [7].

4.3 Architecture

Issues of communication are always of great concern for distributed systems, and even more so in those involving mobile agents. Our architecture employs a centralized, or master database in which: (1) update requests are generated through a priority queuing scheme; (2) data analysis and conflation agent teams are released when necessary; and, (3) data changes are collected. The master database coordinates its update activities through distributed data repositories. Each repository contains a resident object assembly agent for creating a CORBA object from local database information, and a "spy" agent that logs relevant database changes to a data buffer for sending to the master database for analysis.

The master database furthermore contains region-of-interest (ROI) agents that monitor changes to their assigned geographical data regions. These regions can be prioritized for the purpose of acquiring updated information. For example, in the case of a catastrophic oil spill, a ROI agent for the area in the immediate vicinity of the

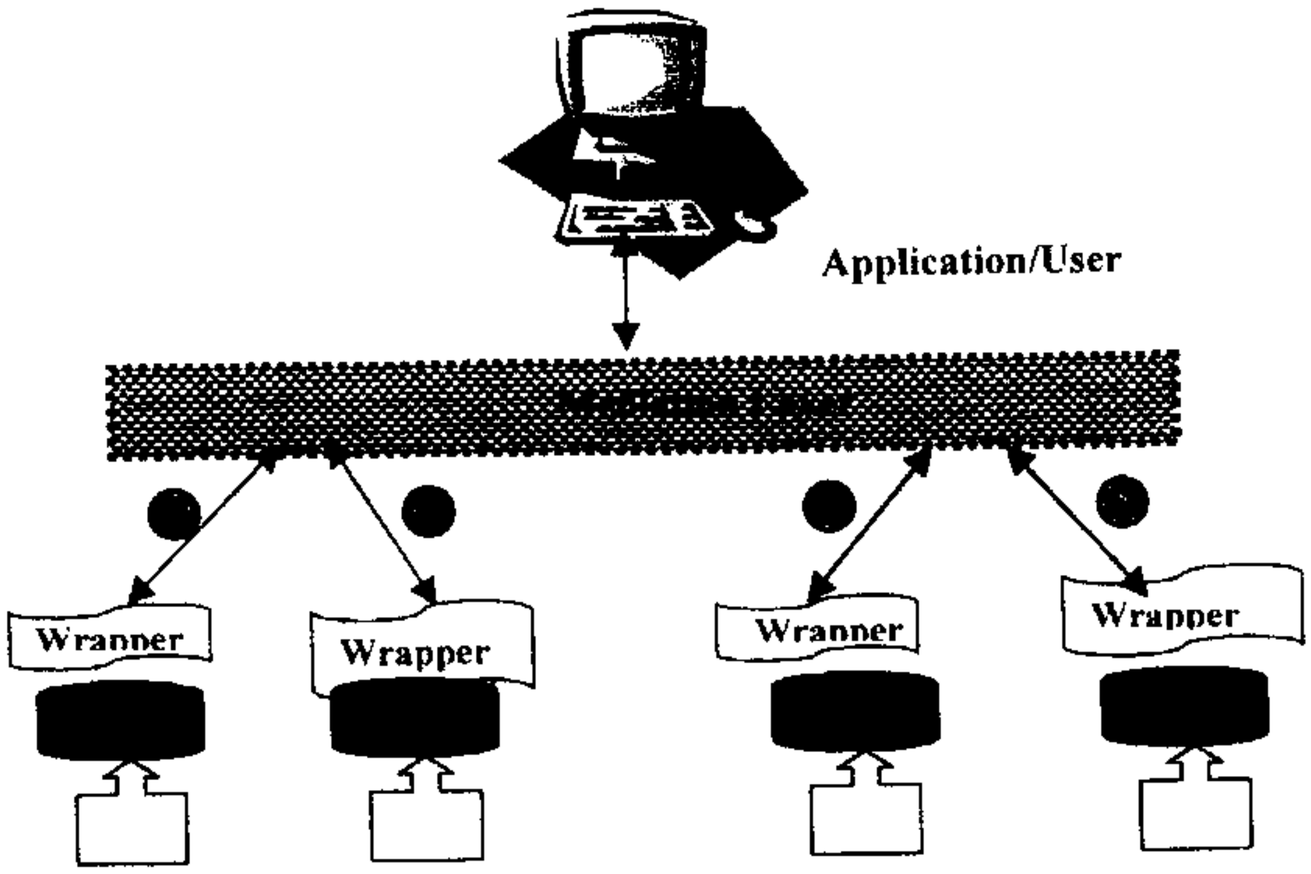

Fig. 2. Three-layer system design

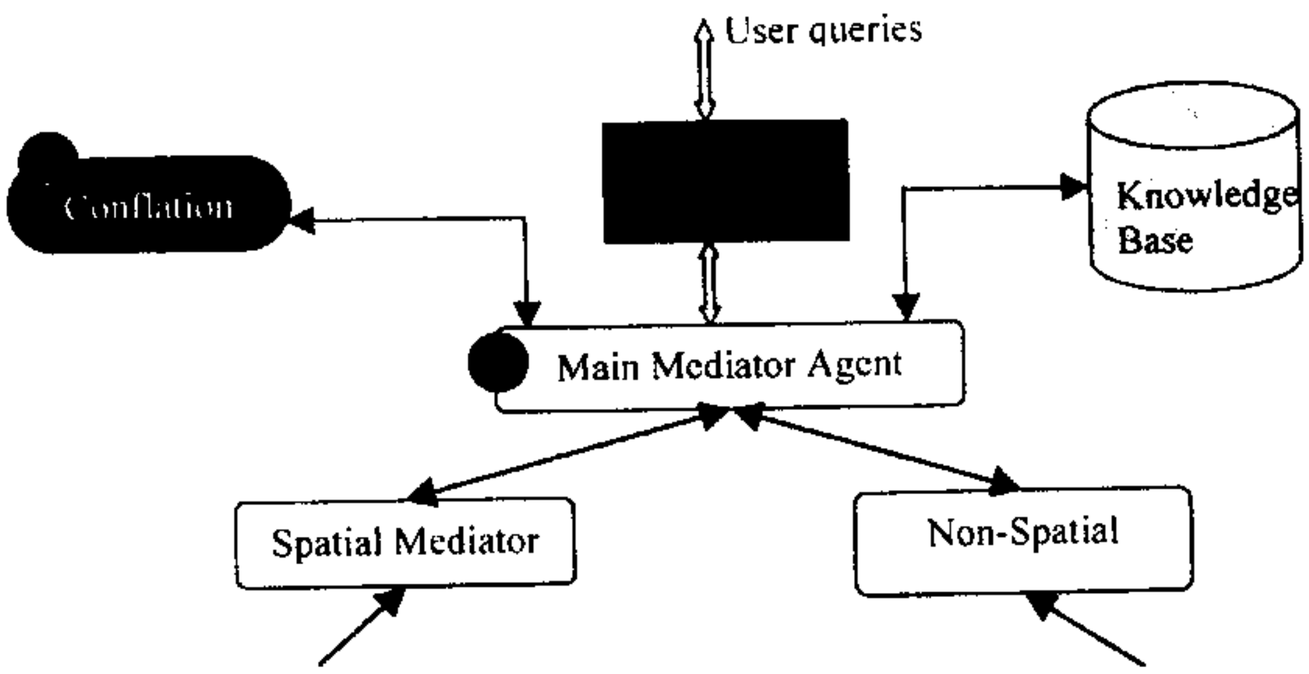

Fig. 3. Mediation layer components

spill would be given a high priority such that information is updated, e.g., every hour. A larger region surrounding the spill might be given a slightly lower priority, such that updates are made every twenty-four hours, and so forth. The resident "spy" agents are responsible for sending logged changes to the master database when the buffer reaches capacity, if no priority requests are received before that time.

