12-9-2006

Semantics-Enabled Framework for Knowledge Discovery from Earth Observation Data

Surya Srinivas Durbha

SEMANTICS-ENABLED FRAMEWORK FOR KNOWLEDGE DISCOVERY FROM EARTH OBSERVATION DATA

By

Surya Srinivas Durbha

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Computer Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

December 2006

SEMANTICS-ENABLED FRAMEWORK FOR KNOWLEDGE DISCOVERY FROM EARTH OBSERVATION DATA

By

Surya Srinivas Durbha

Approved:

Roger L. King
William Giles Distinguished Professor
Associate Dean of College of the
Bagley College of Engineering
(Director of Dissertation)

Nicolas H. Younan
Professor of Electrical and Computer
Engineering (Committee Member)
(Graduate Coordinator)

David Shaw
William Giles Distinguished Professor of
Plant and Soil Science
(Committee Member)

Ioana Banicescu
Professor of Computer Science and
Engineering
(Committee Member)

Kirk Schulz
Dean of Bagley College of
Engineering

Name: Surya Srinivas Durbha

Date of Degree: December 8, 2006

Institution: Mississippi State University

Major Field: Computer Engineering

Major Professor: Dr. Roger L. King

Title of Study: SEMANTICS-ENABLED FRAMEWORK FOR KNOWLEDGE DISCOVERY FROM O EARTH OBSERVATION DATA

Pages in Study: 131

Candidate for Degree of Doctor of Philosophy

Earth observation data has increased significantly over the last decades with satellites collecting and transmitting to Earth receiving stations in excess of three terabytes of data a day. This data acquisition rate is a major challenge to the existing data exploitation and dissemination approaches. The lack of content and semantics based interactive information searching and retrieval capabilities from the image archives is an impediment to the use of the data. The proposed framework (Intelligent Interactive Image Knowledge retrieval-I3KR) is built around a concept-based model using domain dependant ontologies. An unsupervised segmentation algorithm is employed to extract homogeneous regions and calculate primitive descriptors for each region. An unsupervised classification by means of a Kernel Principal Components Analysis (KPCA) method is then performed, which extracts components of features that are

nonlinearly related to the input variables, followed by a Support Vector Machine (SVM) classification to generate models for the object classes.

The assignment of the concepts in the ontology to the objects is achieved by a Description Logics (DL) based inference mechanism. This research also proposes new methodologies for domain-specific rapid image information mining (RIIM) modules for disaster response activities.

In addition, several organizations/individuals are involved in the analysis of Earth observation data. Often the results of this analysis are presented as derivative products in various classification systems (e.g. land use/land cover, soils, hydrology, wetlands, etc.). The generated thematic data sets are highly heterogeneous in syntax, structure and semantics. The second framework developed as a part of this research (Semantics-Enabled Thematic data Integration (SETI)) focuses on identifying and resolving semantic conflicts such as confounding conflicts, scaling and units conflicts, and naming conflicts between data in different classification schemes. The shared ontology approach presented in this work facilitates the reclassification of information items from one information source into the application ontology of another source. Reasoning on the system is performed through a DL reasoner that allows classification of data from one context to another by equality and subsumption. This enables the proposed system to provide enhanced knowledge discovery, query processing, and searching in way that is not possible with key word based searches.

DEDICATION

I would like to dedicate this research to my mother Dr. K. Pushpa, father Dr. D. L. Prasad Rao, for their relentless encouragement to pursue academic excellence and my sister Dr. D. Sita Lakshmi for showing me in innumerable subtle ways the joys of learning.

ACKNOWLEDGMENTS

This dissertation could not have been written without Dr. Roger L. King who not only served as my supervisor but also encouraged and challenged me throughout my academic program. I would like to express my deepest gratitude to him for mentoring me and providing ample academic freedom which helped to let the creative juices flow. As a member of his research team, I have enjoyed the opportunity to work on exciting research projects and continually be inspired by his vision. Also his intellectual humor has helped many a times to keep up my spirits and make this journey pleasant.

I would also like to thank my committee members Dr. Nicolas H. Younan, Dr. David Shaw, and Dr. Ioana Banicescu for serving on my committee and providing valuable feedback and advice.

On this path I have been motivated by many individuals directly or indirectly; my friend Dr. Gautam Cherla consistently provoked and pushed many of my boundaries, though disagreeable on numerous occasions, enabled to explore new horizons. Sridhar Katragadda encouraged me to accelerate my research work and also introduced me to the world of symphony. Satish Chalasani and Thomas George provided me early support in settling down at MSU.

Finally I would also like to acknowledge my deep gratitude to NSF, NASA and NOAA for providing financial support throughout the research and Dr. David Shaw for providing several opportunities within GRI.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1.1 Background

In recent years, U.S. Government Earth remote sensing data collection and archiving has increased significantly. Landsat data alone comprises 434 terabytes of archive (31 years of Landsat 1-5; 165 terabytes, 4 years of Landsat 7; 269 terabytes). Multiple Petabytes of data from Earth Observing Satellite EOS and Pre-EOS are archived by NASA Distributed Active Archive Centers (DAACs).

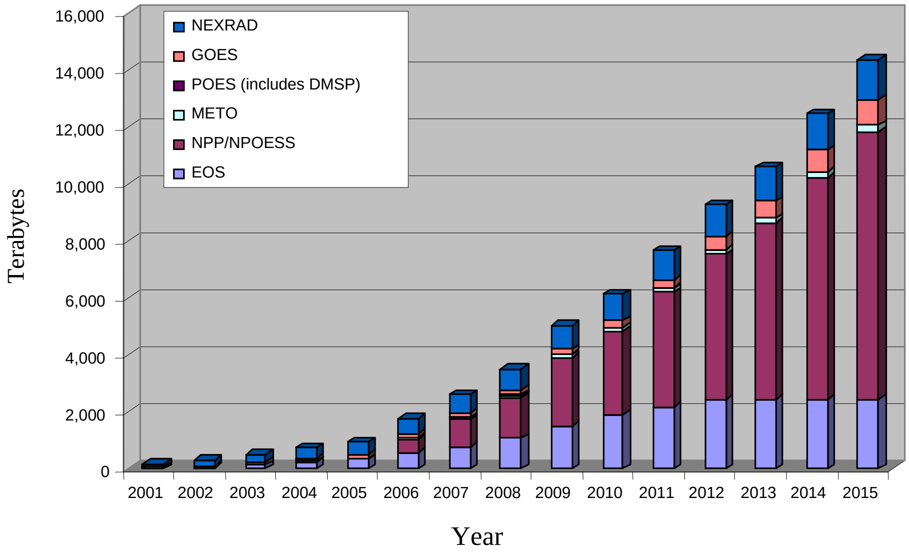


Figure 1.1 Cumulative archive growth of major EO sensor systems

Figure 1.1 depicts the projected cumulative archive growth by data collected from major EO sensor systems over a 15-year period. Since 1998, the science data volume managed by the EOS Data and Information System (EOSDIS) has increased eight-fold, and continues to grow at a rate of over 2 Terabytes per day. The United States Geological Survey (USGS) active archive has increased dramatically to 450 terabytes over the past four years (Figure 1.2). National Oceanic and Atmospheric Administration (NOAA) has archived data at the National Climatic Data Center, National Geophysical Data Center, National Oceanographic Data Center, and the National Coastal Development Data Center. In addition, large amounts of *in situ* data (e.g., AmeriFlux [1], Fluxnet [2]) are collected and archived for guiding, collecting, synthesizing, and disseminating long-term measurements of CO_2, water, and energy exchange from varied ecosystems. Any future efforts to manage carbon sequestration of atmospheric CO_2 in terrestrial or marine systems will also require observations and models to verify changes in stocks.

In the ocean observations domain, a variety of *in situ* sensors and sampling methods are used to collect data (e.g., meteorological, oceanographic, biogeochemical) and assimilate it into the Integrated Ocean Observation System (IOOS). IOOS is envisioned as a coordinated national and international network of observations, data management, and analyses systems that rapidly and systematically acquires and disseminates marine environmental data and information on past, present, and future states of the oceans [3]

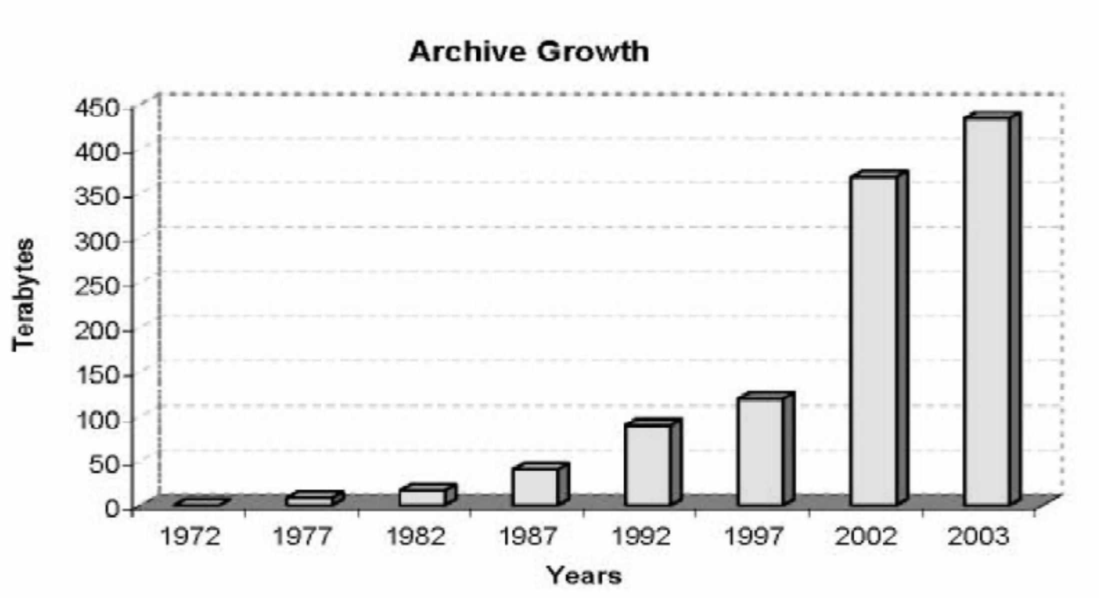


Figure 1.2 Archive growth at Earth Resources Observation Systems (EROS) Data Center

Availability of such a magnitude of data to the users has raised important challenges regarding its archiving, the ability to convert the volumes of data into meaningful information that can be used for decision-making, and dissemination of the generated information. At the end of the data-information channel are diverse groups of users with varying levels of expertise and backgrounds who need to use Earth observations to solve a variety of complex problems. However, usable information, defined as knowledge in this context, is rarely readily available and it becomes the task of the decision maker to extract the clusters of knowledge found in the data. Hence, it is imperative that the information that is generated from Earth observations is usable and relevant to a particular context of the problem-solving environment. Unfortunately, contextual information is rarely captured and percolated through the channels of the knowledge discovery process [4].

One of the primary interests in developing enabling technologies for seamless access of disparate information sources is the ever increasing collection and availability of primary thematic data, including such elements as land cover, sea floor, bottom type, habitat distribution, change detection data, etc. These geospatial datasets offer unique perspectives into the dynamic nature of the geographical phenomenon and consequently, many hydrological, ecological and climatological models use such geographically referenced information as an essential input [5][6].   To overcome the diverse nature of data and represent it in a uniform way, syntactic standardization has long been proposed and a number of metadata standards have been developed worldwide during the last decade [7], which is now widely accepted as the standardized models for both data and metadata. Each of these standards originated in one particular community and was quickly adopted in a variety of domains. An example of the metadata that is specific for the Earth observation data is the Content Standards for Digital Geospatial Meta Data (CSDGMD), developed by the Federal Geographic Data Committee (FGDC) [8]. This comprehensive standard describes nearly 300 separate elements and provides a solid basis for both geographic and environmental data. Other metadata standards to integrate geographic information include the National Spatial Data Infrastructure (NSDI), Geospatial One-stop, and the U.S. Geological Survey’s The National Map as well as standards from the International Standards Organization (ISO) [9].

Although the metadata standards alleviate to a large extent the syntactic heterogeneity of the data, a problem that is still not completely solved is heterogeneity of the intended interpretation of information. In general, the data heterogeneity problems can be divided into three categories [10]

- Syntactic heterogeneity is caused by different logical models (e.g. relational vs. object oriented) or due to different geometric representations (raster vs. vector).
- Schematic heterogeneity occurs because of different conceptual data models (e.g. objects in one database considered as properties in another, different generalization hierarchies).
- Semantic heterogeneity causes most information integration problems. It occurs because of the differences in meaning, interpretation or usage of the same or related data.

This research addresses the semantic heterogeneity in Earth observation data and proposes an enabling framework to

- Translate data into domain specific formalisms.
- Convert the data content into knowledge clusters through explicit specification of the conceptualization through Ontologies (i.e., data models).
- Link users to the knowledge, provide integrated visualization, search and query answering facilities, and to gather information at different levels of granularity, from the sub category to the specific data level.
- Dynamic learning of user defined semantic classes and related tasks, and updating the knowledge base with the newly discovered information.

1.2 Motivation

1.2.1 Why a new system?

This research is driven by the need to facilitate exploitation of huge amounts of Earth observations (EO) data available in a multitude of domains, in a way that would help users at different levels of expertise (Figure 1.3) to explore the vast knowledge hidden in the archived data. The motivating factor for developing a new system that departs significantly from the existing methods is to enable seamless access to imagery and other ancillary data not only to expert users/managers but, also to normal users.

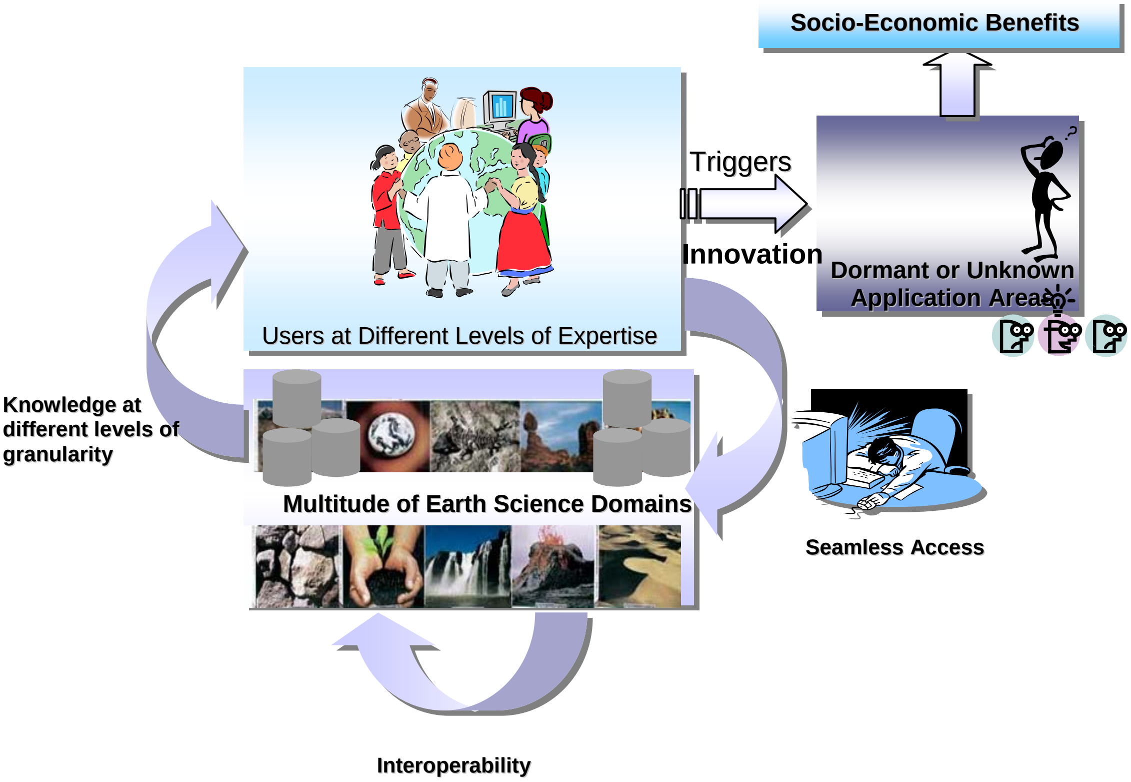

Socio-Economic Benefits
Triggers
Innovation
Dormant or Unknown Application Areas
Users at Different Levels of Expertise
Knowledge at different levels of granularity
Multitude of Earth Science Domains
Seamless Access
Interoperability

Figure 1.3 Need to provide knowledge at different levels of granularity

This would then trigger the use of data in a variety of domains and applications that normally would be unknown or lie dormant due to the lack of proper dissemination of the unprecedented information provided by the current EO systems.

The following sections describe some critical areas that are not currently addressed by available systems.

1.2.2 Data Processing and Dissemination

- Making data available timely (required for emergency response tasks).
- Better categorization / aggregation of content and formulation of custom products and provision of subsetting tools at spatial and temporal levels.
- Support interoperability between various formats of data within an archive and between them.
- Package products based on the meaning and knowledge about the measurements and context of the information sources.
- Creation of machine understandable semantic metadata so that intelligent search engines / agents can automatically process and index the content.
- Dissemination of information through current standards (e.g. OGC Web Map service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Web Catalog Service (CS-W) etc.) driven web services oriented architectures and extended to the semantic web services vision.

1.2.3 User Interfaces

- Pursue semantic web technologies to develop content and semantics driven interactive user interfaces which supports advanced querying that go beyond just keyword based searches.

- Reduce the depth of hypertext linkages to reach a particular goal (i.e. intelligent presentation of content).
- Learning user preferences and providing semi-automated help to fulfill a query requirement based on previous knowledge.

Archived remote sensing imagery is not amenable to automated methods of query and knowledge discovery. At present, information about an image is limited to queries on structural metadata resulting in geographical coordinates, time of acquisition, sensor type, and acquisition mode [11].

Such a limitation in automated exploitation of imagery has placed a severe constraint on the usability of the data by operational users. To overcome this limitation and increase useful exploitation of the data, it is necessary to adopt new technologies that allow the accessibility of remote sensing data based on content and semantics.

Consider a query "*Retrieve all images from sensor X which contains wetlands near a coast in the Eastern part of country Y*". This query requires problem specific discovery of knowledge that is responsive to the needs of an analytical task. Therefore, the need for knowledge discovery (features, complex relationships, and hypotheses that describe potentially interesting regularities) from large heterogeneous networks of observations and information products generated from modeling efforts is important for Earth observation (EO) decision making.

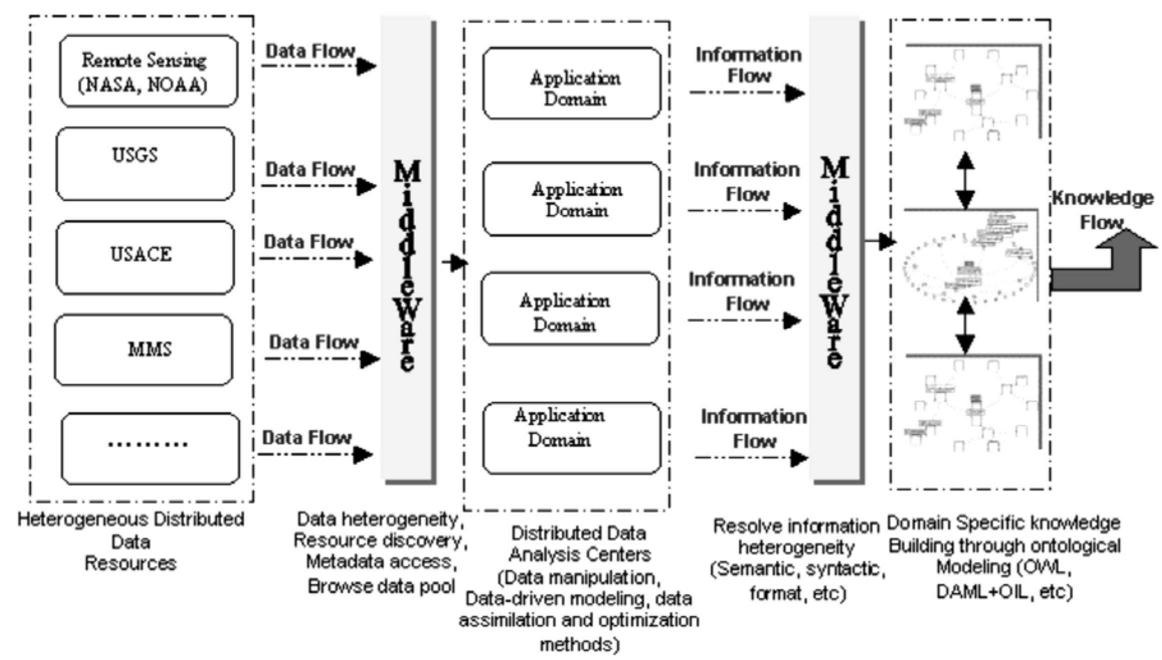

Remote Sensing (NASA, NOAA)
USGS
USACE
MMS
....
Data Flow
Data Flow
Data Flow
Data Flow
Data Flow
MiddleWare
Application Domain
Application Domain
Application Domain
Application Domain
Information Flow
Information Flow
Information Flow
Information Flow
MiddleWare
Knowledge Flow
Heterogeneous Distributed Data Resources
Data heterogeneity, Resource discovery, Metadata access, Browse data pool
Distributed Data Analysis Centers (Data manipulation, Data-driven modeling, data assimilation and optimization methods)
Resolve information heterogeneity (Semantic, syntactic, format, etc)
Domain Specific knowledge Building through ontological Modeling (OWL, DAML+OIL, etc)

Figure 1.4 A process diagram for transforming distributed data resources into knowledge

However, before knowledge can be discovered and shared it has to be formalized in such a manner that it is machine accessible and understandable. Task or context-specific analysis of data requires exploiting the relations between terms used to specify the data, to extract the relevant information, and integrate the results in a coherent form. Figure 1.4 describes the data from various sources (NASA, NOAA, in situ, etc.) that are transformed into information at different application domain data analysis centers. However, to achieve this, middleware is required that provides tools to browse and access the data resources for resolving the heterogeneity problems. Domain specific knowledge building is achieved through ontological modeling that provides functionalities for capturing knowledge.

Image information mining provides advanced support in tasks where the complexity of the regions in the image has intricate shapes and textures. The knowledge about such details could be encoded in predictive models that have varying levels of granularity. These could then be used in real time for exploring the data and generating regions of interest.

1.2.4 Resolving Heterogeneities in Disparate Earth Observation Thematic Data

The semantic heterogeneity problem in EO data, where there is usually no possibility for human interpretation and intervention within a service chain, in such a scenario, formalizing the semantics of geographic information has become indispensable. This proposed framework is described through a motivating example of tackling the problem of semantic heterogeneity in the thematic information sources related to land use and land cover. Explicit semantic description of the contents of the data is required to understand the context. However, the description of data in terms of its semantics which fully describes the data products is a very challenging task and requires formulation of the information sources in ways that would help in automated processing or machine understandability.

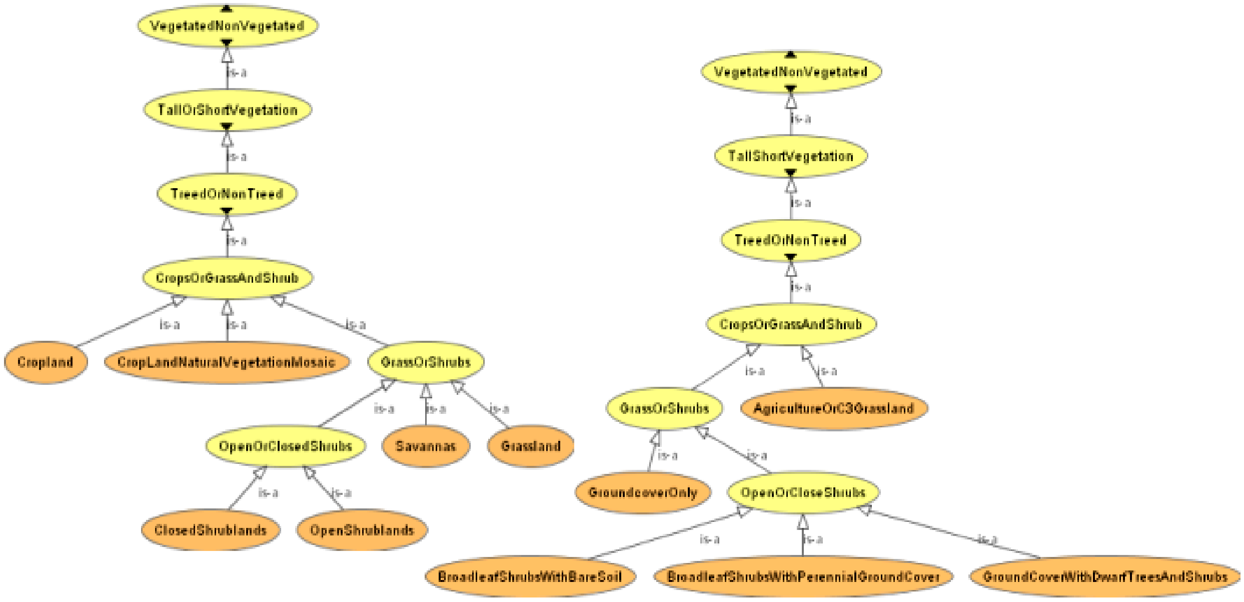


Figure 1.5 Semantic conflicts between classification systems (IGBP and SiB)

The three main causes of semantic heterogeneity are [12]:

- *Confounding conflicts* occur when information items seem to have the same meaning, but differ in reality, e.g. due to different temporal contexts
- *Scaling and Units conflicts* occur when different reference systems are used to measure a value, (e.g. currencies)
- *Naming conflicts* occur when the naming schemes of the information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

There exist several land cover characterization schemes such as the IGBP, USGS, and OGE etc. Each of these classification systems differs in their purpose and level of aggregation. Also the classified data could be available in multiple scales, i.e., a coarse scale or a finer scale. Thus, semantic conflicts arise when data from such sources are used in an overall decision-making scenario. Therefore, it is necessary to adopt technologies that help to overcome the semantic translation problems.

1.3 Contributions of this Research

This research proposes two frameworks each of which provides unique methodologies for semantics-enabled data retrieval and integration. The focus areas are:

- Framework for semantics-enabled knowledge retrieval from remote sensing data archives.
- Framework for semantic reconciliation of disparate Earth observation thematic data.

1.3.1 Framework for Semantics- Enabled Knowledge Retrieval from Remote Sensing Digital Data Archives

This research provides a basis for the content and semantics-based retrieval of knowledge from Earth observation data archives. The proposed system (Intelligent Interactive Image Knowledge Retrieval - I^{3}KR) is built around a concept-based model using domain-dependent Ontologies. The following are the contributions of this research:

1.3.1.1 Architecture

Development of an architecture where the basic concepts of the domain are identified first and generalized later, depending upon the level of reasoning required for executing a particular query. The proposed middleware facilitates the access and exploration of remote sensing digital archives through provision of ontology-based modeling of the concepts involved in the domain of interest, and linking those concepts with predictive models developed through machine learning methods for imagery from different sensors. This architecture is distinctive in the sense that it not only provides an efficient way for intuitive content-based searches, but also adapts open standards (Open Geospatial Consortium) for data dissemination.

1.3.1.2 Content and Semantics

The ontological modeling of the domain specific concepts (e.g., imagery, land cover) that is proposed, enables encoding the definitions of the concepts in a formal way and is used to acquire knowledge in a domain of interest (e.g., Coastal zone). Current systems do not formalize the domain concepts in the form of ontology. Also, I^{3}KR system localizes interesting zones and extracts characteristic information from them and stores this information in a database, which is later used for providing content–based knowledge about the semantic class(s).

1.3.1.3 Primitive Features

Identification and proposal of a unique set of primitive feature extraction algorithms corresponding to color, texture, and shape that are useful in an image information mining context.

1.3.1.4 Color Space Transformations

Extraction of primitive features based on L*a*b* color space. L*a*b* provides almost perfect separation of brightness and color information. This allows fine control over highlight and shadow; it also excels at distinguishing shades of green foliage. The a and b channels of Lab are good sources of masks for separating uniquely colored objects because they depend only on hue, and will, therefore, trace a true outline of an object in variable lighting. Hence, it provides unique advantages over the use of normal RGB color space used in existing systems.

1.3.1.5 Region-Based Approach

I^3KR is a region-based system that departs significantly from the few existing image mining systems used with Earth remote sensing imagery which are pixel based. A region-based retrieval system segments images into regions (objects), and retrieves images based on the similarity of the regions. Several advantages are achieved by this architecture, such as savings on computation, time, and resources.

1.3.1.6 Feature Data Structure Retrieval and Dimensionality Reduction

The large files sizes of data after feature extraction needs methods to reduce the data so that it is manageable in practice. This system addresses this in multiple ways:

- Region-based architecture provides significantly less amounts of data after feature extraction.
- Uses a more sophisticated algorithm for feature dimensionality reduction by a nonlinear Kernel Principal Component Analysis (KPCA). The prevalent systems use flat clustering methods such as K-means and variations of it for data reduction.
- The system facilitates feature interactions and also selection of an optimal set of features through a wrapper-based genetic algorithm approach. In a wrapper

approach, the feature subset selection algorithm exists as a wrapper around the induction algorithm. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as a part of the function evaluating feature subsets. Current systems do not address this issue and assume the features to be optimal. Thus, this work reduces the number of features, which in turn provides a reduction in data size.

1.3.1.7 Model Development and Concept Assignment

Support Vector Machines (SVM) based development of semantic models is used in this research for learning.

The choice of this algorithm is driven by

- SVMs consistent superior performance over other competing methods in a variety of domains.
- Can work on sparse data and significantly faster.
- Strong statistical background
- SVM exhibits inherent advantages due to their use of the structural risk minimization principle in formulating cost functions and of quadratic programming during model optimization. These advantages lead to a unique optimal and global solution compared to conventional artificial neural network models.

Current systems use a Bayesian learning approach, where the modeling is done based on the prior information available in a domain. In real life inference problems it is often impossible to elicit the actual prior knowledge.

I^{3}KR system uses the domain knowledge through Ontologies that provides not only the advantage of domain specific knowledge, but allows interoperating between different domains.

1.3.1.8 Probabilistic Outputs

Current systems provide semantic labels to the classified output, but generally do not provide a confidence value of the identified class. I^{3}KR system provides a probabilistic output that helps to:

- Provide feedback about the strength of the classified object.
- Rank the classified output with respect to their relevance to the user query.
- Combine the classification outputs for an overall decision making scenario.
- Select concepts from application ontologies.

1.3.1.9 Rapid Image Information Mining

This work provides the ability for rapid image information mining (RIIM) for disaster response and assessment in near real time scenarios. The computationally intensive tasks of feature extraction and model generation are considerably reduced by the wrapper-based approach for feature selection and generation adopted in this research. This is vital for emergency response activities. The RIIM system provides capabilities for a first assessment of the disaster situation through the querying of the actual content in the remote sensing images which is currently limited by queries only at the image metadata level.

1.3.2 Framework for Semantic Reconciliation of Disparate Earth Observation Thematic Data

The objective of this research is to provide methodologies for seamless integration of disparate thematic databases. This work proposes a framework for Semantics Enabled Thematic data Integration (SETI). The following summarizes the contributions for this part of the work.

1.3.2.1 Problem Formulation

In this research the integration problem between disparate EO Thematic data archives is formulated in terms of a semantic reconciliation problem. Current data integration frameworks in EO domain consider only the syntactic elements that address the differences in logical and conceptual data models while completely ignoring the semantic conflicts. The resolution of these conflicts allows the conversion of data into information and actionable intelligence.

1.3.2.2 Semantic Conflicts Identification

The semantic conflicts are identified in terms of confounding conflicts, scaling and units conflicts, and naming conflicts. In particular these conflicts are put in the context of resolving the heterogeneities in data in land cover classification schemes such as IGBP, USGS, SiB, OGE etc.

1.3.2.3 Application Ontologies Development

In SETI Ontologies were developed for each of the seven classification schemes (International Geosphere Biosphere Programme (IGBP) [13], United States geological

Survey (USGS) [14], Olson Global Ecosystems (OGE) [15], Simple Biosphere model (SiB) [16], Simple Biosphere model2 (SiB2) [17], Biosphere Atmosphere Transfer Scheme (BATS) [18], etc). The core attributes of the seasonal land cover were used to define object properties (e.g. hasBiome, hasBiomeCode, hasStructure, hasFoliage etc.) for each land cover class. Current systems only work at the database level without regard to the semantics, which impedes in interoperating between thematic data. Also, the search on the database is limited to key word searches or through structured query language (SQL) based queries. This research formulates this issue in a different way as a knowledge representation problem and acquires a knowledge base. The knowledge base uses an ontological approach to specify its structure (entity types and relationships) and its classification scheme. An ontology, together with a set of instances of its classes, constitutes a knowledge base that is amenable to intelligent reasoning and querying that go beyond key word based searches.

1.3.2.4 Shared Ontology Development

For the integration of the classification systems a separate, more expressive terminology is required. The semantics of this terminology may be specified by a logic-based ontology, which then is called a shared ontology or reference ontology. This research adopts the shared ontology concept and develops a reference ontology for land cover classification schemes, which is envisaged to be a meta-standard in the future.

1.3.2.5 Concept- Based Retrieval and Integration of Thematic Information

The shared ontology approach provides the necessary framework for Description Logic (DL)-based reasoning across heterogeneous information sources. SETI uses a DL reasoner that allows classification of data from one context to another by equality and subsumption. This provides an ability to search each concept in the application ontology belonging to an information source with the subsumed concepts in the shared ontology to check if they satisfy the concept definitions and then retrieves those concepts that match the semantics. This methodology of concept-based searches is not available in the current EO systems for data retrieval.

1.4 Document Overview

This dissertation is organized as follows; Chapter 2 describes the current state of knowledge in image information mining applications and also describes the various systems in the area of semantics-driven knowledge management solutions. The emphasis is on applications that could be put in the context of Earth science applications. Chapter 3 describes in depth the proposed methodologies for image information mining focusing on feature extraction, feature selection, feature generation, and classification using a variety of machine learning algorithms. Also, a procedure is described that enables the linking of model generated objects to relevant semantics in an ontology. Chapter 4 is focused on the proposed framework for semantics-enabled reconciliation of disparate thematic data; the proposed methodology is described through a motivating example of resolving semantic conflicts between various land cover classification schemes. Chapter 5 presents the results from both the developed frameworks. In particular, it describes the developed interface and the retrieval of knowledge from real world datasets comprising both raw remote sensing data and also processed information (thematic data). Chapter 6 concludes with some recommendations for future directions in semantics-driven knowledge management solutions.

CHAPTER II

LITERATURE REVIEW

2.1 Content-Based Information Retrieval (CBIR)

Content-Based Image Retrieval (CBIR) systems have mostly been developed outside the remote sensing domain, and the adoption of such systems to remote sensing image mining is challenging due to the unique content present in imagery. New methods for analysis must be developed for Earth remote sensing data. Typical features based on color, texture, shape, region, and appearance have different interpretations in remote sensing imagery as opposed to images in other domains (e.g., photo catalogs). Typical features are color, texture, shape, region, and appearance [19]. Some of the CBIR systems include IBM QBIC System [20], MIT Photobook System [21], and Virage System [22]. Due to the massive growth in the information content in images, region-based features have recently been developed to address the partial matching capability of CBIR. A region-based retrieval system segments images into regions (objects), and retrieves images based on the similarity of the regions. Typical region based systems include Berkeley BlobWorld [23], UCSB Netra [24], and Columbia VisualSEEK [25].

2.2 Recent Approaches in Image Information Mining

Image information mining is a relatively new idea in remote sensing where previous efforts have been focused on developing general-purpose image content retrieval systems that are interactive and have some level of intelligence built into them. The Knowledge Driven Information Mining in Remote Sensing Image Archives (KIM) system [11] and the EO domain specific Knowledge Enabled Services (KES) are examples of such a system. The KIM / KES prototype technique for information mining differs from traditional classification methods (e.g., a host of supervised or unsupervised methods). It is based on extracting and storing basic characteristics of image pixels and areas, which are then selected (one or more and weighted) by users as representative of the searched feature. Knowledge discovery and data mining based on hierarchical segmentation has also been proposed [26]. This approach provides capabilities for exploring the intrinsic properties of a region by a segmentation hierarchy, with the goal of developing heuristics for an automatic labeling of image regions. It also affords the opportunity for knowledge discovery on image data represented as a segmentation hierarchy.

This research proposes a framework for semantics-enabled knowledge discovery from Earth Observation (EO) data archives. The goal is to facilitate complex and more advanced, context-sensitive query processing over distributed data archives. This is achieved through the modeling of the information sources by domain specific ontologies, which are capable of capturing knowledge structures.

Ontology is defined as “a shared, formal conceptualization of a domain” [27]. Hence, ontologies can be used for data exploration / data integration tasks (because of their potential to describe the semantics of sources), to solve heterogeneity problems, and to provide varied levels of querying which facilitates knowledge discovery at different levels of granularity.

2.2.1 Metadata and Interoperability

Several metadata standards have been developed to address syntactic standardization [7]. A metadata standard that originated in the environmental community and was specifically designed for environmental and geospatial data is the Content Standards for Digital Geospatial Metadata (CSDGM), developed by the Federal Geographic Data Committee (FGDC) [8]. Extensions for FGDC have been developed that provide additional information particularly relevant to remote sensing (e.g., the geometry of the measurement process, the properties of the measuring instrument, the process of raw readings into geospatial information, and the distinction between metadata applicable to an entire collection of data and those applicable to component parts). NASA's Earth Observing System Data and Information System (EOSDIS) Core System (ECS) has developed metadata standards for the EOS data [28].

The main modules are Collection (50), Granule (26), Data Originator (34), Contact (16), Temporal (19), Spatial (57), Document (39), and Delivered Algorithm Package (47) (numbers in the parenthesis represents the number of elements in each the modules). One of the important goals of this standard was to allow data searches by scientists from diverse disciplines (e.g., atmospheric chemists, hydrologists, oceanographers), but also make the data accessible to non-experts (e.g., policy makers, educators). The structural metadata standards, such as the one developed by ECS, enables a user to have a variety of requirements for searching and ordering of the data

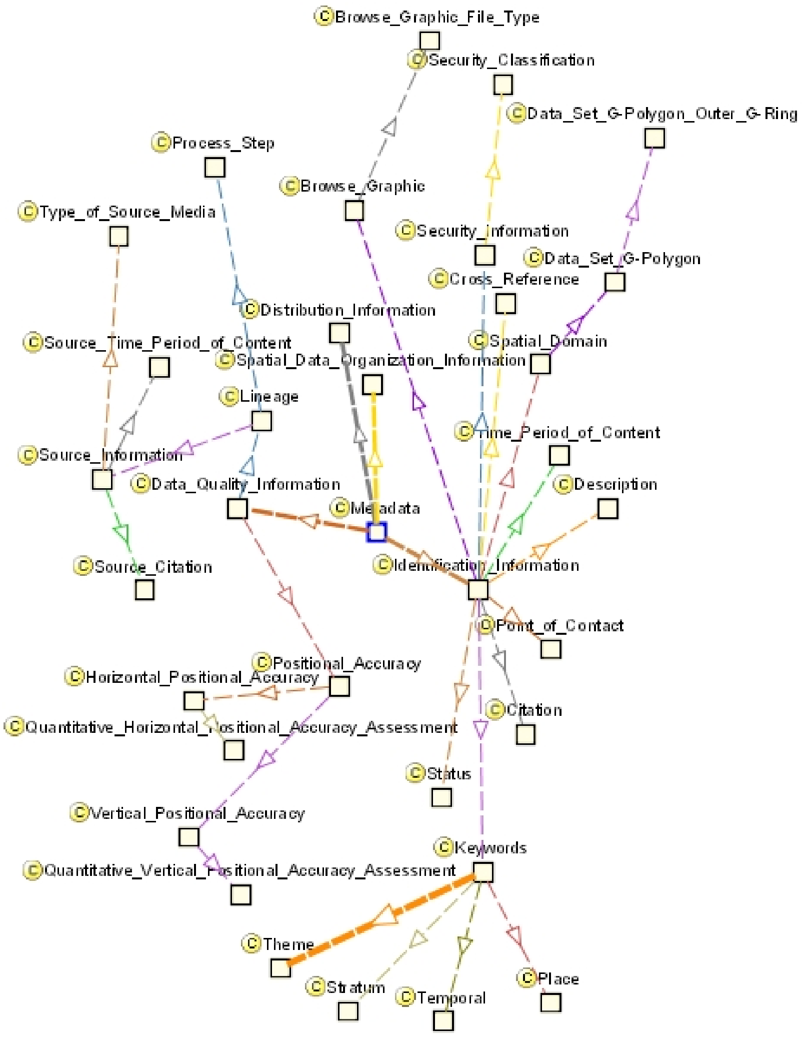


Figure 2.1 A portion of the FGDC-CSDGM in ontology form [29]

(e.g., a single granule or collection of granules). They also can provide browse or descriptive information prior to ordering the parent data (e.g., production history, storage format, production algorithms). All of these search requirements are satisfied by an exhaustive set of metadata elements. However, it is important to realize that these metadata standards allow us to structure the file contents, but they do not provide a semantic description of the domain of the information source. A more recent approach is the use of ontologies to make the conceptualization of a domain explicit. Figure 2.1 shows a portion of the FGDC-CSDGM conceptualization in ontology form [29]. The advantage of this approach is that it represents a standard that is widely accepted by the Earth science research community.

2.3 Semantics-Based Reconciliation of Disparate Information Sources

The importance of resolving semantic differences has recently gained wide attention in a variety of domains due to the progress in techniques to model, capture, represent and reason about semantics; gradual progress in attention from data to information, and increasingly towards knowledge acquisition and management. Ontologies are often used as interlinguas (an artificial language designed to be used for machine translation) for providing interoperability [30] for they serve as a common data format for data interchange. Ontologies help to solve the problem of implicit hidden knowledge by making the conceptualization of a domain explicit. Ontologies are useful for many different applications that can be classified into several areas [31]. Each of these areas has different requirements on the level of formality and the extent of explication provided by the ontology [32].

Many funded research projects have been initiated by the international community. Some of the frameworks for ontology-based applications are KAON [33], On-To-Knowledge [34] and Web-ODE [35]. In these systems, the middleware serves to hide the ontology sources from domain-specific application clients. Other major integration and retrieval systems are OntoBroker system, which implements the basic functionality of a single ontology information integration and retrieval system [36]. The Observer system is a multiple-ontology system, and uses query rewriting technique to translate between different ontologies [37]. The BUSTER [38] system uses a hybrid approach, in which it uses the shared terminology in query formulation and processing. A computational approach that compares concepts from unconnected and independent ontologies has been described in [39]. However, most of the above systems cater to the needs of the applications in domains like document repositories, office data repositories, web sites and other e-commerce applications and do not address specifically the requirements of the Earth science area. Hence, there is an urgent need to focus on the development of integrated systems that help in meaningful data sharing which is indispensable in this domain.

2.4 Semantic Web Technology and Relevance to Earth Observations

Semantic interoperability requires resolving various context-dependent incompatibilities, i.e. semantic conflicts. The context refers to the knowledge that is required to reason about another system for the purpose of answering a specific query. Therefore, it is important to provide contextual knowledge of domain applications in order to ensure semantic interoperability [40].

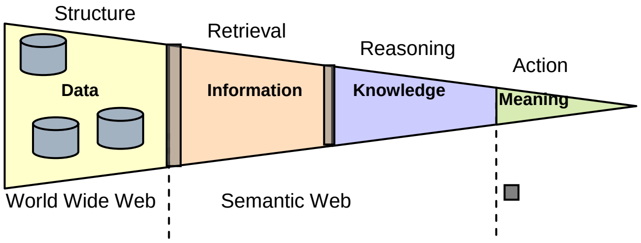


Figure 2.2 Process model for Semantic web driven knowledge discovery

The Semantic Web (Figure 2.2) is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [41]. It allows data to be shared and reused across application, enterprise, and community boundaries. Using web languages, such as RDF [42], DAML+OIL [43], and OWL [44] it is possible to create semantically rich data models. These models are made up of triples (subject-predicate-object), where subjects and objects are entities, and predicates indicate relationships between those entities. Implicit in these models is more information than can usually be found in their text representation [45]. Figure 2.3 depicts the components of ontology driven applications.

Earth Observations are obtained from a multitude of sources and requires coordination among different agencies and user groups to come to a shared understanding on a set of concepts involved in a domain. The realization of seamless interoperability and EO data integration is thus dependent on the resolution of conflicts arising from data represented in different data models, data sets from heterogeneous sources that differ in data modeling, scale, thematic content, contexts, meaning, etc. Thus, to enable computer programs to automatically generate transformations between different terminology systems is the core of the dream of the Semantic Web.