Conflation is orchestrated in a distributed manner, with the process being initiated through a request from the master database. When agents arrive at a distributed data site, pertinent queries are run and results posted to a local "bulletin board." Local conflation agents are then instantiated, which send requests as appropriate to object-creating and query agents. The final conflation results are sent back to the master database.

5 Summary and future work

In this paper, we have presented a framework involving the use of mobile agents with fuzzy matching and integration capabilities for the basis of a large-scale environmental system. The complex environmental planning and analysis requirements for such systems today involve collection and integration of geographical data from multiple, diverse sources. Although the use of fuzzy logic is currently concentrated in the conflation agents and the integration agents, there is certainly room for expansion in the use of these techniques to other aspects of the system, particularly the region-of-interest agents, to better enable them to make priority updating decisions. This methodology based upon cooperative intelligent agents is proposed as the future direction for autonomous, integrative environmental databases.

We are currently implementing a prototype of the described system framework. For testing purposes, a refinement to the generalized agent framework presented in this paper will be implemented on a single server, integrated with the Naval Research Laboratory's Geospatial Integrated Database (GIDB) [8]. Initially, data sources will be explicitly provided, and performance of the agents on these test data sets will be monitored, and improvements (re-design) made in response to observed performance. Parameters targeted for collection include existence/

location of data bottlenecks, optimal number of agents for a particular hardware/software configuration, and measures of resulting data currency.