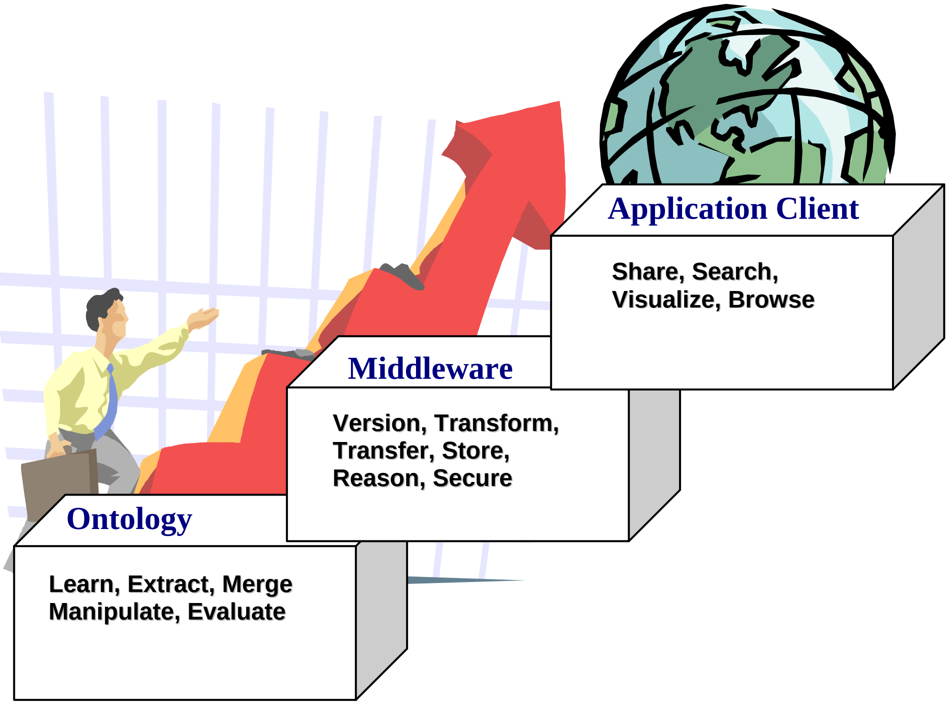


Figure 2.3 Components of ontology driven applications

The major philosophical difference between the Semantic Web and the World Wide Web is that the Semantic Web is supposed to provide machine-accessible meaning for its constructs, whereas in the World Wide Web this meaning is provided by external mechanisms. This meaning is largely based on the meaning of names which, in the Semantic Web, are URI references [42].

The use of Description logic-based ontologies has been one of the primary applications of the Semantic Web, which is a specific form of formal logic that can be run efficiently on a computer. Hence, this research pursues this line of thought and focuses on formulating the EO data integration problem in a knowledge representation framework instead of the prevalent database management system approach.

CHAPTER III

FRAMEWORK FOR SEMANTICS-ENABLED KNOWLEDGE RETRIEVAL FROM REMOTE SENSING DIGITAL DATA ARCHIVES

3.1 Introduction

This research proposes to develop an ontology middleware system that serves as a flexible and extendable platform for knowledge management solutions. The middleware facilitates the access and exploration of remote sensing digital archives through provision of ontology-based modeling of the concepts involved in the domain of interest and linking those concepts with predictive models developed through machine learning methods for imagery from different sensors. The proposed Intelligent Interactive Image Knowledge Retrieval (I^3KR) middleware serves to provide the following functionalities:

- An ontology server providing the basic storage services.
- Mechanisms for knowledge management.
- Support for integration of variety of reasoning modules suitable for various domains.

3.1.1 Approaches to Ontology Integration

In scientific discovery applications, it is necessary to examine data in different contexts, from different perspectives, and at varying levels of granularity. Since no single global ontology would satisfy the requirement, a shared ontology approach is proposed for this work. There are different approaches to ontology integration.

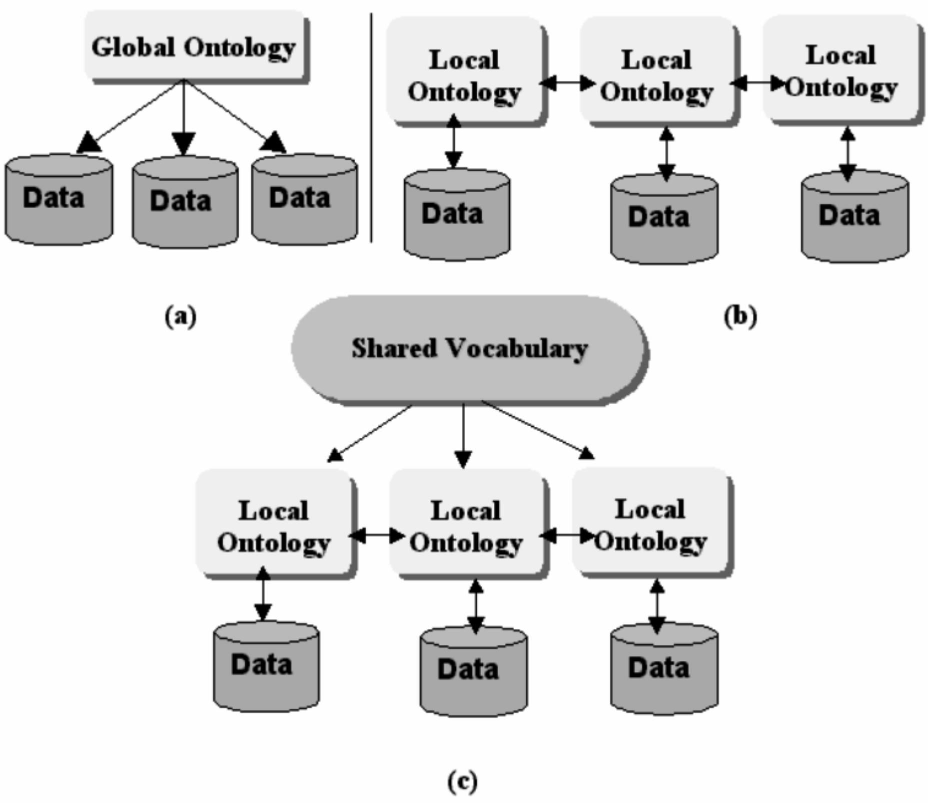


Figure 3.1 Shared ontology approach (Adapted from [45])

As shown in Figure 3.1 (a), independent data sources can be related to a single global ontology. However, this approach can be applied only to integration problems where all the data sources provide nearly the same view of the domain.

In addition, single ontology approaches are susceptible to changes in information sources that can affect the conceptualization of the domain represented by the ontology.

Figure 3.1 (b) illustrates a multiple ontology approach where each source is represented by its own ontology. No common or minimum ontology commitment is needed and each of the source ontologies can be developed without respect to other sources or their ontologies.

This architecture is useful to simplify the integration tasks and supports change (i.e., adding or removing sources). However, the lack of a common ontology makes it difficult to compare different source ontologies. A hybrid ontology approach consisting of a global shared ontology that encompasses all the local application level ontologies for a domain of interest (e.g., coastal zone) is adopted for this work. Recent studies [46] have suggested the advantages of this approach to be:

- New sources can be added easily without the need of modification.
- Supports acquisition and evolution of Ontologies.

The Ontology Web Language (OWL-DL) [44] is used to build the ontologies**.** Domain-specific ontologies help to define concepts in a finer granularity. These fine-grained concepts then allow us to determine specific relationships among features (e.g., shape, texture, color) in images that may be used to classify those images.

Three kinds of inter-relationships are used to create the ontology: IS-A, Instance Of, and Part-Of. These correspond to key abstraction primitives in object-based and semantic models**.** In Figure 3.2, the shared vocabulary is conceptualized in the form of a coastal zone ontology containing general terminologies encompassing the coastal zone. This enables the integration of the application ontologies based on the shared vocabulary of terms. Thus, water bodies that are classified by the International Geosphere-Biosphere

Programme (IGBP) land cover classification scheme can be used to explore the types of water bodies (e.g. river, lakes) by using the hydrology ontology. Further, if it is identified as a lake, it can be classified according to the trophic state (Eutrophic, Hypereutrophic, Oligotrophic, etc). Ontologies for Landsat and MODIS imagery based on the Anderson classification system [14] were developed. Further ontologies for land cover characteristics have been conceptualized in the IGBP ontology and concepts in the hydrology domain have been formalized.

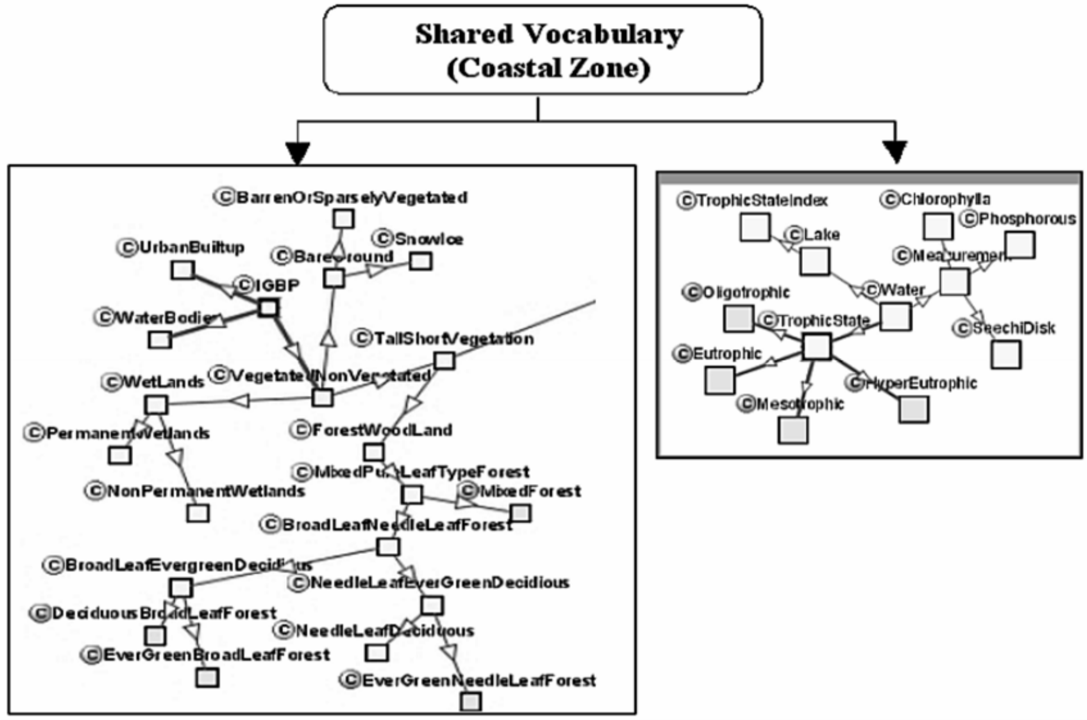


Figure 3.2 Integration of the application Ontologies (shown above are portions of IGBP and hydrology Ontologies) using shared vocabulary

The ontologies were modeled using Protégé-2000 [47], an open source ontology and knowledge base editor**.** Exploration of ontologies at various levels of granularity necessitates defining classes by restricting their property values. Then, by a combination of various restrictions, they are inherited into subclasses. The combinations of these restrictions define all conditions that must hold for individuals of the given class. Given below are the necessary and sufficient conditions that an information entity has to fulfill in order to belong to that concept.

3.1.2 Necessary Conditions

Concepts are described by a set of necessary conditions in terms of values of some properties. Thus, there are properties that are characteristic for a concept and can therefore always be observed for the instances of that class. However, they only apply in one direction: If we know that an object is a lake, then we can deduce that its tone is dark on a Near Infrared (NIR) band / False Color Composite (FCC) image, but we cannot deduce that a dark tone always belongs to a lake (i.e., it could be a shadow).

3.1.3 Necessary and Sufficient Conditions

An entity automatically belongs to the concept if it shows sufficient characteristic properties. Stronger, bi-directional relationships can be achieved by defining necessary and sufficient conditions for a class. Thus, by building necessary & sufficient conditions, intelligent tools (classifiers) can find additional characteristics of these classes.

Below are two examples for necessary and sufficient conditions in the two application domains as shown in Figure 3.2.

- ***Deciduous broadleaf forest*** (∃hasBiome {Mountains BorealConiferousForests SemiEvergreenAndDecidiousForests SchlerophyllousWoodlandsWithWinterRain

TemperateDecidiousForest}) ? (∃hasBiomeCode {B12 B10 B2 B7 B5}) ? (∃hasFoliage {SummergreenEvergreen DroughtDecidious Summergreen}) ? (∃hasRegion {TropicalSubtropical Other TemperateArctic}) ? (∃hasStructure{BroadleafForestAndWoodland MediumTallForest LowOpenForestWoodland})

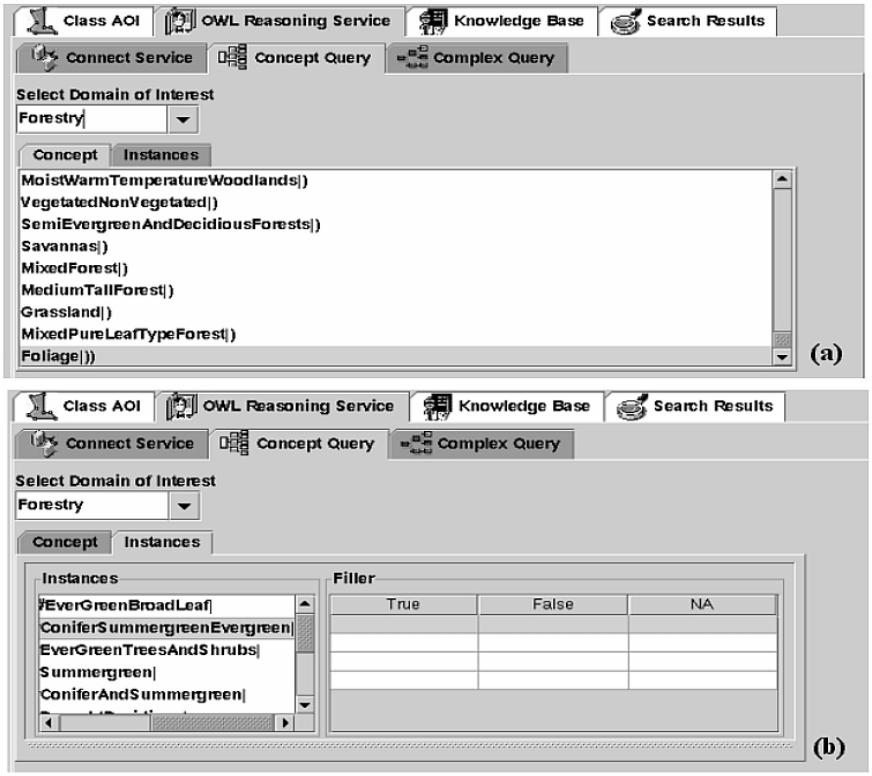
- ***Eutrophic lake*** (∃maxChlorophyplla э 60) ? (∃minChlorophyplla э 10) ? (∃maxPhosphorous э 100) ? (∃ minPhosphorous э 25) ? (∃maxSeechiDisk э 2) ? (∃minSeechiDisk э 0.5)

In the above expressions concepts have been defined for deciduous broadleaf forest and eutrophic lakes. In the case of the former concept, restrictions have been imposed on the object properties (e.g. hasBiome, hasBiomeCode, hasFoliage, hasRegion, and hasStructure) to formulate a defined concept. *Object properties* link individuals to individuals whereas *Datatype properties* link individuals to data values. When we define a property there are a number of ways to restrict the relation. The domain and range can be specified. The property can be defined to be a specialization (subproperty) of an existing property, etc. Several restrictions can be defined for properties such as allValuesFrom, someValuesFrom, Cardinality, hasValue etc.

For example the someValuesFrom restriction on the hasFoliage property of the class deciduous broadleaf forest is restricted to at least one value from SummergreenEvergreen, DroughtDecidious, and Summergreen foliage type.

The above process of building relationships will help in answering queries such as “Find all Eutrophic Lakes in year 2000 in Landsat ETM+ imagery for a particular area

X”. The application ontologies (e.g., hydrology, forestry) themselves make the concepts in the data source explicit. The hybrid ontology approach adopted in this work enables the development of application ontologies from a shared vocabulary (e.g., coastal zone, coastal hazards). Once the user selects the relevant concepts, the DL reasoning engine [48] executes the searches by automatically mapping between the query concepts of different application ontologies within the same domain. The Racer reasoner used in this work allows concept consistency checking and concept (re)classification based on inferencing. This proves to be very useful for determining subsumption relations and the identification of equivalence and disjointness between concepts. Reasoning between concepts is done within the so called TBox. Racer allows also Abox reasoning, based on individuals. All reasoning is done within the framework of Description Logics.



Forestry

MoistWarmTemperatureWoodlands|)
VegetatedNonVegetated|)
SemiEvergreenAndDecidiousForests|)
Savannas|)
MixedForest|)
MediumTallForest|)
Grassland|)
MixedPureLeafTypeForest|)
Foliage|))

Forestry

EverGreenBroadLeaf|
ConiferSummergreenEvergreen|
EverGreenTreesAndShrubs|
Summergreen|
ConiferAndSummergreen|

True	False	NA

Figure 3.3 Middleware depicting the concept query interface

To provide access to the ontologies, a concept query interface was developed, which allows access to the concepts of the shared vocabulary and application ontologies (Figure 3.3). The interface permits reasoning about possible matches with simple and complex concept searches. Once a user selects a concept (e.g., foliage), the corresponding instances are displayed in a list. This is useful when the user is uncertain about the exact semantics of the concepts for which he/she is looking. Once the user selects the relevant concepts, the DL reasoning engine executes the searches by automatically mapping between the query concepts of different application ontologies.

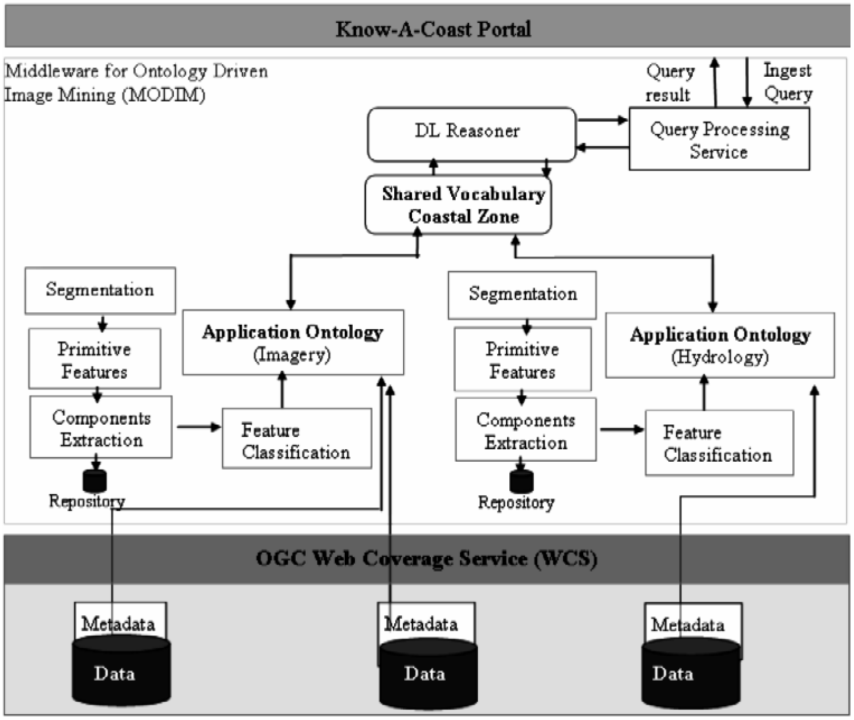


Figure 3.4 Framework for ontology driven image mining

As shown in Figure 3.4, the application ontologies (e.g., hydrology, land use/cover, imagery) make the concepts in the data source explicit. Once a query is ingested into the query processing service, it is processed and converted into a form usable by the DL reasoner. The DL reasoner allows classification of data from one context to another by equality and subsumption. Subsumption means that if concept B satisfies the requirements for being a case of concept A, then B can automatically be classified below A [49]. For example, if the user query is to retrieve all Evergreen Broadleaf forest then Tropical rain forest, Tropical degraded forest and Seasonal Tropical forest are subsume match.

This procedure enables query processing and searching in a way not possible with keyword-based searches. The Open Geospatial Consortium, Inc. (OGC) has developed an architectural framework for geospatial services on the web [50]. It specifies the scope, objectives, and behavior of a system and its functional components. It also identifies behaviors and properties that are common to all such services, but also allows extensibility for specific services and service types.

The framework in this work has been built upon the existing OGC Web Coverage Service (WCS), which enables a user with a service that has the capability to extract only the necessary data that meets his/her, requirements.

3.1.4 Primitive Features Extraction and Predictive Models Development

The task of content-based retrieval from remote sensing images begins at the primitive level, where the regions in an image are indexed based on the color, shape and texture of each region. These are machine-centered features and require the association to a meaningful set of concepts at the higher level. This association is achieved by mapping the keywords and concept descriptors by a higher-level domain specific ontology. This enables reasoning against the ontology and the ability to examine the relationships among the identified objects and associate the proper concepts with the image.

In I^{3}KR, a region-based approach is adopted, which starts by applying a segmentation algorithm [51] to the tiled image (Figure 3.5). The goal is to assign a semantic meaning to the generated regions by mapping them to concepts in the domain-specific ontologies.

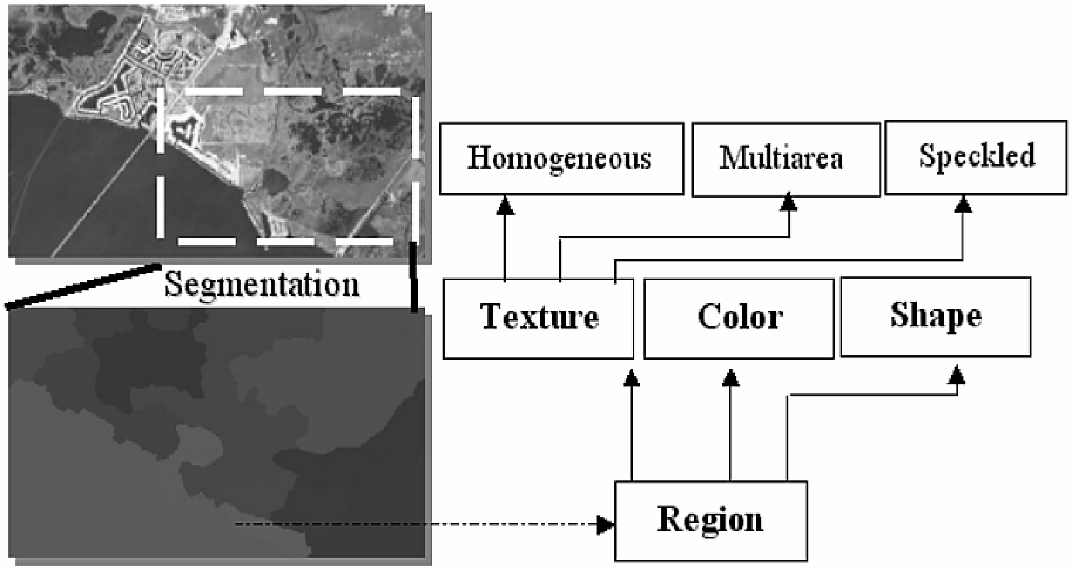


Figure 3.5 Unsupervised segmentation of the image and subsequent feature extraction using texture, color, and shape parameters

3.1.5 Kernel Principal Component Analysis (KPCA)

The feature extraction task produces large volumes of data that are difficult to manage and requires the estimated image parameters to be compressed [11]. A kernel PCA, proposed as a nonlinear extension of a PCA [52], [53] computes the principal components in a high dimensional feature space F, which is nonlinearly related to the input space. A kernel PCA is based on the principle that since a PCA in F can be formulated in terms of the dot products in F, this same formulation can also be performed using kernel functions without explicitly working in F (Figure 3.6). A kernel PCA has been shown to provide better performance than a linear PCA in several applications [54].

Given a set of centered m samples [53], kernel PCA diagonalizes the estimate of the covariance matrix of the mapped data $\Phi(x_i)$

$$C = \frac{1}{m}\sum_{i=1}^{M}\Phi(x_i).\Phi(x_i) \tag{1}$$

Finding the eigenvalues for the covariance matrix C ,

$$\lambda w = Cw \tag{2}$$

for eigenvalues $\lambda \geq 0$ and eigenvectors $w \in F$.

As $Cw = (1/M)\sum_{i=1}^{M}(\Phi(x_i).w)\Phi(x_i)$ (3)

all solutions w with $\lambda \neq 0$ lie within the span of $\Phi(x_1),......\Phi(x_M)$ i.e., the coefficients $\alpha_i\,(i = 1,....,M)$ exist such that

$$w = \sum_{i=1}^{M}\alpha_i\Phi(x_i) \tag{4}$$

Denoting an $m \times m$ matrix K by

$$K_{ij} = k(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{5}$$

then the kernel PCA problem becomes

$$m\lambda K\alpha = K^2\alpha \equiv m\lambda\alpha = K\alpha \tag{6}$$

where α denotes a column vector with entries $\alpha_1,......\alpha_m$

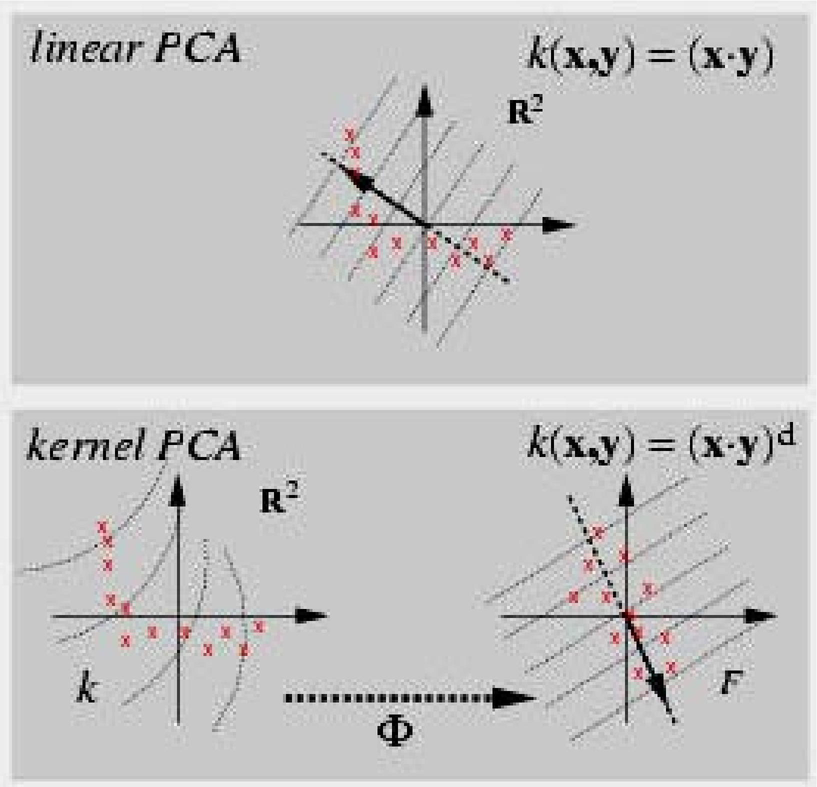

linear PCA
$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})$
$\mathbf{R}^2$
kernel PCA
$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})^d$
$\mathbf{R}^2$
k
Φ
F

Figure 3.6 Kernel-PCA implicitly performs a linear PCA in some high dimensional feature space that is nonlinearly related to input space (adapted from [53]).

The primitive features that were extracted from each region are used to perform an unsupervised classification (using KPCA), which extracts components of features that are nonlinearly related to the input variables. This process also reduces the data size**.** The resulting components are stored in a database.

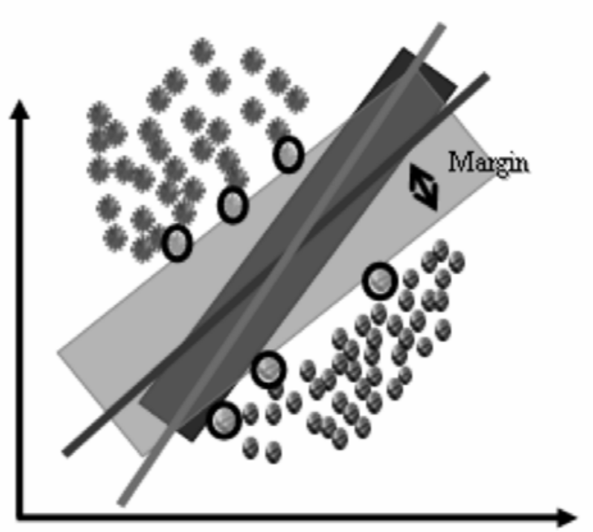


Figure 3.7 Linearly separable case; only support vectors (dark circled) are required to define the optimally defined hyperplane.

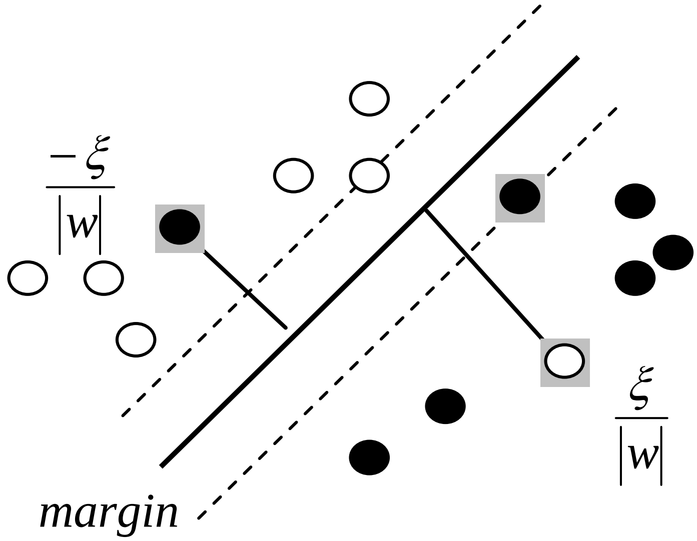


Figure 3.8 Non-Linearly separable case; only support vectors (dark circled) are required to define the optimally defined hyperplane.

3.1.6 Support Vector Machines

Support Vector Machines (SVMs), as originally introduced by Vapnik within the area of statistical learning theory and structural risk minimization [55], have proven to work successfully on many applications of nonlinear classification and function estimation. SVMs can be used for both classification and regression problems. Some applications of SVMs for classification are isolated handwritten digit recognition [56], object detection [57], and face detection in images [58]. The problems are formulated as convex optimization problems, usually quadratic programs, for which the dual problem is solved. Within the models and the formulation one makes use of the kernel trick, which is based on the Mercer theorem related to positive definite kernels [59]. One can plug in any positive definite kernel (e.g., linear, polynomial, or Radial Basis Function (RBF)) for a support vector machine classifier.

We try to find an optimal hyperplane that separates two classes (Figure 3.7 & 3.8). In order to find an optimal hyperplane, we need to minimize the norm of the vector w, which defines the separating hyperplane. This is equivalent to maximizing the margin between two classes. Given a set of instance-label pairs,

$(x_i, y_i), i = 1,...,l$ where $x_i \in R^N$

Let the decision function be

$$f(x) = sign(w.\phi(x) + b) \tag{7}$$

To maximize the margin (distance between hyperplane and the nearest point) the SVM [55] [60] requires the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \qquad (8)$$

subject to

$$y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \qquad (8a)$$

$$\xi_i \geq 0 \qquad (8b)$$

Where $w \in R^N$, $b \in R$ define a linear regressor in the feature space, which is nonlinear in the input space. In addition ξ and C, respectively are the positive slack variable and the penalization applied to the errors.

The dual solution to this problem is to maximize the quadratic from

$$L_D(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i . x_j) + \sum_{i=1}^{n} \alpha_i \qquad (9)\ \text{s.t}$$

$$0 \leq \alpha_i \leq C i = 1,...n \qquad (9a)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (9b)$$

Here $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function.

Normally the training data are separated into two parts; one is used for training and the other is used for testing. An improved version of handling training sets is cross-validation. In V-fold cross-validation, we first divide the training set into V subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining V-1 subsets. Thus, each instance of the whole training set is predicted once and the cross-validation accuracy is the percentage of data that are correctly classified. The cross-validation procedure can help alleviate over-fitting the data. I^{3}KR uses five-fold cross-validation. As shown in Figure 3.9 better boundary delineation could be achieved by performing KPCA followed by a SVM classification.

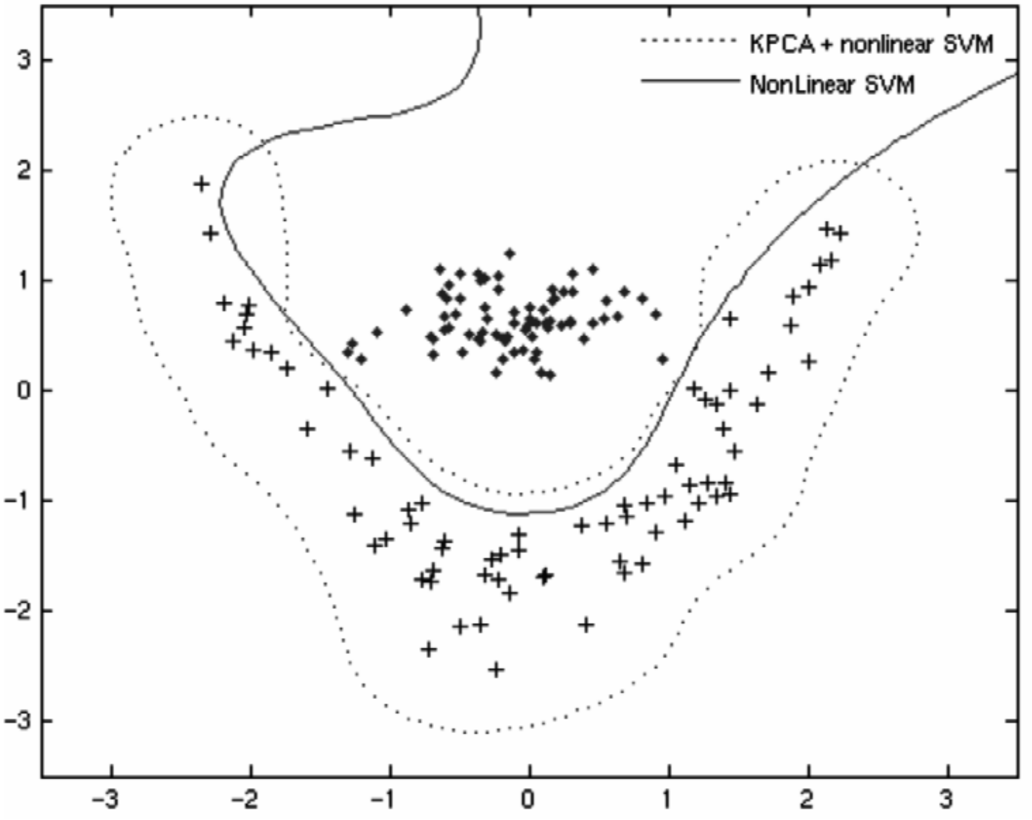


Figure 3.9 KPCA followed by SVM classification provides better boundary delineation

3.1.7 Probabilistic Outputs from SVM

The outputs from a binary SVM do not allow for post processing of the result. Calibrated posterior probability $P(class|input)$ is very useful in practical recognition scenarios [61]. It has particular relevance in the proposed approach to image mining from remote sensing archives. The following are the advantages of having probabilistic outputs:

- Provides a feedback about the strength of the classified object.
- Ranks the classified output with respect to their relevance to the user query.
- Combine the classification outputs for an overall decision-making function.
- Useful for concept selection from application ontology.

Instead of predicting the class label, the posterior class probability $p(y=1|x)$ can be approximated by a sigmoid function [62]

$$p(x) = \frac{1}{1+\exp(Af(x)+B)} \tag{10}$$

with parameters A and B. The best values for them are estimated using maximum likelihood estimation from a training set (f_i, y_i) given by

$$\min_{z=(A,B)} F(z) \tag{11}$$

where

$$F(z) = -\sum_{i=1}^{l} \left(t_i \log(p_i) + (1-t_i)\log(1-p_i)\right) \tag{12}$$

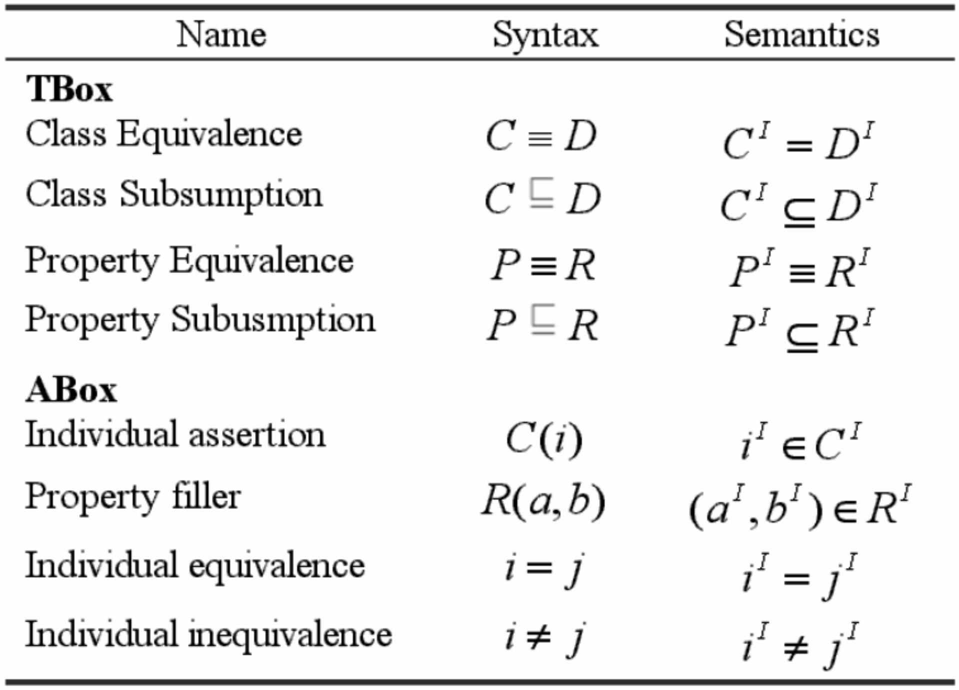
$p_i = \dfrac{1}{1+\exp(Af_i + B)}$, $f_i = f(x_i)$ and t_i are target probabilities defined as $t_i = \dfrac{y_i + 1}{2}$

The posteriori probability is a measure of how probable an image is of a particular type [11]. We calculated the posterior probabilities of the predicted land cover types given a particular image region.

3.1.8 DL- Based Concept Selection

The model generated by the SVM is used to predict an unknown region. The object obtained as a result of the prediction has to be assigned to the proper concept in the hierarchy of the domain specific ontology. The goal is to map the predicted objects to the ontology concepts through Description Logic. In the DL systems [63], a knowledge base consists of an ABox and a TBox (originally from “Terminological Box and Assertional Box” respectively). The TBox stores a set of universally quantified assertions, stating general properties of concepts and roles (e.g., “Deciduous Broadleaf forest has at least one structure.”). A typical assertion in an ABox is one stating that an individual is an instance of a certain concept (e.g., one can assert that Lake X is an instance of a “Lake, which is Eutrophic.”).

Table 3.1 Description Logic (DL) axioms [63][4]



Name	Syntax	Semantics
TBox		
Class Equivalence	$C \equiv D$	$C^I = D^I$
Class Subsumption	$C \sqsubseteq D$	$C^I \subseteq D^I$
Property Equivalence	$P \equiv R$	$P^I \equiv R^I$
Property Subusmption	$P \sqsubseteq R$	$P^I \subseteq R^I$
ABox		
Individual assertion	$C(i)$	$i^I \in C^I$
Property filler	$R(a,b)$	$(a^I, b^I) \in R^I$
Individual equivalence	$i = j$	$i^I = j^I$
Individual inequivalence	$i \neq j$	$i^I \neq j^I$

We can distinguish four kinds of assertions for a TBox and an ABox. Class assertions (Table 3.1) express that an individual is a member of a class. Property fillers express that two individuals are related to each other through a given property.

A classification problem is characterized by the determination of membership relations between an object under consideration and a set of predefined concepts [64]. The match between observations (model predicted objects) and membership conditions is performed using knowledge that associates properties of the objects with their classes. This can be formalized in the following way [65]

- Let C be a set of solution classes (concept predicates (Water Body, Vegetation, etc.))

- Let O be the set of observations (i.e., the necessary conditions for concept membership) { $N^c | c \in C$ })
- Let R be a set of classification rules (sufficient conditions for class membership { $S^c | c \in C$ })

Then a classification task is to find a solution class $c_i \in C$ such that

$$O \wedge R \Rightarrow c_i(X) \tag{13}$$

Therefore, a single information entity can be translated from one context into another by finding a concept definition in the target structure satisfying the above expression. In I^{3}KR, the classification is handled by the middleware that integrates the concepts of the current domain by sending a request to the DL reasoner. Since the concepts in the application ontologies are formed from a global shared ontology, after reclassification all the sub concepts of the query will form the result.

3.2 Methodology for Rapid Image Information Mining

This section presents approaches for the development and application of image information mining components for the following:

i) The application of image information mining in coastal disaster events, with particular emphasis on the image mining of post hurricane events. Results are presented using imagery from Landsat ETM+ of post hurricane Katrina flooding.

ii) Proposes an image information mining system that is fast and reliable with the capability to perform the tasks of identifying affected regions with minimal expert supervision.

iii) Previous efforts in image information mining area have focused mainly on the reduction of features using clustering approaches [11] [66], but little has been reported on the selection of best feature subsets. This research enables predictive model development that goes in conjunction with feature selection and feature generation.

3.2.1 Feature Selection and Generation for Image Information Mining Applications

Feature selection is defined as the selection of a subset of features to describe a phenomenon from a larger set that may contain irrelevant or redundant features [67]. Feature selection techniques usually involve a criterion function and a search algorithm. The former aims at evaluating the separability of classes for a given subset of features. And the latter identifies the subset of features that maximize the adopted criterion [68]. Several separability indexes have been proposed in the remote sensing literature [69-73].

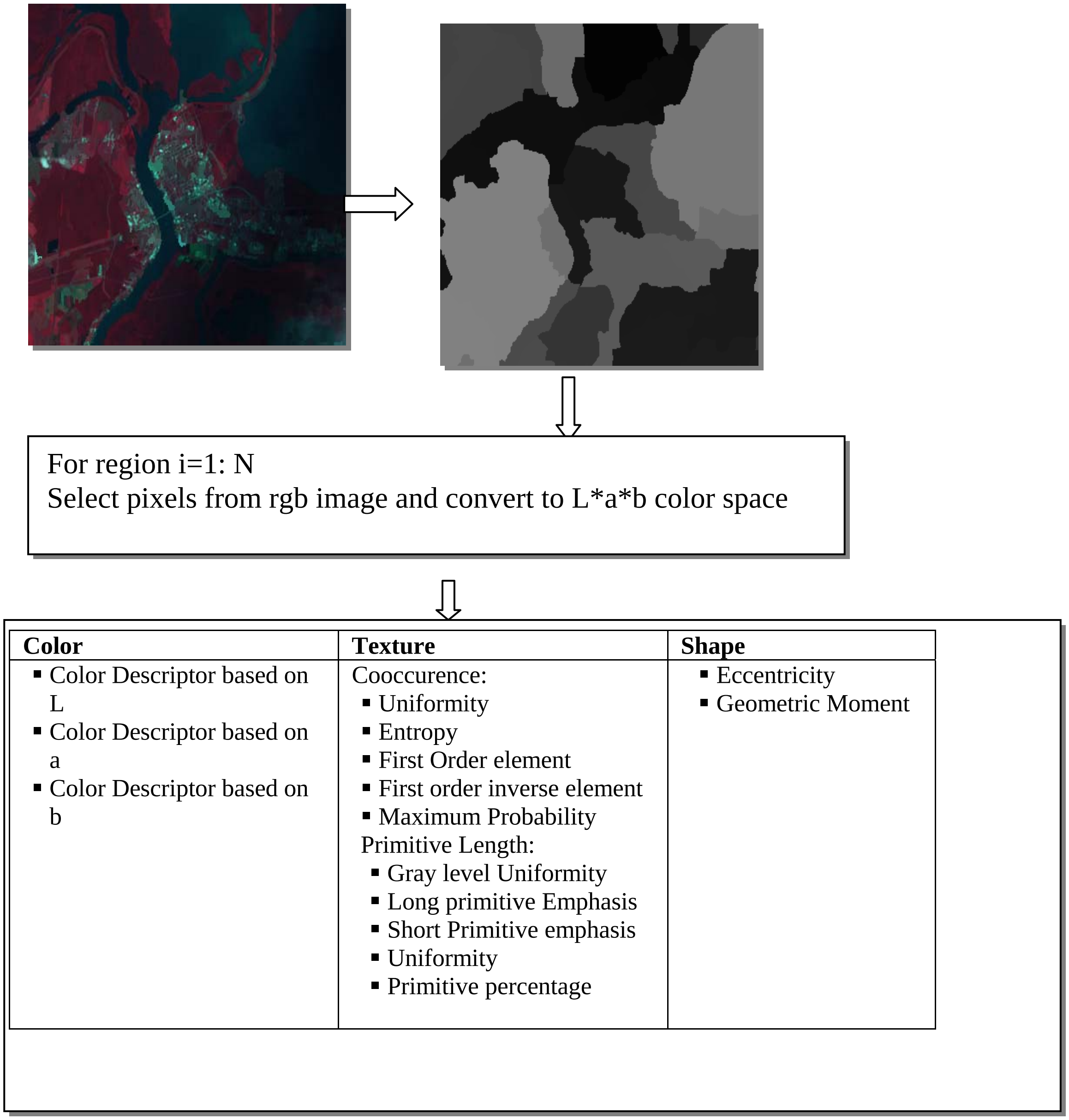
These indexes are generally based on the average distance among the classes, and are computed by using the statistical distance between the pair of classes and are dependent upon the set of features considered. Lorenzo Bruzzone [68] argues that criterion functions based on the average pair-wise distances without taking into consideration the costs associated with classes are not appropriate for selecting features that minimizes the total classification cost.

A criterion function based on the Bayes rule for minimum cost (BRMC) has been proposed [68] and uses a neural network as the induction algorithm. However, recently it has been noted that the feature selection stage and classification stage are not independent because the goal is correct classification with a corresponding feature pattern extracted with the intermediate step of feature extraction and dimensionality reduction [74]. Hence, it is recommended to couple feature selection with effective classification techniques. Wrappers-based feature selection is a methodology that has had a long history within the literature on statistics and pattern recognition [75], but its use within machine learning is relatively recent, and to the best of our knowledge no studies have been made for their applications in the remote sensing domain. In a wrapper approach, the feature subset selection algorithm exists as a wrapper around the induction algorithm. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as a part of the function evaluating feature subsets [76] [77] [78]. While giving good results in terms of accuracy of the final classifier, wrapper approaches are computationally expensive.