References

1. **Bordoga G, Kraft D, Pasi G** (1999) Fuzzy set techniques in information retrieval. In: Fuzzy Sets in Approximate Reasoning and Information Systems, Bezdek J, Dubois D, Prade H (eds), Kluwer, Boston MA, pp. 469–510
2. **Bogdoga G, Pasi G** (1993) Fuzzy linguistic approach generalizing boolean information retrieval, J Am Soc Inf Sci **44**: 70–82
3. **Burrough P** (1989) Fuzzy mathematical models for soil survey and land evaluation, J Soil Sci **40**: 477–492
4. **Burrough P, Frank A (eds)** (1996) Geographic Objects with Indeterminate Boundaries, GISDATA Series Vol. 2, Taylor and Francis, London, UK
5. **Burrough P, MacMillian R, Van Dursen W** (1992) Fuzzy classification methods for determining land suitability from soil profile observation and topography, J Soil Sci **43**: 193–210
6. **Burrough P, van Gaans P, MacMillian R** (2000) High-resolution landform classification using fuzzy *k*-means, Fuzzy Sets and Systems **113**: 37–52
7. **Cobb M, Chung M, Miller V, Foley H, III, Petry F, Shaw K** (1998) A rule-based approach for the conflation of attributed vector data, GeoInformatica **2**(1): 7–35
8. **Cobb M, Foley H, III, Wilson R, Chung M, Shaw K** (1998) An OO database migrates to the web, IEEE Software **15**(3): 22–30
9. **Cobb M, Petry F** (1998) Modeling spatial data within a fuzzy framework, J Am Soc Infor Sci **49**(3): 253–266
10. **Cobb M, Petry F, Robinson V** (eds.) (2000) Special issue: Uncertainty in geographical information systems and spatial data, Fuzzy Sets and Systems **113**(1)
11. **Cross V, Firat A** (2000) Fuzzy objects for geographical information systems, Fuzzy Sets and Systems **113**(1): 19–36
12. **Dale M** (1988) Some fuzzy approaches to phytosociology: ideals and instances, Folia Geobot. Phytotaxon **23**: 239–274
13. **Defense Mapping Agency** (1993) Military Standard: Vector Product Format, MIL-STD-2407, Defense Mapping Agency, Fairfax, VA
14. **Egenhofer M, Fonseca F** (1999) Ontology-driven geographic information systems, 7th ACM Symposium on Advances in GIS, Kansas City, MO, pp. 14–19
15. **Fisher P, Pathirana S** (1990) Evaluation of fuzzy membership of land cover classes in the suburban zone, Remote Sensing of Environment **34**: 121–132
16. **Flowerdew R** (1991) Spatial data integration. In: Geographical Information Systems: Principles and Applications, Maguire DJ, Goodchild MF, Rhind DW (eds), Longman Scientific and Technical, Great Britain, Vol. 1, pp. 375–387
17. **Foley H, III, Petry F, Cobb M, Shaw K** (1997) Utilization of an expert system for the analysis of semantic characteristics for improved conflation in geographic information systems, Proceedings of the 10th International Conference On Industrial and Engineering Applications of AI, Atlanta, GA, pp. 267–275
18. **Foley H, III, Petry F, Cobb M, Shaw K** (1997) Using semantic constraints for improved conflation in spatial databases, Proceedings of the 7th International Fuzzy Systems Association World Congress, Prague, pp. 193–197
19. **Foody G** (1992) A Fuzzy sets approach to representation of vegetation continua from remotely sensed data, Photogrammetric Eng and Remote Sensing **58**: 221–225
20. **Gale S** (1972) Inexactness, fuzzy sets and the foundation of behavioral geography, Geograph Anal **4**: 337–349
21. **George R, Buckles B, Petry F, Yazici A** (1992) Uncertainty modeling in object-oriented geographical information systems, 1992 Proc. Conf Database & Expert System App., Seville, Spain, pp. 77–86
22. **Glass G** (1999) Objectspace Voyager 3.0 Overview, Objectspace, Inc. http://www.objectspace.com/voyager
23. **Goodchild M** (2000) Uncertainty in geographic information systems, Fuzzy Sets and Systems **113**: 3–5
24. **Guesgen H, Albrecht J** (2000) Imprecise reasoning in geographic information systems, Fuzzy Sets and Systems **113**(1): 121–131
25. **Kotz D et al.** (1997) Agent TCL: targeting the needs of mobile computing, IEEE Internet Computing **1**(4): 58–67
26. **Lagacherie P, Andrieux P, Bouzigues R** (1996) Fuzziness and uncertainty of soil boundaries: from reality to coding in GIS. In: Geographic Objects with Indeterminate Boundaries, Burrough P, Frank A (eds), Taylor and Francis, London, UK, pp. 275–286
27. **Lawton G** (1999) Putting agents to work, Knowledge Management
28. **Leung Y** (1979) Locational choice: a fuzzy set approach, Geograph Bull **19**: 28–34
29. **Liu Z-Q, Satur R** (1999) Contextual fuzzy cognitive map for decision support in geographic information systems, IEEE Trans Fuzzy Systems **7**: 495–507
30. **Mackay DS, Robinson V** (2000) A multiple criteria decision support system for testing integrated environmental models, Fuzzy Sets and Systems **113**: 53–67
31. **Mackay DS, Robinson V** (1998) Model self-evaluation and detection of semantic error in a spatially explicit ecosystem process model, Proc IPMU, Paris, pp. 588–596
32. **McBratney A, De Gruijter J** (1992) A continuum approach to soil classification by modified fuzzy *k*-means with extragrades, J Soil Sci **43**: 159–175
33. **Morris A, Petry F** (1998) Design of fuzzy querying in object-oriented spatial data and GIS, Proc NAFIPS 98, Pensacola, FL, pp. 211–215
34. **Morris A, Petry F, Cobb M** (1998) Fuzzy object-oriented database modeling of spatial data, Proc IPMU Conference, Paris, pp. 604–611
35. **Pipkin J** (1978) Fuzzy sets and spatial choice, Ann Assoc Amer Geograph **68**: 196–204
36. **Risbey J, Kandlikar M, Patwardhan A** (1996) Assessing integrated assessments, Climate Change **34**: 369–395
37. **Roberts D** (1989) Fuzzy systems vegetation theory, Vegetatio **83**: 71–80
38. **Robinson V** (1988) Implications of fuzzy set theory for geographic databases, Comp Environ Urban Systems **12**: 89–98
39. **Robinson V** (1990) Interactive machine acquisition of a fuzzy spatial relation, Comp Geosci **6**: 857–872
40. **Robinson V, Frank A** (1985) About different kinds of uncertainty in geographic information systems, Proc AUTOCARTO 7 Conference
41. **Saalfeld A** (1988) Conflation: automated map compilation, Int J GIS **2**(3): 217–228
42. **Saint-Joan D, Mezzadri-Centeno T** (1998) GEODES: a fuzzy expert system for spatial decision support with GIS, Proc IPMU, Paris, pp. 612–619
43. **Sarjakoski T** (1996) How many lakes, islands, and rivers are there in finland? a case study of fuzziness in the extent and identity of geographic objects. In: Geographic Objects with Indeterminate Boundaries, Burrough P, Frank A (eds), Taylor & Francis, London, UK, pp. 299–312
44. **Shaw K, Chung MJ, Wilson R, Ladner R, Cobb MA, Lovitt T** (1999) An object-oriented database approach for urban warfare, Proceedings of Urban and Regional Information Systems Association, URISA '99, Chicago, IL, pp. 272–283
45. **Shepherd IDH** (1991) Information integration and GIS. In: Geographical Information Systems: Principles and Applications, Maguire DJ, Goodchild MF, Rhind DW (eds), Longman Scientific and Technical, Great Britain, Vol. 1, pp. 337–360
46. **Stoms D** (1987) Reasoning with uncertainty in intelligent geographic information systems, Proc GIS 87 – 2nd Annual