Other methods, such as filter methods, are much faster than wrappers but produce disappointing results because they ignore the induction algorithm [77]. However, wrapper-based methods can be effectively used in image information mining applications which are region-based instead of pixel based. The proposed RIIM is a region based framework that encodes knowledge at the regional level instead of pixel level, hence the computation cost is reduced making it an effective medium for incorporating wrapper based methods. The RIIM system adopts a Genetic algorithm-based wrapper approach for feature selection and generation. Genetic algorithms (GAs) are randomized search and optimization techniques guided by the principles of evolution and natural genetics. They are efficient, adaptive and robust search processes, producing near optimal solutions and have a large amount of implicit parallelism.

The utility of GAs in solving problems that are large and multimodal and highly complex has been demonstrated in several areas [79]. GAs have been used to search for feature subsets in conjunction with several classification methods such as neural networks [80, 81], decision trees [82], k-nearest neighbors [83-86] and Naïve Bayes [87, 88].

The rapid image information mining (RIIM) system uses machine learning to address the task of exploring remote sensing imagery based on its content. The process of knowledge extraction from the imagery starts with the creation of tiles of the full scenes of the images and then performing an unsupervised segmentation on each tile. We use hierarchical segmentation algorithm [51] to delineate regions of interest which are then used for primitive features extraction. Before performing the low-level feature extraction, each region is converted from RGB color space to CIE L*a*b* color space. This color conversion has been dictated by the fact that L*a*b color space provides more perceptual color information. Fifteen primitive features based on color, texture and shape were extracted. Figure 3.10 depicts the low level feature extraction algorithms used in this study.



For region i=1: N
Select pixels from rgb image and convert to L*a*b color space

Color	Texture	Shape
▪ Color Descriptor based on L ▪ Color Descriptor based on a ▪ Color Descriptor based on b	Cooccurence: ▪ Uniformity ▪ Entropy ▪ First Order element ▪ First order inverse element ▪ Maximum Probability Primitive Length: ▪ Gray level Uniformity ▪ Long primitive Emphasis ▪ Short Primitive emphasis ▪ Uniformity ▪ Primitive percentage	▪ Eccentricity ▪ Geometric Moment

Figure 3.10 Low level primitive features extracted from the image

Thus, the selection of relevant features, and the elimination of irrelevant ones, is one of the central problems in machine learning, and many induction algorithms incorporate some approach to address this issue. Numerous search algorithms have been used to search for feature subsets [89]. The application of evolutionary learning algorithms to pattern recognition is becoming increasingly common. A variety of researchers have used evolutionary algorithms to perform feature selection [90]. The majority of these approaches begin with a large pool of potential features and an evolutionary process is used to evolve a population of feature subsets drawn from the pool. The subsets are evaluated using a standard classifier.

Feature selection algorithms can be categorized into exponential, randomized and sequential algorithms. Exponential algorithms (e.g., branch & bound, exhaustive) have exponential complexity in the number of features and are frequently prohibitively expensive to use. Randomized algorithms include genetic and simulated annealing search methods and attain high accuracies. Sequential search algorithms have polynomial complexity and add or subtract features and use a hill-climbing strategy.

Sequential forward selection (SFS) begins with zero attributes, evaluates all feature subsets with exactly one feature, and selects the one with the best performance. It then adds to this subset the feature that yields the best performance for the subsets of the next larger size. This cycle repeats until no improvement is obtained from extending the current subset. Sequential Backward Selection (SBS) begins with all features and repeatedly removes a feature whose removal yields the maximal performance improvement. The sequential forward selection and its backward counterpart are suboptimal methods that obtain a chain of nested subsets of features in a straight forward manner. This nesting effect constitutes one of their main drawbacks [91].

The algorithms cannot correct previous additions (deletions) of features. Also feature interaction is not taken into consideration in these methods. Feature interaction is characterized as a situation in which the effect of a feature on the target (semantic class) depends on the value of other features [92].

In this work a wrapper approach that uses a modified genetic algorithm was used for the incremental selection and generation of new features [93]. It uses an attribute-based induction algorithm for the evaluation of the features at hand. Inductive learning involves the process of *learning by example* (i.e., where a system tries to induce a general rule from a set of observed instances). Computational studies of Darwinian evolution and natural selection have led to numerous models for solving optimization [94-98]. GAs comprises of a subset of these evolution-based optimization problems techniques focusing on the application of selection, mutation, and recombination to a population of competing problem solutions [99-100]. The population is usually formed from a constant number of individuals representing samples from the search space.

3.2.2 Chromosome

In the RIIM system a chromosome is defined as an individual whose length is the same as the total number of features corresponding to each segmented region in the image. Each bit of the chromosome is initialized by a randomly selected 0 or 1. The fitness of the chromosome during the evolutionary process is calculated by considering only features that have 1s.

New individuals for the next generation are formed by applying two genetic operators; crossover and mutation to the individuals from the current generation. In each generation, half of those individuals with higher fitness values survive, and the others are extinguished. Two parents are selected from the survived individuals, and children are generated by a two-point crossover.

3.2.3 Crossover

The crossover process defines how genes (chromosomes) from the parents have been passed to the offspring. In each generation, once two individuals are selected as the parents, a gene from each parent is broken into several segments and recombined with gene segments from the other parent based on a predefined crossover probability. After the crossover operation, every two parents will produce two children. The above selection and crossover process will continue to run in each generation until the number of children equals the population size. At the end of each generation, it is useful to pass a certain number of the best individuals directly to the next generation, which is called elitism. In elitism, the best individual from the current generation is copied directly to the next generation, and is used for fast convergence.

3.2.4 Mutation

The mutation process simulates the natural disturbance during crossover. It is a bit-by-bit operation based on the mutation probability (mutation rate). Mutation rate is generally selected based on the population size and other factors, such as selection method and with or without an elitism policy. The mutation operation follows immediately after the crossover operation; Figure 3.11 depicts the two points cross over and mutation process used in GA. Parents that will produce new individuals are chosen according to their fitness. Better individuals are more likely to pass their genes to the next generation. Therefore, each generation should have a better overall fitness.

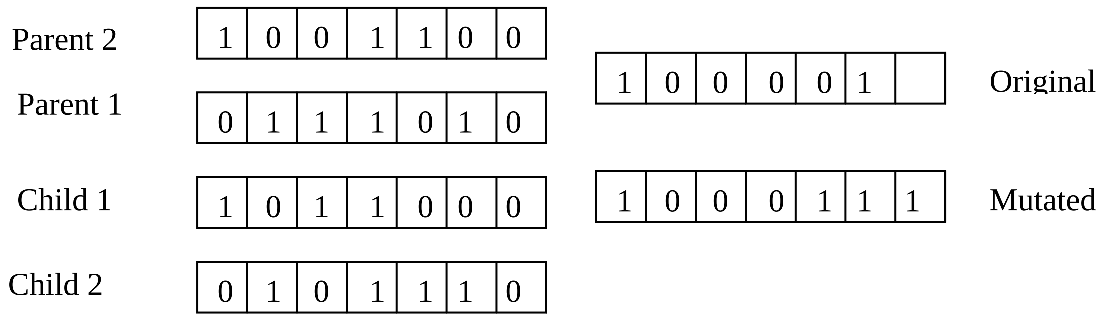


Figure 3.11 Cross over and mutation

3.2.5 Hybrid Wrapper- Based Genetic Algorithm Approach for Feature Selection and Generation

This algorithm combines the positive search properties of conventional genetic algorithms with the option to adapt the search space incrementally. In the wrapper approach the feature subset selection algorithm exists as a wrapper around the induction algorithm. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm itself as part of the function evaluating the feature subsets.

As shown in the Figure 3.12, the outer cross-validation wrapper randomly splits the original data set into n equally sized parts. For each run, the ith part is kept as a test set while the remaining n-1 parts are passed to the genetic algorithm and subsequently to the final SVM learner whose learned model is tested on the ith part of the dataset.

The inner cross-validation trains the model on the training data training set and optimizes the choice of an attribute set using the disjunct evaluation data test set to avoid a bias in the selection of the attribute set. For reliable performance estimation of the complete operator chain for feature selection and classification learning the training evaluation and test data sets need to be disjunct, which is guaranteed here by the nested cross validations [93]. The combined feature selection and generation of new features using a wrapper based approach employs an attribute-based induction algorithm for the evaluation of the feature sets.

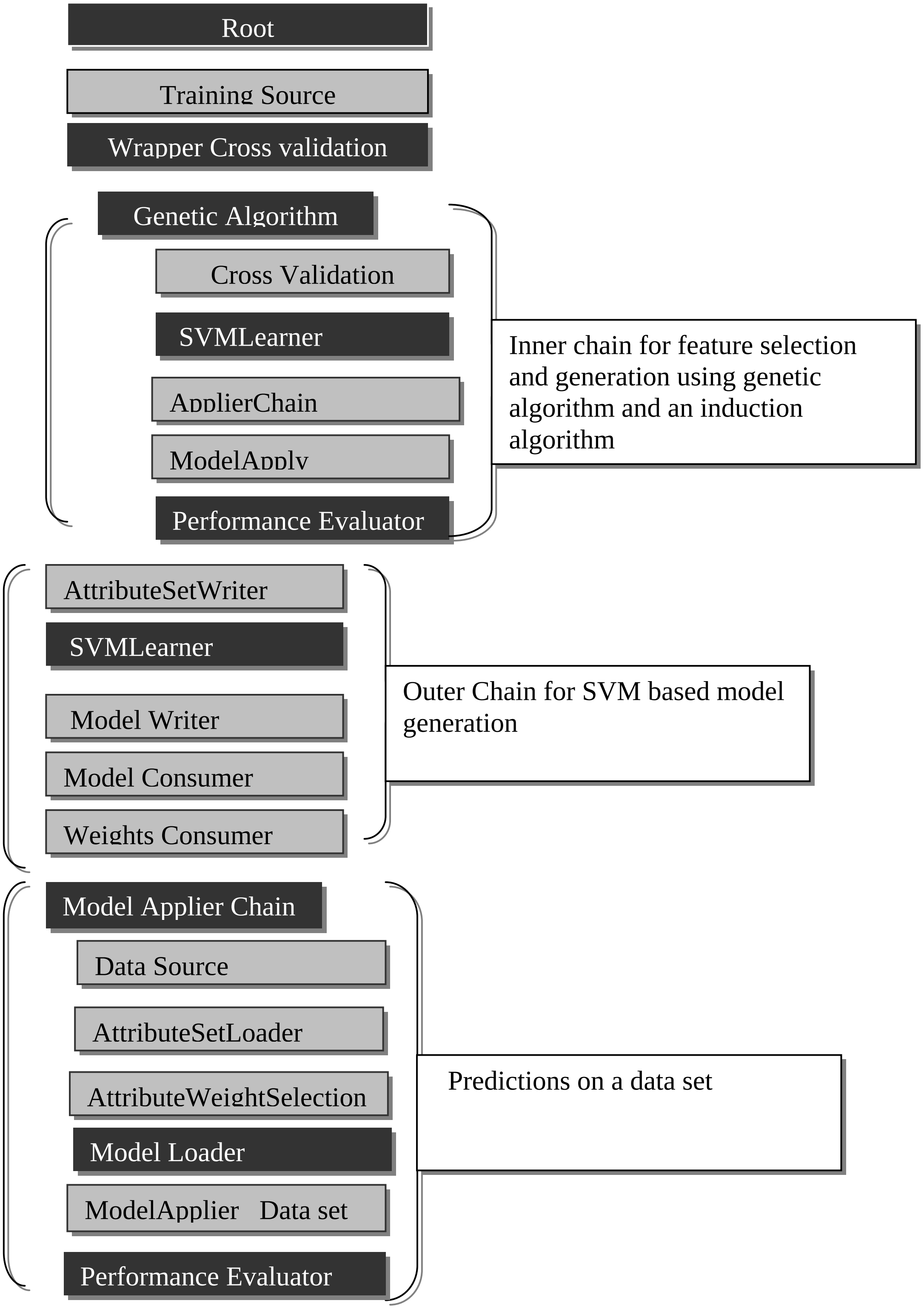


Figure 3.12 Algorithm for Wrapper-based approach for feature selection, generation, model creation and performance evaluation

In the combined feature selection and generation case, in addition to the standard mutation operator the crossover operator is modified to make it a variable length crossover operator which is based on the variable length genetic algorithms [101] a new operator that produces new features is also introduced. This operator uses a type restriction on the generator operator (e.g. Boolean, mathematical) to combine a given set of compatible features, resulting in new features generation. Figure 3.13 depicts this modified feature generator operator. For instance, the suitable features from *Colordescriptor1, uniformity, entropy, eccentricity* could be combined with an addition operator which produces a new feature and added to the original feature set. The set of the compatible features is not limited to the original features, but can contain compound features that have already been created by the generator [93]. The above methodology has been adopted for feature selection and generation in this study.

3.2.6 Materials

To evaluate the presented wrapper-based approach on hurricane-related events, data from Landsat ETM+ data (path 22, row 39, Aug 30, 2005) corresponding to post-Katrina hurricane, and Landsat ETM + data (path 23, row 36, Sep 22, 1999) that is not related to the hurricane (pre-hurricane) were used in this study. This would help to identify training sites of different land covers that are specific to post hurricane areas (e.g. flooded fastlands) and training sites that are present, in general, during other times (agriculture, fallow, etc). Such a strategy helps to develop predictive models which have the capability for image information mining from affected areas and compare the land cover classes in the pre and post-hurricane events and also the evaluation over a period of time.

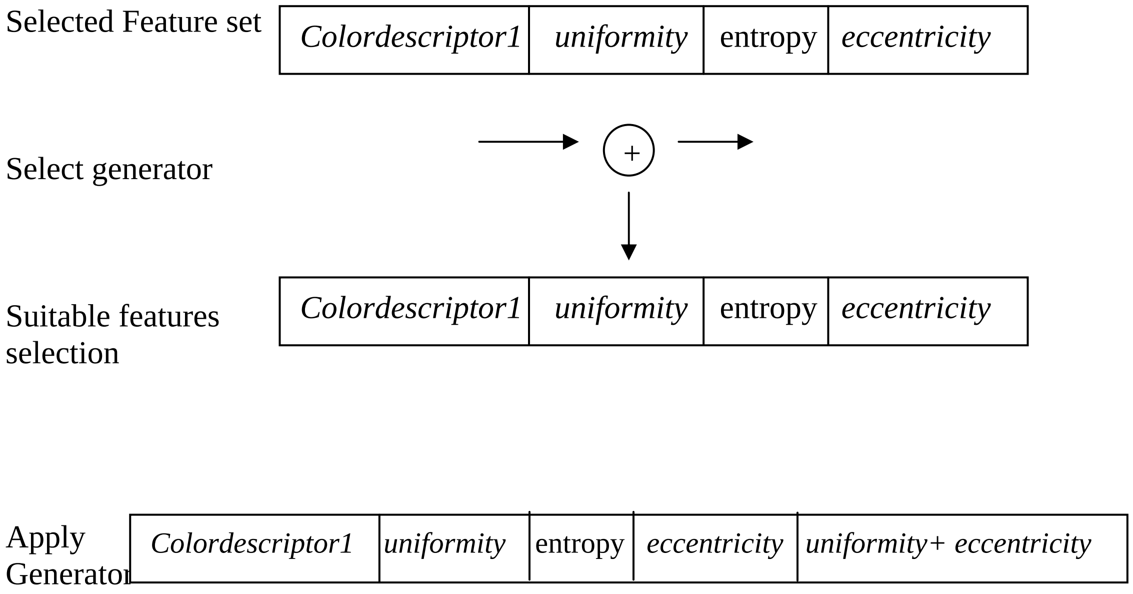


Figure 3.13 Modified feature generator used in the combined approach (Modified from [93])

The database consists of primitive features from 7,117 segmented regions extracted from 60 tiles (each of 967 x 915 dimension) corresponding to post-hurricane Landsat ETM+ data and primitive features from 4592 segmented regions extracted from 60 tiles (each of 719 x 575 dimensions) corresponding to pre-hurricane data. The bands 4, 3, 2 corresponding to near infrared, red, and green were selected and the false color composites (FCC) were derived from these bands.

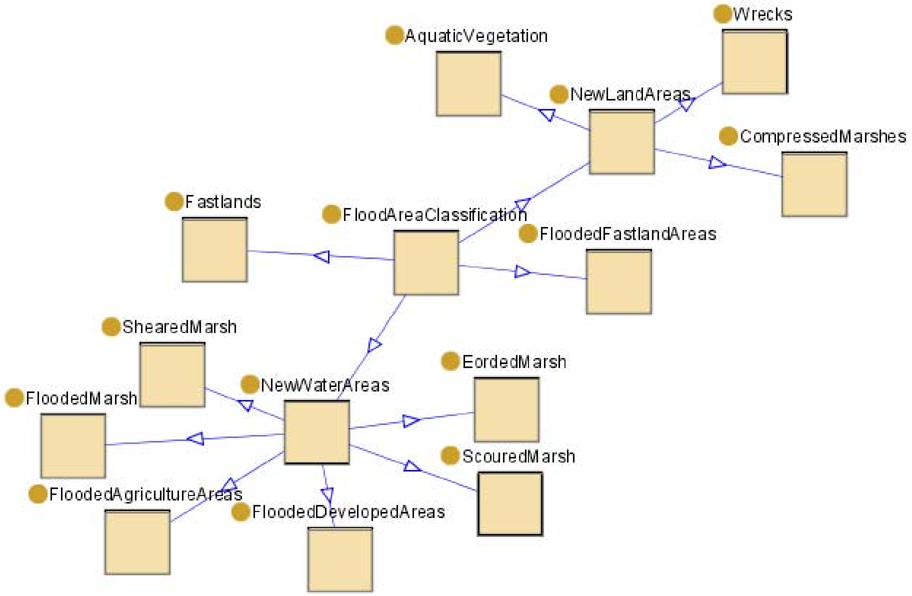


Figure 3.14 USGS wetlands classification

Table 3.2 presents the number of training samples used for each semantic class. The flooded area classes selected for this study were based on the USGS-NWRC classification system (Figure 3.14), which provides specific land cover types that need to be assessed in a post-hurricane scenario.

Table 3.2 Training samples used in the study; each sample corresponds to a region (area depends on segmentation granularity) in the image.

Semantic Class	Number of samples
Flooded Fastlands (includes flooded agriculture and developed areas)	100
Flooded Vegetation (includes flooded marshes)	175
Agriculture	156
Fallow	385
Forests	150
Clouds	120
Water bodies	100
Total Training data	1086

The results from the above described framework are presented in Chapter 5.

CHAPTER IV

FRAMEWORK FOR SEMANTICS-ENABLED THEMATIC DATA RETRIEVAL

4.1 Introduction

In this section, the focus is on the semantic heterogeneity problem in Earth Observation (EO) data, where there is usually no possibility for human interpretation and intervention within a service chain. In such a scenario, formalizing the semantics of geographic information has become indispensable.

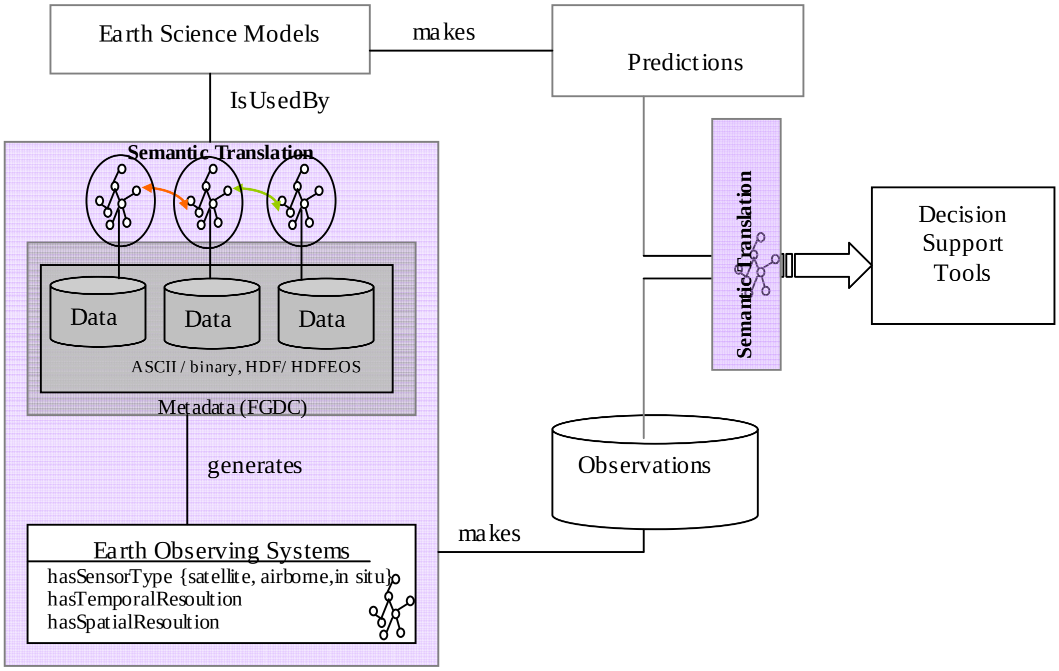


Figure 4.1 Need for resolving semantic heterogeneities in integrated systems network

This framework will be presented through a motivating example (resolving semantic heterogeneities between various land cover classification schemes) of tackling the problem of semantic heterogeneity in the thematic information sources related to Land Cover. This has particular significance in the overall integrated system approach (Figure 4.1) where the key to understanding the model data requirements is the domain-specific conceptualization of the data (e.g. ontological modeling), and formulating it in a way that the context of the source is well understood. This would enable to transform data into different contexts as required by a specific Earth science model or a decision support tool.

4.2 Semantic Conflicts

An explicit semantic description of the contents of the data is required to understand the context. However, the data description in terms of its semantics, which fully describes the data products, is a very challenging task and requires formulation of the information sources in ways that would help in automated processing or machine understandability. The three main causes of semantic heterogeneity are [12]:

- *Confounding conflicts* occur when information items seem to have the same meaning, but differ in reality, e.g. due to different temporal contexts.
- *Scaling and Units conflicts* occur when different reference systems are used to measure a value, (e.g. currencies).
- *Naming conflicts* occur when the naming schemes of the information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

The investigation of *confounding conflicts* is significant to EO data as the ability to collect imagery of the same area of the Earth's surface at different periods of time is one of the most important elements for applying remote sensing data. Change detection studies are routinely performed from such multi-temporal data, whose output is again highly context-sensitive and depends on the project specific goals. For example, change detection studies for deforestation accounting for greenhouse gases and change detection studies that look at deforestation and the associated urban sprawl. If two such information sources exist, then it is necessary to identify whether a value is an intrinsic and permanent property of some instance, or if it depends on some evaluation context and, in the latter case, by associating this value with its context it is possible to achieve interoperability.

Scaling conflicts frequently occur in the EO thematic data representation. Land use / land cover information, is one of the major sources of geographic information available today. It is highly heterogeneous in syntax, structure and semantics [10]. The heterogeneities arise because land use/land cover data are produced and provided by a variety of agencies having different definitions, standards and applications of the data. Solving the problem of semantic heterogeneity (e.g., the categorical land cover types in various land cover classification systems), is difficult, but very useful for information sharing. The intent of the different classification is mainly to reduce the information load by abstracting from details. For instance, there exists several land cover characterization schemes such as the International Geosphere Biosphere Programme (IGBP), United States Geological Survey (USGS), Olson Global Ecosystems (OGE), Simple Biosphere model (SiB), Simple Biosphere 2 (SiB2) model, Biosphere Atmosphere Transfer Scheme (BATS). Each of these classification systems differs in their purpose and level of aggregation. Also, the classified data could be available in multiple scales, i.e., a coarse scale or a finer scale. Thus, semantic conflicts arise when data from such sources are used in an overall decision-making scenario.

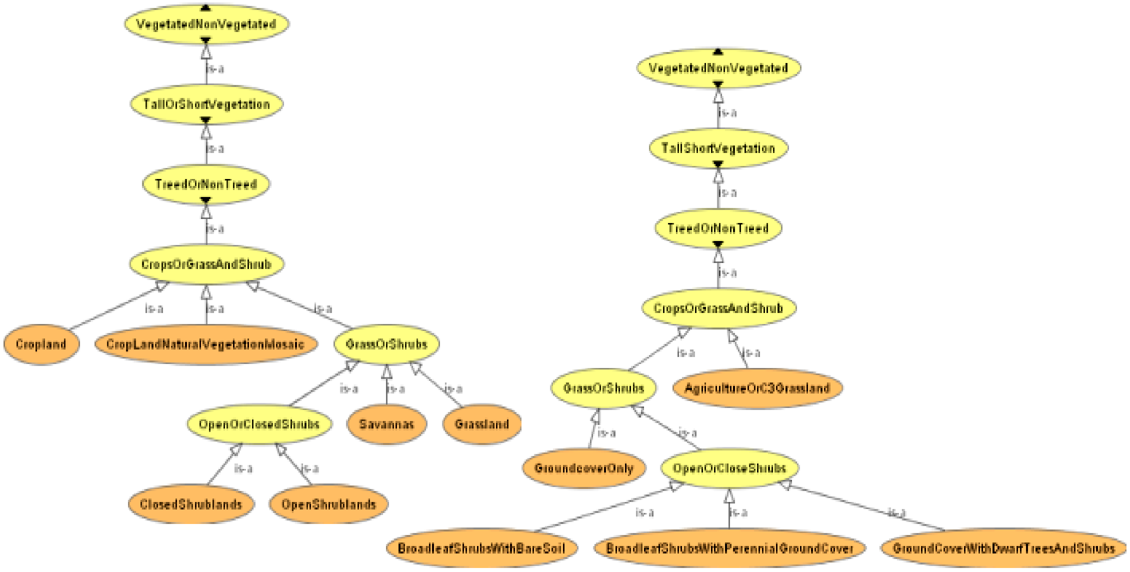


Figure 4.2 Semantic heterogeneities arising in terms of scaling, aggregation, and overlap between classification systems (IGBP and SiB)

A *naming conflict* is a commonly observed conflict in land cover classification schemes (e.g., the class *Evergreen Needleleaf forest* in an IGBP scheme corresponds to *Evergreen Needleleaf Trees* in the BAT Scheme and *Evergreen Needleleaf Vegetation* in the Running Vegetation life forms scheme). Similarly, the class *grassland* in the IGBP scheme corresponds to *Ground cover only* or *Ground cover with trees and shrubs* depending on the context in which it is used in the SiB scheme and *Annual grass vegetation* in the Running Vegetation Life forms scheme. Thus, it can be observed that the intended purpose of one classification scheme differs from another scheme. Although they are used to abstract similar details, they tend to produce vast heterogeneities. Figure 4.2 depicts the semantic heterogeneities arising in terms of scaling, aggregation, overlap, and naming conflicts between the IGBP classification scheme and the SiB classification scheme. Thus, it can be seen that there exists a semantic translation problem for integration of information sources that are in different classification systems. This is pursued as the motivating example to demonstrate how the emerging semantic web technologies can be adopted in the EO domain.

4.3 Integration Problem

The integration problem between disparate thematic data archives is finding the right data that matches a given criteria.

The above problem can be formally defined as [102]

Let $IS_1 = \langle\langle S_1, C_1, d_1\rangle, I_1, M_1\rangle$ and $IS_2 = \langle\langle S_2, C_2, d_2\rangle, I_2, M_2\rangle$ be information sources; then a bilateral integration problem is equivalent to finding a membership $M : I_1 \bigcup I_2 \times C_1$ such that for all $x \in I_1 \bigcup I_2$ and $c_i \in C_1$:

$$(x, c_i) \in M \quad iff \quad x : d_1(c_i) \tag{14}$$

Where S_1, S_2 are the source ontologies C_1, C_2 are set of class names and d is a mapping that assigns a class definition over the terms from S to every class term in C and I is a set of information items.

Consider two repositories in two different classification systems such as, IGBP (IS1) and SiB (IS2). Then, a sample query is finding all the data corresponding to broadleaf evergreen or deciduous forest in both the information sources. Such a query can be efficiently answered only if the semantics of both the information sources is well understood. However, even if both the information sources are sufficiently conceptualized by two distinct ontologies (e.g., IGBP and SiB), comparing them is rather a challenging task due to the great variation among the level of detail and logic of different ontology representations. In general two types of ontologies are prevalent; an ontology that is a collection of categories organized by a partial order that is induced by inclusion and a more detailed ontology called the axiomatized ontology which is a terminological ontology whose categories are distinguished by axioms and definitions stated in logic or in some language that could be automatically translated into logic [103]

The general approach for data integration in axiomatized ontologies is to map the local terms of distinct ontologies onto a single shared ontology. The Semantic similarity is then determined as a function of some distance relation between two terms in the hierarchical structure underlying the ontology [104-107]. Other systems build a shared ontology by integrating the existing ontologies [38] [63] [108-110].

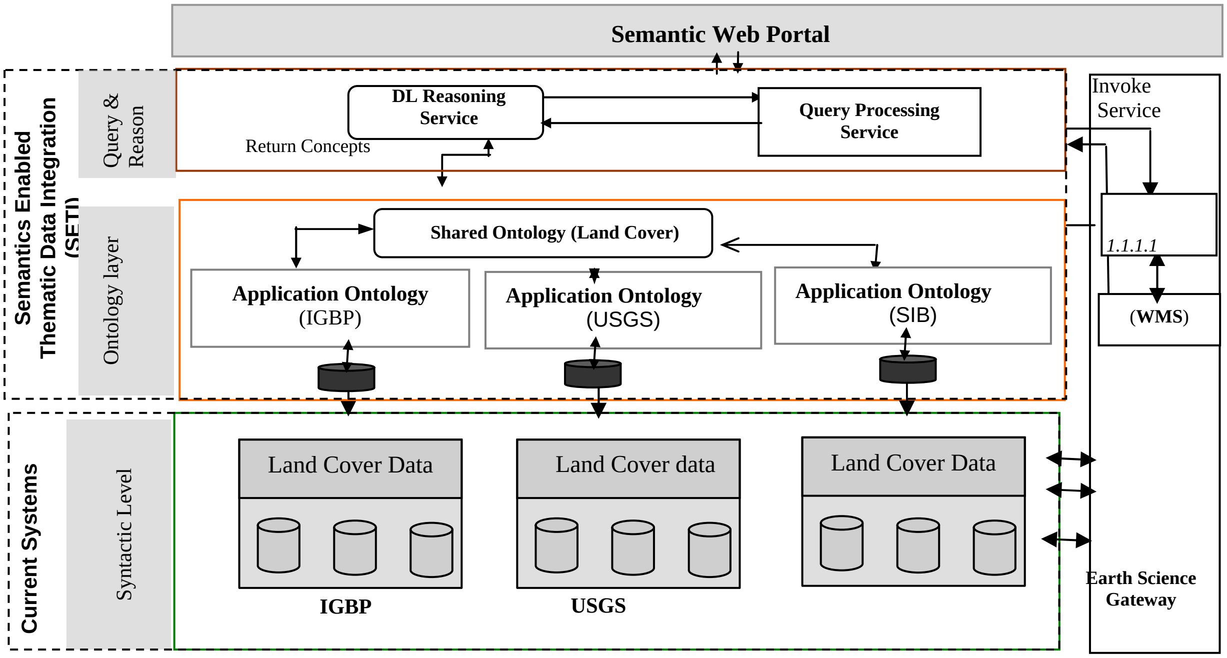


Figure 4.3 Framework for **S**emantics **E**nabled **T**hematic Data **I**ntegration (SETI)

This research pursues the shared ontology approach for the proposed framework by integrating the ontologies developed for each of the classification schemes (IGBP, BATS, etc.) and building a shared ontology for their integration.

4.4 Shared Ontology Approach

Assuming that ontologies are used to capture the context of the information entities, and then as we move from one context to another there is a requirement to integrate ontologies. In this work, we adopt a hybrid ontology approach (see Figure 4.3) consisting of a global shared ontology that encompasses all the local application-level ontologies for a domain of interest (land cover).

As shown in Figure 4.3, by using the application ontologies (e.g. IGBP, SIB, USGS), it is possible to perform terminological reasoning over the definition of classes in them by considering the set of axioms from the shared ontology, the definitions of relations and the set of class definitions [102]. A brief overview of the shared ontology development is provided in the next section.

The global land cover characterization has been completed for use in a wide range of continental to global scale environmental studies using the Advanced Very High Resolution Radiometer (AVHRR) Normalized Difference Vegetation Index (NDVI) composite covering 1-km AVHRR data spanning April 1992 through March 1993 [111]. This database provides a unique view of the broad patterns of the biogeographical and ecoclimatic diversity of the global land surface, and presents a detailed interpretation of the extent of human development [112]. The global database is available on a continent-by-continent basis [113]. In this study the North America Land cover characterization database is used to demonstrate the framework (i.e., SETI) presented in the previous section.

4.4.1 North America Land Cover Characterization

The North American land cover database is one portion of a global land cover characteristics database that was developed on a continent-by-continent basis. All continents in the global database share the same map projections (Interrupted Goode

Homolosine and Lambert Azimuthal Equal Area) and have 1-km nominal spatial resolution.

Table 4.1 Derived global land cover data sets in the Global Land cover database(adapted from [112])

Classification scheme	**Number of classes**	**Intended application**
Olson global Ecosystems (Olson. 1994)	94	Carbon Cycle Studies
IGBP DISCover Land Cover Legend (Belward. 1996)	17	Global Change
Biosphere-Atmosphere Transfer Scheme (BATS) (Dickinson et al. 1986)	20	Land Atmosphere Interactions (Climate models)
Simple Biosphere Model (SiB) (Sellers et al, 1986)	16	Land-atmosphere interactions (Climate models)
Simple Biosphere Model2 (SiB2) (Sellers et al. 1996)	10	Land-atmosphere interactions (Climate models)
USGS Land Use/Land Cover System (Anderson at al. 1976)	24	Multi-purpose
Global Remote Sensing Land cover (Running et al . 1994)	8	Biogeochemical modeling

The database consists of 7,793 rows and 11,329 columns and the core attributes of each of the seasonal land cover are [112]:

- Land cover descriptions
- Seasonal characteristics
- Site characteristics (elevation ranges, biome representation, and other relevant local descriptors)
- Multi-temporal NDVI statistics.

Ontologies have been developed for each of the classification schemes as depicted in Table 4.1. The core attributes of the seasonal land cover have been used to define object properties (e.g., hasBiome, hasBiomeCode, hasStructure, hasFoliage) for each land cover class. Below is provided a brief description of the ontology development process and also the steps involved in the shared ontology development, which is necessary in this framework for semantic interoperability.

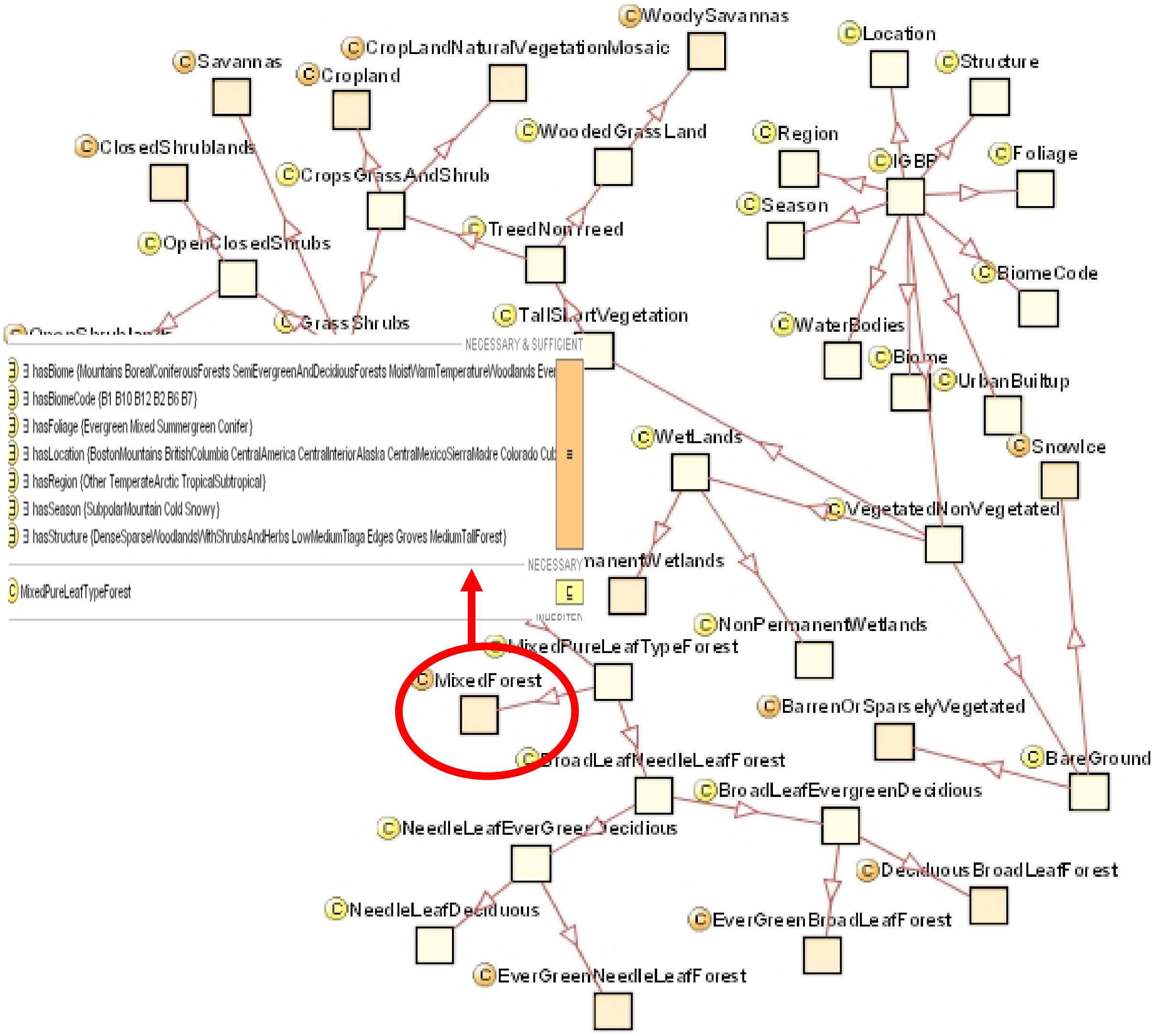
4.4.2 Ontology Development

4.4.2.1 Web Ontology Language (OWL)

The Web Ontology Language (OWL) is a current W^{3}C standard for developing ontologies. OWL has three versions: OWL lite, OWL DL, and OWL full. Each of these versions caters to the different requirements of the users and is a function of its expressiveness. The OWL DL version to used to develop the ontologies, which provides maximum expressiveness, without losing computational completeness (all entailments are guaranteed to be computed) and decidability (all computations will finish in finite time) of reasoning systems [114]. OWL DL is so named due to its correspondence with Description Logics (DL), a field of research that has studied a particular decidable fragment of first order logic.

The open source ontology and knowledge base editor Protégé [47] has been employed to develop the ontologies. Exploration of ontologies at various levels of granularity necessitates defining classes by restricting their property values. Then, by combination of various restrictions, they are inherited into subclasses. The combinations of these restrictions define all conditions that must hold for individuals of the given class

[4]. Several ontologies were developed for different land cover schemes (Figures 4.2-4.6). Each of these land cover schemes are selected for a particular project based on the project-specific goals and routinely much of the information regarding the land cover status of a region is disseminated though these land cover schemes. Hence, achieving interoperability between these land cover classification schemes is challenging. Thus, the conceptualization of these schemes in an ontology would provide a distinct way to understand the actual meaning of a class in a particular scheme and hence, help to identify similar class in another classification scheme. As can be seen in the Figures 4.4-4.8, there are several classes that have subtle differences in terms of meaning, interpretation, scaling, and naming attributes. So it is important to conceptualize the intended definition of what we mean by a particular concept. As shown in Figure 4.4, a defined concept is obtained by defining the necessary and sufficient conditions.



∃ hasBiome {Mountains BorealConiferousForests SemiEvergreenAndDecidiousForests MoistWarmTemperatureWoodlands Ever

∃ hasBiomeCode {B1 B10 B12 B2 B6 B7}

∃ hasFoliage {Evergreen Mixed Summergreen Conifer}

∃ hasLocation {BostonMountains BritishColumbia CentralAmerica CentralInteriorAlaska CentralMexicoSierraMadre Colorado Cub

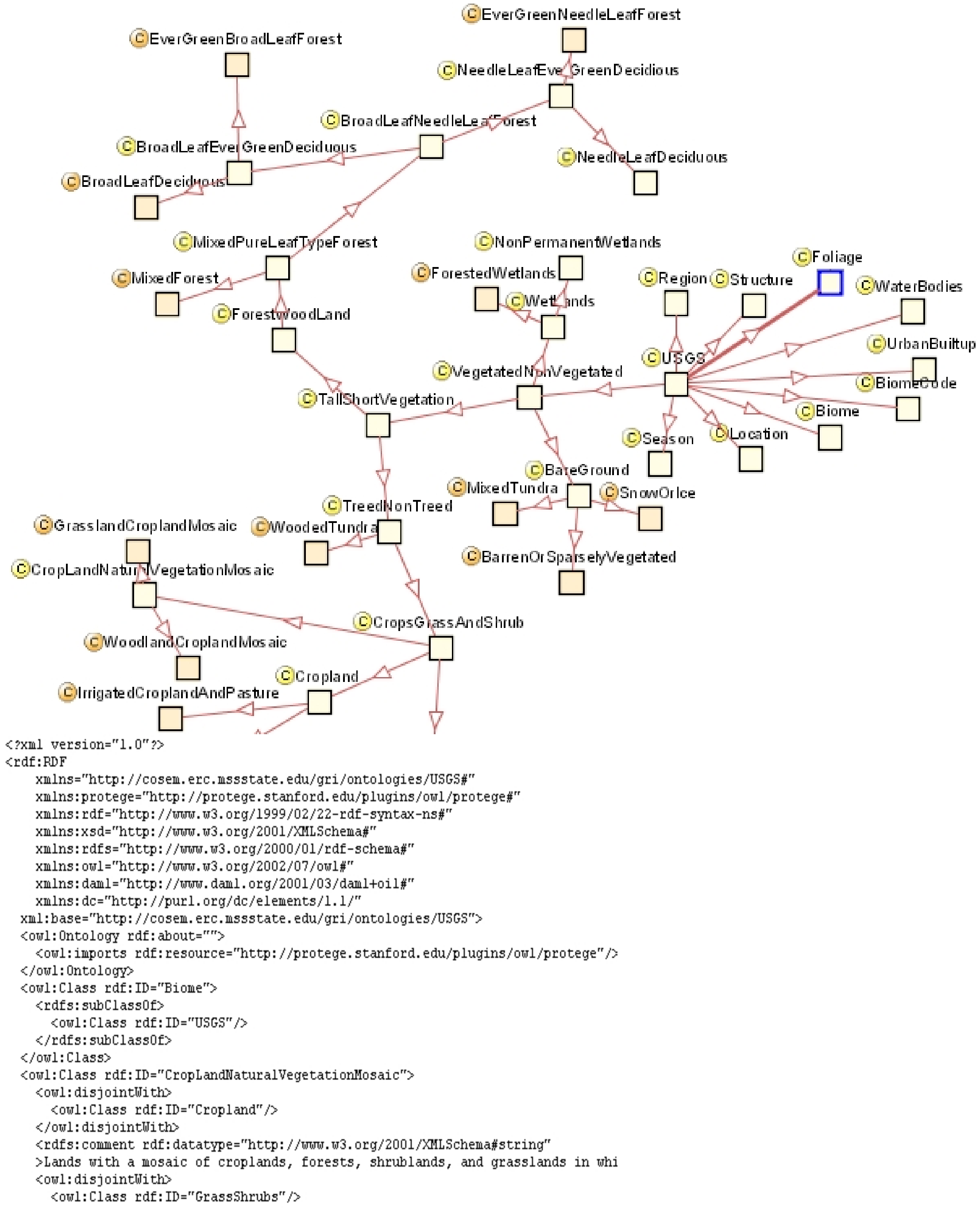
∃ hasRegion {Other TemperateArctic TropicalSubtropical}

∃ hasSeason {SubpolarMountain Cold Snowy}

∃ hasStructure {DenseSparseWoodlandsWithShrubsAndHerbs LowMediumTiaga Edges Groves MediumTallForest}

MixedPureLeafTypeForest

Figure 4.4 International Geosphere Biosphere Programme(IGBP) Land Cover



```xml
<?xml version="1.0"?>
<rdf:RDF
    xmlns="http://cosem.erc.msstate.edu/gri/ontologies/USGS#"
    xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
  xml:base="http://cosem.erc.msstate.edu/gri/ontologies/USGS">
  <owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
  </owl:Ontology>
  <owl:Class rdf:ID="Biome">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="USGS"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="CropLandNaturalVegetationMosaic">
    <owl:disjointWith>
      <owl:Class rdf:ID="Cropland"/>
    </owl:disjointWith>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Lands with a mosaic of croplands, forests, shrublands, and grasslands in whi
    <owl:disjointWith>
      <owl:Class rdf:ID="GrassShrubs"/>
```

Figure 4.5 United States Geological Survey(USGS) Land Use/Land Cover System

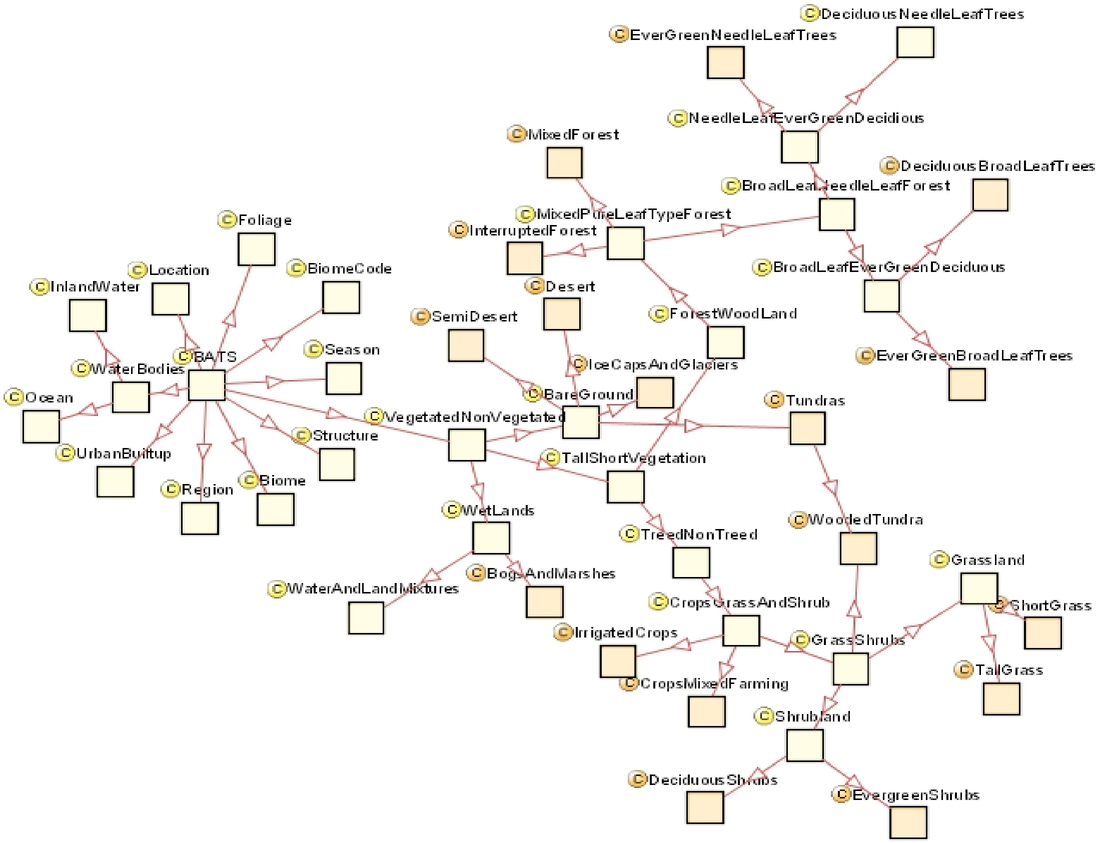


Figure 4.6 Biosphere-Atmosphere Transfer Scheme (BATS)

During the process of running the classifier to check for consistency and classification of the taxonomy in the development of the OGE ontology, it has been observed that the classifier has been able to find additional characteristics that resulted in an inferred hierarchy as shown in Figure 4.9 The concepts *Dry Tropical Woods*, *Tropical Rain Forest*, *Tropical Degraded Forest* and *Seasonal Tropical Forest* were originally modeled as sub concepts of *broadleaf evergreen or deciduous* class, after running the classifier, it checked for the concept definitions and provided an inferred hierarchy.