Int Conf On Geographic Information Systems, Amer Soc for Photogrammetry and Remote Sensing, Falls Church, VA, pp. 693–699
47. **Usery EL** (1996) A conceptual framework and fuzzy set implementation for geographic feature. In: Geographic Objects with Indeterminate Boundaries, Burrough P, Frank A (eds), Taylor and Francis, London, UK, pp. 71–85
48. **van Asselt M, Rotmans J, den Elzen M, Hinderink H** (1997) Uncertainty in integrated assessment: a cultural perspective-based approach, GLOBO Report Series #9, RIVM, Bilthoven, The Netherlands
49. **Wang F** (2000) A fuzzy grammar and possibility theory-based natural language user interface for spatial queries, Fuzzy Sets and Systems 113(1): 147–159
50. **Yager R** (1992) On the specificity of a possibility distribution, Fuzzy Sets and Systems 50: 279–292
51. **Yager R, Filev D** (1994) Essentials of Fuzzy Modeling and Control, J Wiley, New York
52. **Yager R, Petry F** (2001) Constructing intelligent fusion procedures using fuzzy systems modeling, Int J Intelligent Systems (To appear)
53. **Yazici A, Petry F, Pendergraft C** (2001) Fuzzy modeling approach for integrated assessments using cultural theory, Tk J Electrical Eng Comp Sci 9(1): 31–42
54. **Yen J, Lenagri R** (1999) Fuzzy Logic, Prentice-Hall, New York
55. **Zadeh L** (1979) A theory of approximate reasoning. In: Machine Intelligence, Hayes J, Michie D (eds), Vol. 9, Halstead Press, New York, pp. 149–194
56. **Zadeh L** (1978) Fuzzy logic = computing with words, Fuzzy Sets and Systems, 1: 3–28