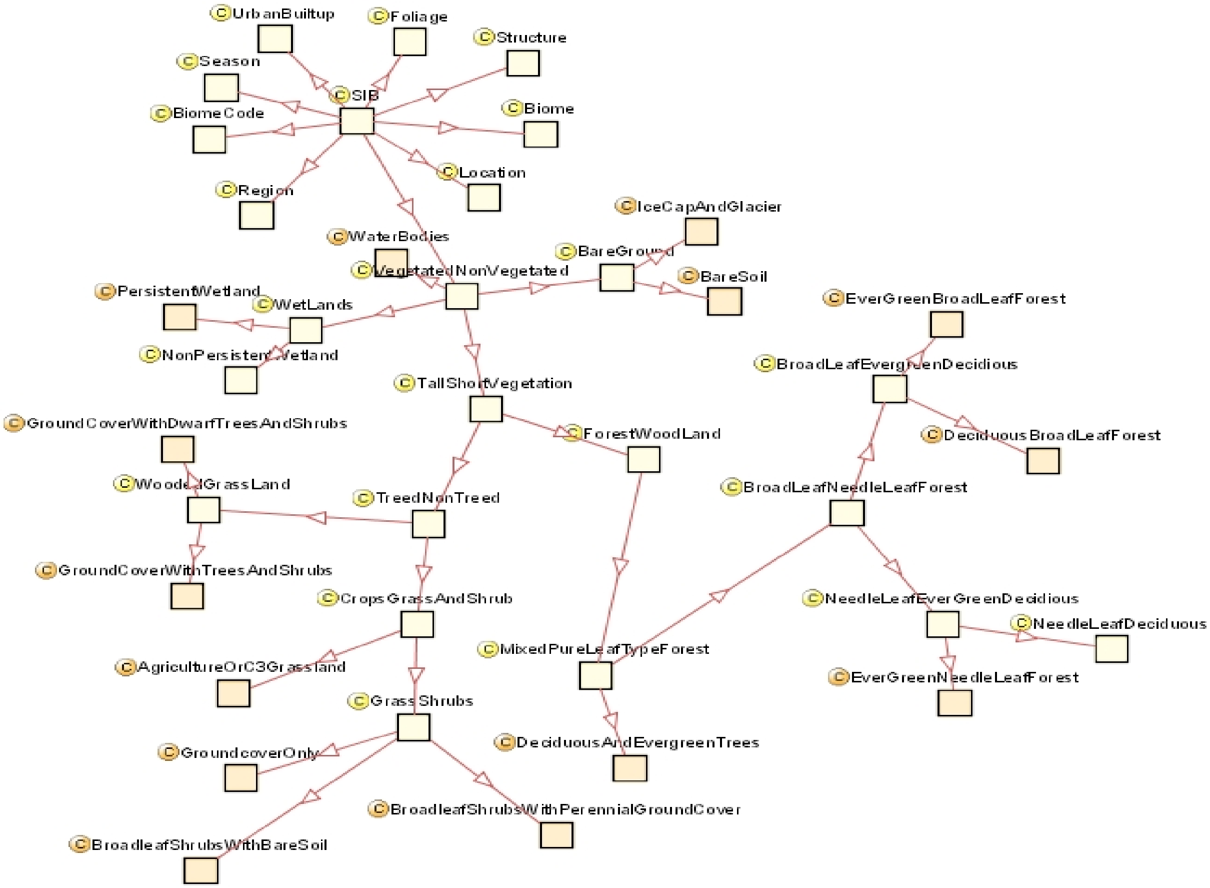


Figure 4.7 Simple Biosphere Model

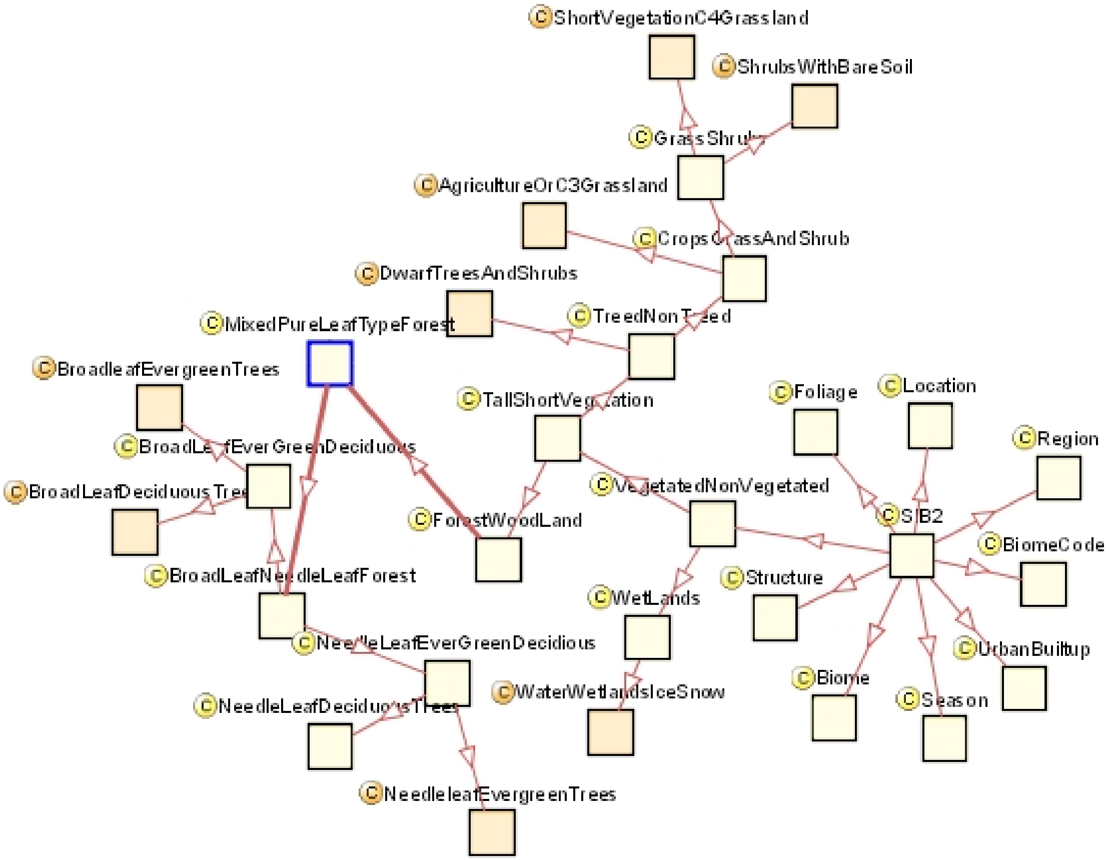


Figure 4.8 Simple Biosphere 2 Model

Such modeling of land cover concepts in OWL-DL helps in reasoning such as classification and retrieval by a description logic reasoner [48].

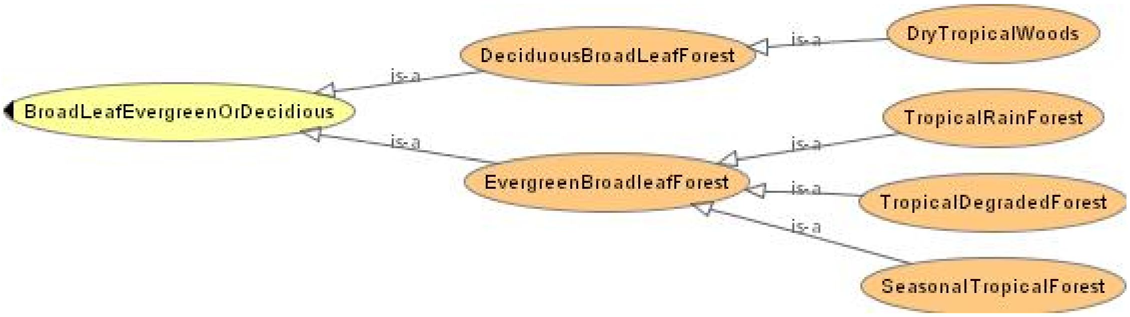


Figure 4.9 Inferred hierarchy returned by the classifier for OGE ontology

4.4.2.2 Shared Ontology Development

The classification systems described earlier overlap and complex cases of semantic heterogeneity as discussed previously arise. Due to their informal and specific character, the standards specifying the semantics of the terminologies are not powerful enough to resolve those heterogeneities. For the integration of these classification systems a separate, more expressive terminology is required. The semantics of this terminology may be specified by a logic-based ontology, which then is called a shared ontology or reference ontology. The semantics of the reference ontology may be specified by a standard, which is often called a meta-standard [115].

The Olson Global Ecosystems scheme is used as the starting point for developing the shared ontology because it [112]: (1) has sufficient thematic detail (94 potential classes) and was developed for global applications; (2) has been used for large area modeling and has links to landscape productivity, particularly carbon stocks; (3) recognizes anthropogenic elements of the landscape; (4) recognizes landscape mosaics that occur at coarse resolutions; and (5) includes attributes on climate and physiognomy, and implies floristic elements.

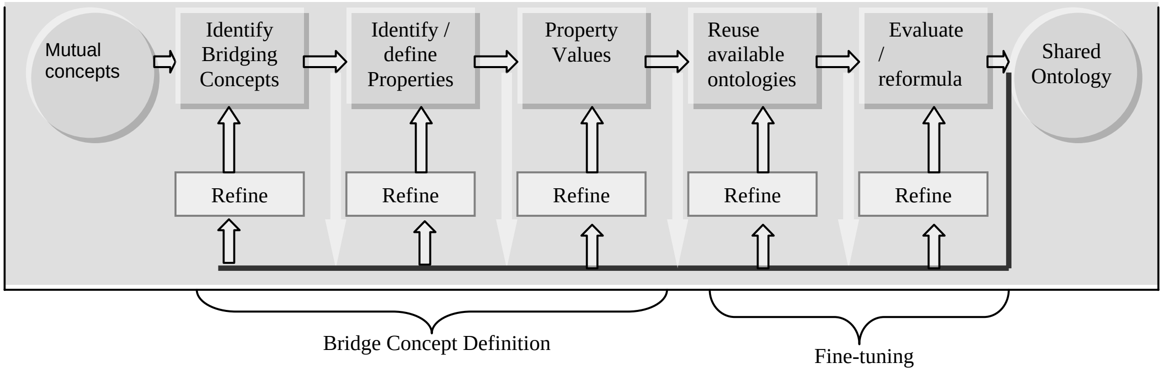


Figure 4.10 Evolutionary prototyping lifecycle in a shared ontology development

The bridging concepts have been identified between each of the application level ontologies and are then used to define a more general defined concept that sufficiently describes the common concepts. The shared ontology is a very general ontology that covers all possible refinements (Figure 4.10).

Drawing parallels from the software engineering field, the development of the shared ontology normally follows an evolutionary prototyping life cycle (Figure 4.9). In this life cycle, one can go back from any stage to any stage of the development process [49]. The evolutionary prototyping approach dynamically responds to changes in user needs and accommodates subsequent unpredictable changes in requirements, as the development process progresses [116]. As long as the ontology does not satisfy evaluation criteria and does not meet all requirements during specification, the prototype is modified.

4.4.2.3 Concept- Based Retrieval and Integration of Thematic Information

The shared ontology approach presented in the previous section provides the necessary framework for DL-based reasoning across heterogeneous information sources. The task of finding a set of classes satisfying the query in the information source is based on the retrieval of direct subclasses and super classes in the shared ontology as we have only that knowledge available, since it has already been classified. The direct super classes are retrieved when the concept has negation in it (e.g. retrieve all forests that are *not* mixed forest) otherwise the direct subclasses are retrieved. The reclassification of the information item for one information source into the application ontology of another source can be formalized as [102] [117].

Let $IS_1 = \langle\langle S, C_1, d_1\rangle, I_1, M_1\rangle$ and $IS_2 = \langle\langle S, C_2, d_2\rangle, I_2, M_2\rangle$ be information sources, S is the shared ontology, x is an information item, and $x \in I_2$, d is a mapping that assigns a class definition over the terms from S to every class term in C, then for every $c_1 \in C_1$, we can define $M^{'} : I_2 \times C_1 \rightarrow \{0,1,?\}$ (approximate classifier) such that:

$$M^{'}(x, c_1) = 1 \ \ if \ \ x : \left(\bigvee_{c \in \mathrm{glb}\, IS2} d2(c) \right) \tag{15}$$

$$M^{'}(x, c_1) = 0 \ if \ \ x : \neg\left(\bigwedge_{c \in \mathrm{lub}\, IS2} d2(c) \right) \tag{16}$$

$$M^{'}(x, c_1) = ?\, , \ otherwise \tag{17}$$

Thus using the above greatest lower bounds (glb) and lowest upper bounds (lub) depending on the query concept whether it is a negation or otherwise, it is possible to retrieve information from heterogeneous sources, by considering their semantic descriptions. Using a DL reasoner which allows classification of data from one context to another by equality and subsumption (Subsumption means that if concept B satisfies the requirements for being a case of concept A, then B can automatically be classified below A [49]), the above procedure was adopted to search each concept in the application ontology belonging to an information source with the subsumed concepts in the shared ontology to check if they satisfy the concept definitions and then retrieve those concepts that match the semantics. For example, a query on retrieval of Broadleaf Evergreen or Deciduous type forests from two different information sources such as IGBP and SiB would return results containing the actual concepts that have been searched in application ontologies related to each of these thematic data repositories (e.g. IGBP and SiB). The DL based reasoner uses the definition of concepts from the shared ontology which essentially are the subsumed concepts of the query concept and then searches for all the concepts that satisfy the criteria in each of the application ontologies.

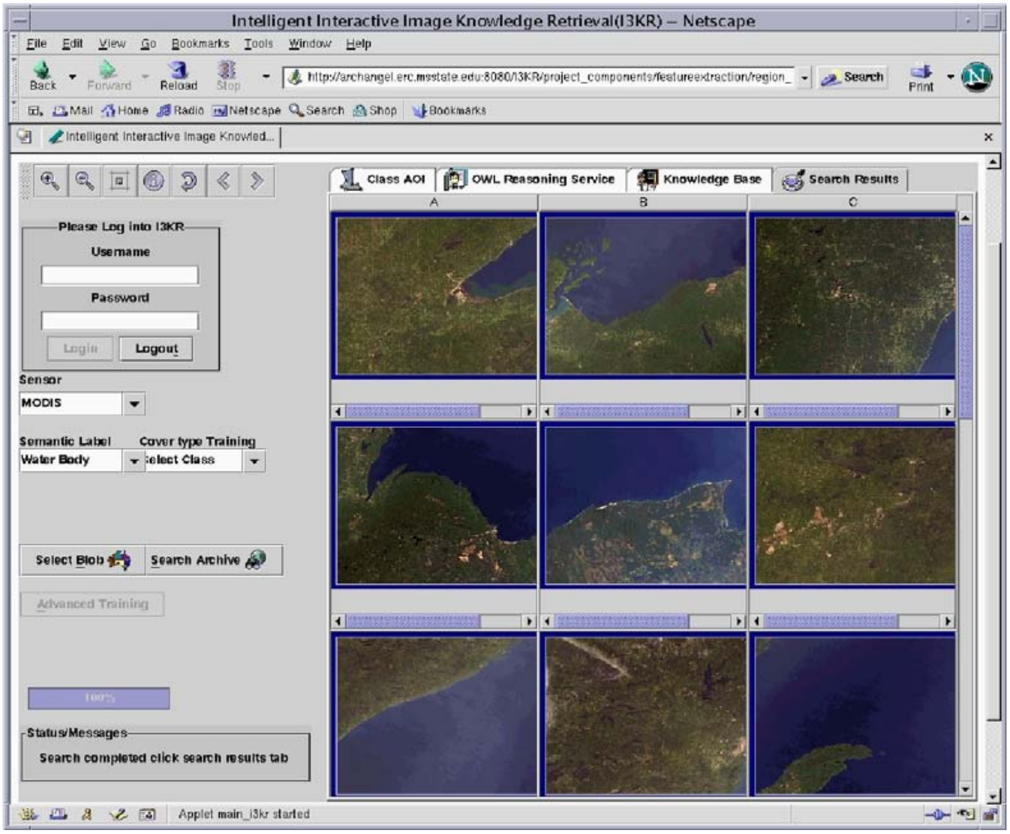
The retrieval of instances of the concepts that satisfied the query is the final result of this process. This procedure enables query processing and searching in a way not possible with keyword-based searches.

CHAPTER V

RESULTS

This chapter presents the results from the frameworks described in chapters 4 & 5.

5.1 Results of the Intelligent Interactive Image Knowledge Retrieval (I3KR) Framework



Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape

http://archangel.erc.msstate.edu:8080/I3KR/project_components/featureextraction/region_

Intelligent Interactive Image Knowled...

Sensor: MODIS

Semantic Label: Water Body

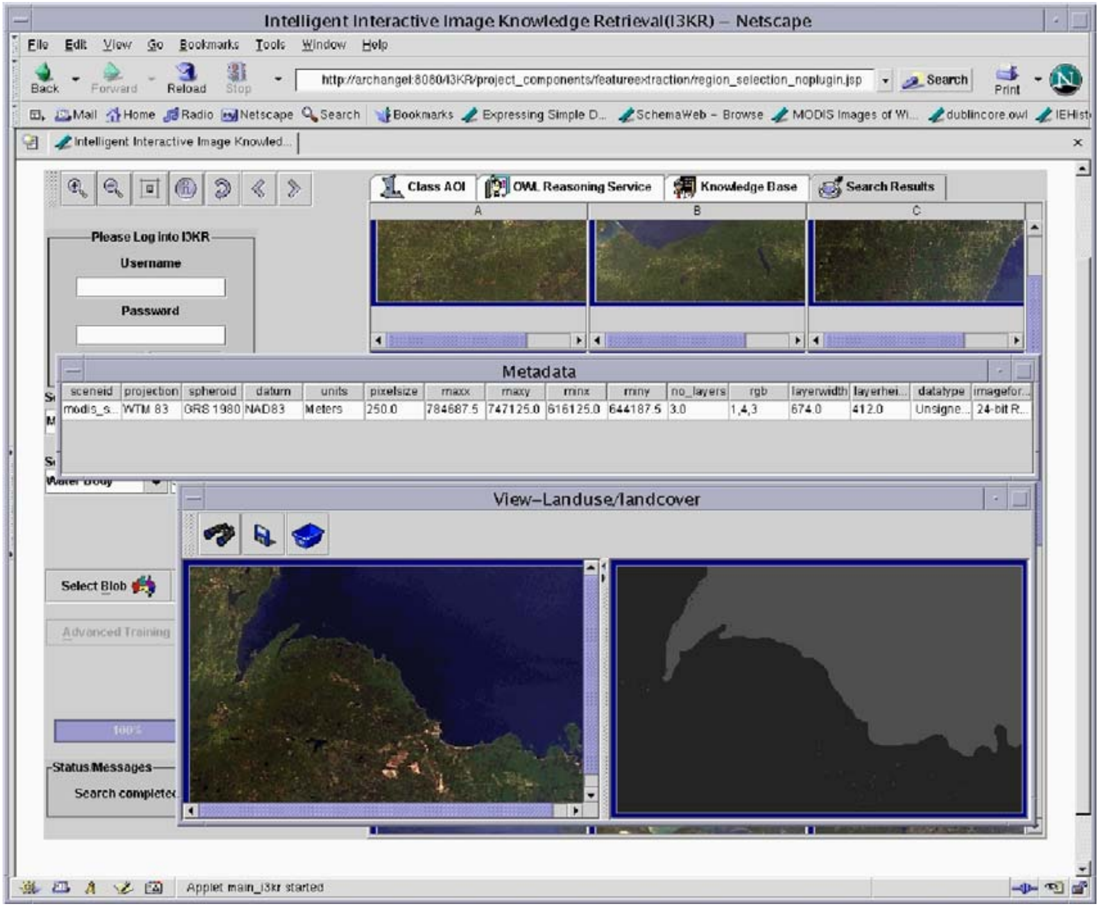
Cover type Training: :elect Class

Status/Messages: Search completed click search results tab

Applet main_i3kr started

Figure 5.1 Results of a semantic query

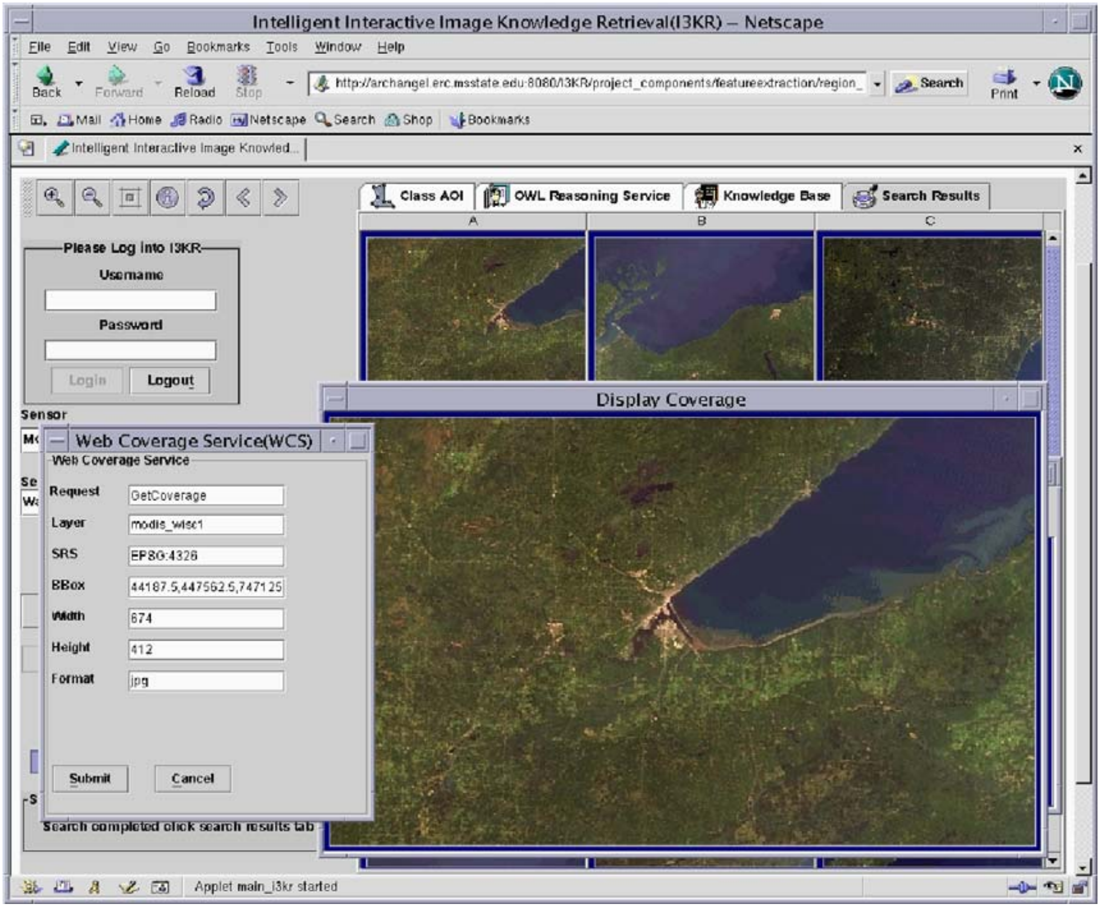
Both the systems were implemented in JAVA [118]. The user interface is provided through an applet that runs in a browser. The I^{3}KR system provides a number of modules (including reasoning services, Area of Interest (AOI) selection, and knowledge base browsing and querying) that have been integrated a GUI.


Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp
Search
Print
Mail Home Radio Netscape Search Bookmarks Expressing Simple D... SchemaWeb – Browse MODIS Images of Wi... dublincore.owl IEHist
Intelligent Interactive Image Knowled...
Class AOI
OWL Reasoning Service
Knowledge Base
Search Results
A
B
C
Please Log into I3KR
Username
Password
Metadata
Water Body
View–Landuse/landcover
Select Blob
Advanced Training
100%
Status Messages
Search complete
Applet main_i3kr started

sceneid	projection	spheroid	datum	units	pixelsize	maxx	maxy	minx	miny	no_layers	rgb	layerwidth	layerhei...	datatype	imagefor...
modis_s...	WTM 83	GRS 1980	NAD83	Meters	250.0	784687.5	747125.0	616125.0	644187.5	3.0	1,4,3	674.0	412.0	Unsigne...	24-bit R...

Figure 5.2 I^{3}KR system depicting metadata of the image and also more details can be seen about the retrieved area of interest

The user is provided with an integrated environment where it is possible to interact with the system in a variety of ways. For example, the user can execute a semantic query and visualize the results (Figure 5.1) or they can browse the existing knowledge base and then look for concepts that are relevant to their conjecture.


Intelligent Interactive Image Knowledge Retrieval(I3KR) – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://archangel.erc.msstate.edu:8080/I3KR/project_components/featureextraction/region_
Search
Print
Mail Home Radio Netscape Search Shop Bookmarks
Intelligent Interactive Image Knowled...
Class AOI
OWL Reasoning Service
Knowledge Base
Search Results
A
B
C
Please Log into I3KR
Username
Password
Login
Logout
Sensor
Display Coverage
Web Coverage Service(WCS)
Submit
Cancel

Web Coverage Service

Request GetCoverage

Layer modis_wisc1

SRS EPSG:4326

BBox 44187.5,447562.5,747125

Width 674

Height 412

Format jpg

Search completed click search results tab

Applet main_i3kr started

Figure 5.3 Retrieval of the images from the archive through Web Coverage Service (WCS)

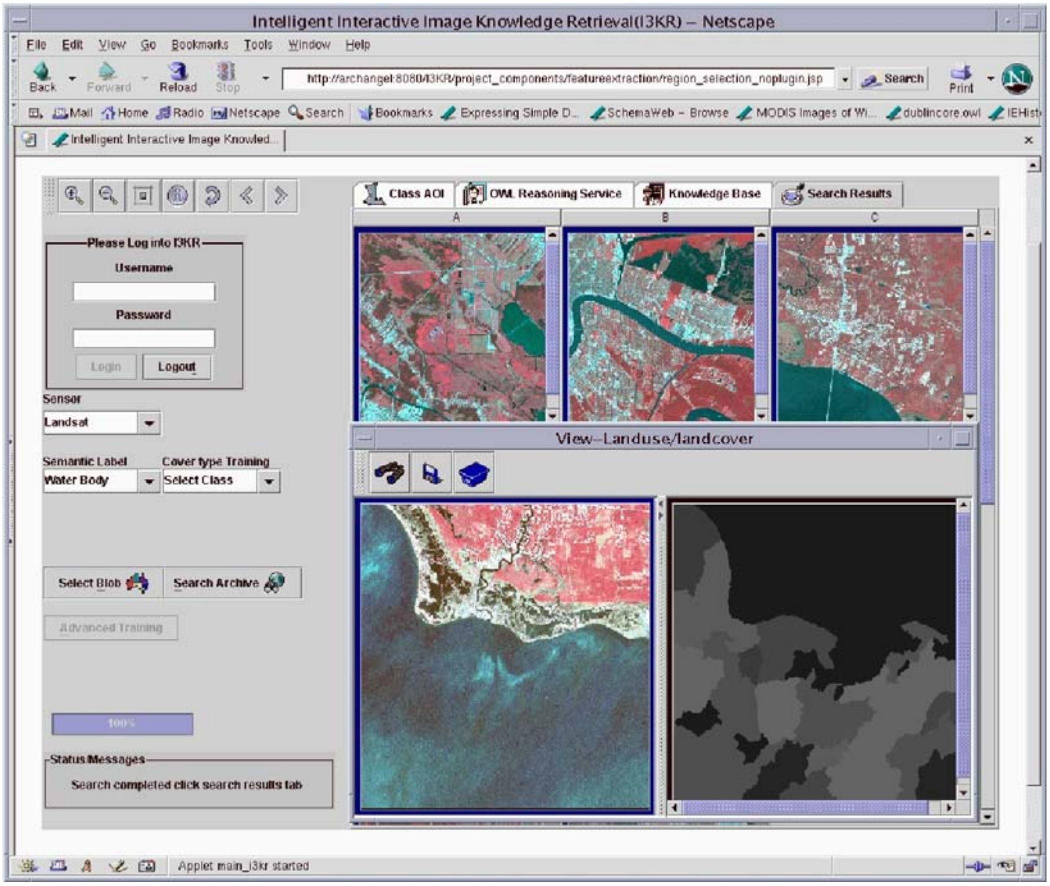
In addition, since the user might not know the exact semantics of the information that they are looking for, the exploration of the ontology through the concept query interface gives the ability to search and explore at different conceptual levels.

The system also provides functionalities to store user-learned knowledge. Currently the image archive consists of tiled images from MODIS and Landsat sensors.

Figure 5.1 depicts the retrieved images from a semantic query about what MODIS imagery in the archive contained water bodies. It is possible to explore the results of the query further by clicking on the image of interest. An image view window opens which depicts more details of the cover type of interest along with the metadata (Figure 5.2).

The OGC Web Coverage Service (WCS) (Figure 5.3) provides the user with a service that has the capability to extract only the necessary data that meet their requirements. This also enables search of distributed archives and helps alleviate data transfer bottlenecks over the network. The user can explore the full scene interactively by passing the WCS parameters like Spatial Reference System (SRS), Bounding Box (BB), width, height and format (jpg, tiff, Geo-tiff etc). Once the knowledge has been discovered by mining through the archives, the WCS can also be used to facilitate decision-making by analyzing data from multiple sensors (e.g., MODIS, Landsat) at different resolutions of the same region by separate requests to distributed archives (e.g., NASA, NOAA).

Figure 5.4 depicts the retrieval of images from a Landsat data archive. Finer details within a cover type (water body) are evident. These various levels of segmentation help an analyst in knowledge discovery. Figures 5.5 & 5.6 depict the results from a semantic query on agriculture and forest, respectively.



http://archangel:8080/I3KR/project_components/featureextraction/region_selection_noplugin.jsp

Sensor: Landsat

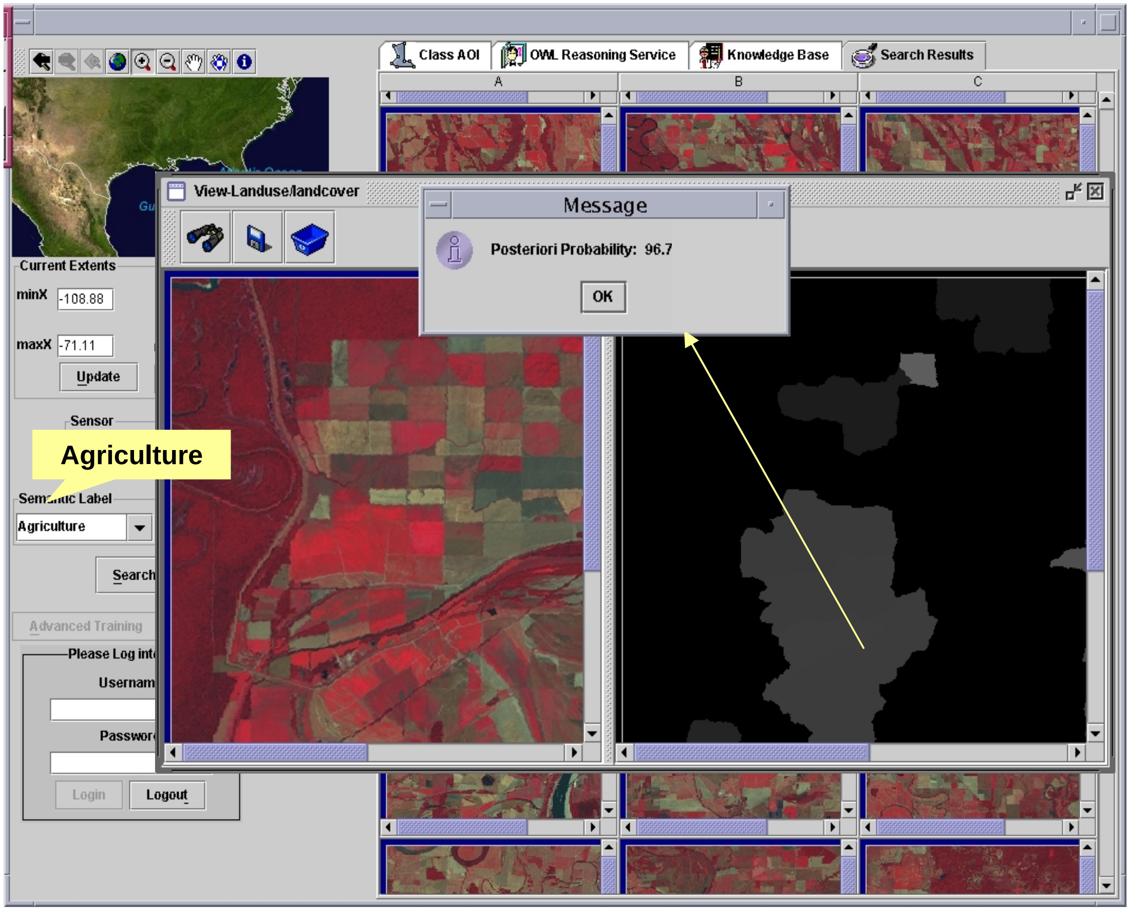
Semantic Label: Water Body

Cover type Training: Select Class

Status Messages: Search completed click search results tab

Applet main_i3kr started

Figure 5.4 Varying levels of segmentation details within a cover type



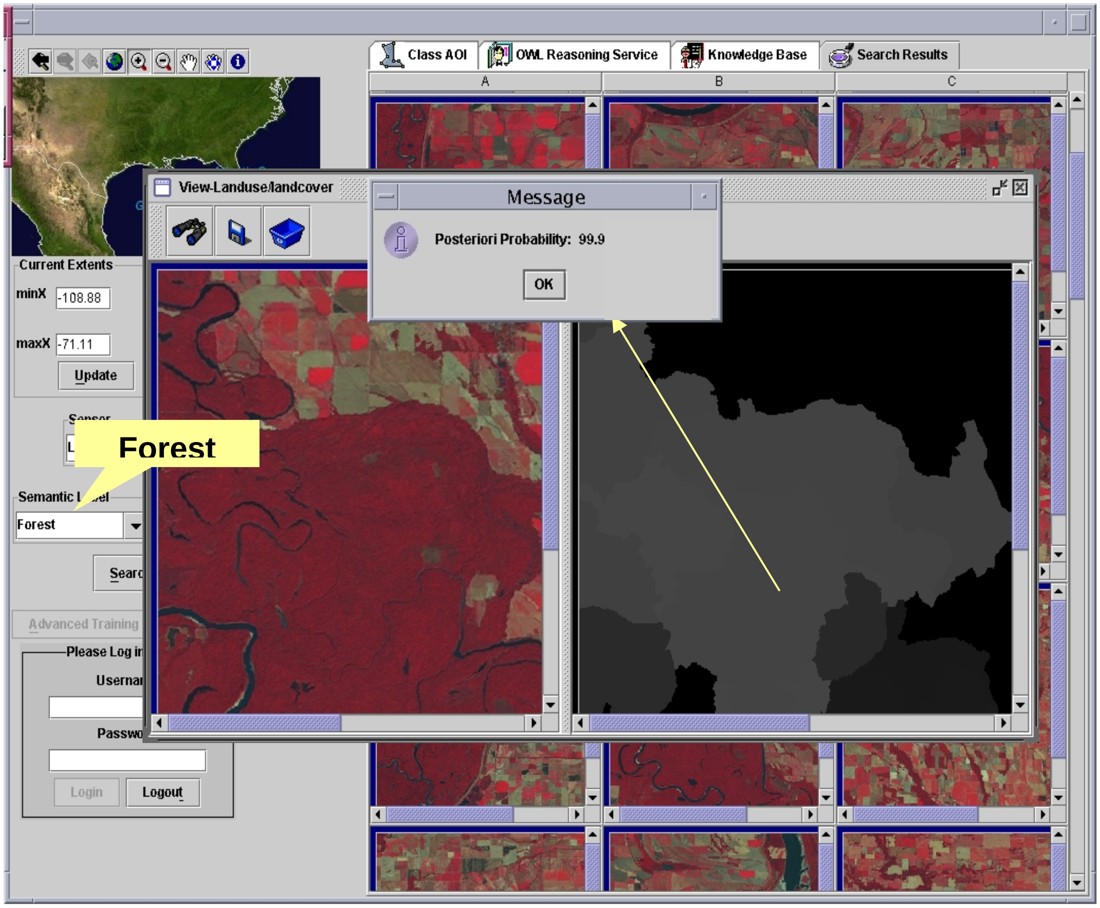
Posteriori Probability: 96.7

minX -108.88

maxX -71.11

Semantic Label Agriculture

Figure 5.5 I^{3}KR system depicting the results of a query on agriculture



Posteriori Probability: 99.9

minX -108.88

maxX -71.11

Semantic Label: Forest

Figure 5.6 I^3KR system depicting the results of a query on forest

5.2 Results from GA- Based Feature Selection and Feature Generation

Several experiments were conducted to formulate the semantic models. In the wrapper-based approach of feature selection/generation the performance evaluation was conducted using two nested cross-validations. The inner cross validation was used to find an optimal subset of features and the outer cross-validation was used to evaluate the performance of this subset of features.

The genetic algorithm parameters were set to 100 for the number of generations, 15 for population size, 0.5 for the crossover probability, 0.2 for the mutation (feature selection probability), and 0.5 for the feature generation probability. The induction algorithm used was a SVM for classification with complexity C=1000, epsilon=0.1 and using a RBF kernel. A recent study by Keerthi and Lin shows that if RBF is used with model selection, then there is no need to consider the linear kernel [61]. The kernel matrix using a sigmoid may not be positive definite and in general its accuracy is not better than RBF.

Table 5.1 Accuracy, precision, recall and F-measure obtained using only feature selection by GA

Class	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Water bodies	98.80	93.80	87.76	90.53
Agriculture	97.63	98.83	98.18	98.50
Fallow Land	96.49	98.38	94.81	96.56
Forest	98.53	97.84	94.44	96.11
Flooded vegetation	91.04	89.83	100	94.64
Flooded fastlands	96.92	93.75	83.33	88.24

Table 5.2 Features selected by GA (Only Feature Selection)

Class	Features Selected by GA
Agriculture	*Color:* ColorDescriptor1,ColorDescriptor2, *Texture(Cooccurence):* First order inverse element, uniformity *Texture(primitive length):* Gray level Uniformity *Shape:* Geometric Moment
Water bodies	*Color:* ColorDescriptor1, Colordescriptor2, *Texture(Cooccurence):* First order element, entropy, uniformity, *Texture(primitive length):* long primitive emphasis, primitive percentage
Flooded vegetation	*Color:* Colordescriptor1,colordescriptor3 *Texture(Cooccurence):* entropy *Texture(primitive length):* short primitive emphasis, long primitive emphasis, primitive percentage *Shape:* eccentricity
Fallow	*Color:* ColorDescriptor1, ColorDescriptor2,ColorDescriptor3 *Texture(Cooccurence):* entropy
Forest	*Color :* ColorDescriptor1,Colordescriptor2 *Texture(Cooccurence):* Maximum probability, first order element *Shape:* eccentricity
Flooded fastlands	*Color:* Colordescriptor1,Colordescriptor2,Colordescriptor3, *Texture(Cooccurence):* Maximum probability *Texture(primitive length):* Short primitive emphasis Uniformity *Shape:* eccentricity

Table 5.3 Accuracy, precision, recall and F-measure obtained using both feature selection and generation by GA

Class	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Agriculture	97.76	98.84	98.37	98.59
Fallow Land	97.43	98.67	96.36	97.50
Forest	96.80	92.86	90.28	91.55
Flooded vegetation	92.54	92.86	98.11	95.41
Flooded fastlands	96.12	87.50	82.35	84.85
Water bodies	99.07	95.56	89.58	92.47

Table 5.4 Features selected by GA (Feature Selection and generation) Note: only three features are shown here.

Class	Selected features	Features Generated from
Agriculture	*Color:* Colordescriptor1	*Color:* Colordescriptor1
	Color: Colordescriptor2	*Color:* Colordescriptor2
	Texture (Cooccurence): First order inverse element	*Texture (Cooccurence):* First order inverse element
	Texture (Primitive Length): Short primitive emphasis	*Texture (Primitive Length):* Short primitive emphasis
	Texture (Primitive Length): Gray level uniformity	*Texture (Primitive Length):* Gray level uniformity
	Gen1	*(geometric moment, maximum probability)
	Gen2	+(first order inverse element, uniformity)
	Gen3	*(*(geometric moment, maximum probability)*(geometric moment, maximum probability))
Water bodies	*Color:* Colordescriptor1	*Color:* Colordescriptor1
	Color: Colordescriptor2	*Color:* Colordescriptor2
	Texture (Cooccurence): First order inverse element	*Texture (Cooccurence):* First order inverse element
	Texture (Cooccurence): First order element	*Texture (Cooccurence):* First order element
	Texture (Primitive Length): Short Primitive emphasis	*Texture (Primitive Length):* *Short Primitive emphasis*
	Shape: Geometric Moment	*Shape:* Geometric Moment
	Gen1	+(primitive percentage, +(primitive percentage, *(entropy, long primitive emphasis)))

Table 5.4 (continued)

Class	Selected features	Features Generated from
Flooded vegetation	*Texture (Primitive Length):* Long primitive emphasis	*Texture (Primitive Length):* Long primitive emphasis
	Gen1	+(*(*(geometric moment, gray level uniformity), uniformity), +(*(*geometric moment, gray level uniformity), uniformity), long primitive emphasis))
	Gen2	+(*(*(geometric moment, gray level uniformity), uniformity), long primitive emphasis)
	Gen3	+(+(*(*geometric moment, gray level uniformity), uniformity)+(+(+(geometric moment gray level uniformity), uniformity), long primitive emphasis)), +(+(geometric moment, gray level uniformity), uniformity))
	Gen4	+(long primitive emphasis, *(+(*(*(geometric moment, gray level uniformity),uniformity), +(*(*(geometric moment, gray level uniformity), uniformity), long primitive emphasis)), *(*(geometric moment, gray level uniformity), uniformity), long primitive emphasis)), *(*geometric moment, gray level uniformity), uniformity)))

5.2.2 Precision, Recall and F-measure

In content-based image retrieval (CBIR), *recall* and *precision* measurements are most often used to illustrate how many relevant (target) and irrelevant (misdirected) images are contained in the highest ranked images [1]. In region-based image information mining, instead of accounting for the number of images retrieved, it is more relevant to account for the number of correct regions that are retrieved, which indirectly would correspond to the images retrieved. Hence, precision in this case is the proportion of relevant regions to all the regions retrieved. If R is the set of returned regions and S the set of regions relevant to the query, then:

$$precision = \frac{|S \cap R|}{|R|} \qquad (18)$$

Recall is the proportion of relevant regions that are retrieved, out of all relevant regions

$$recall = \frac{|S \cap R|}{|S|} \qquad (19)$$

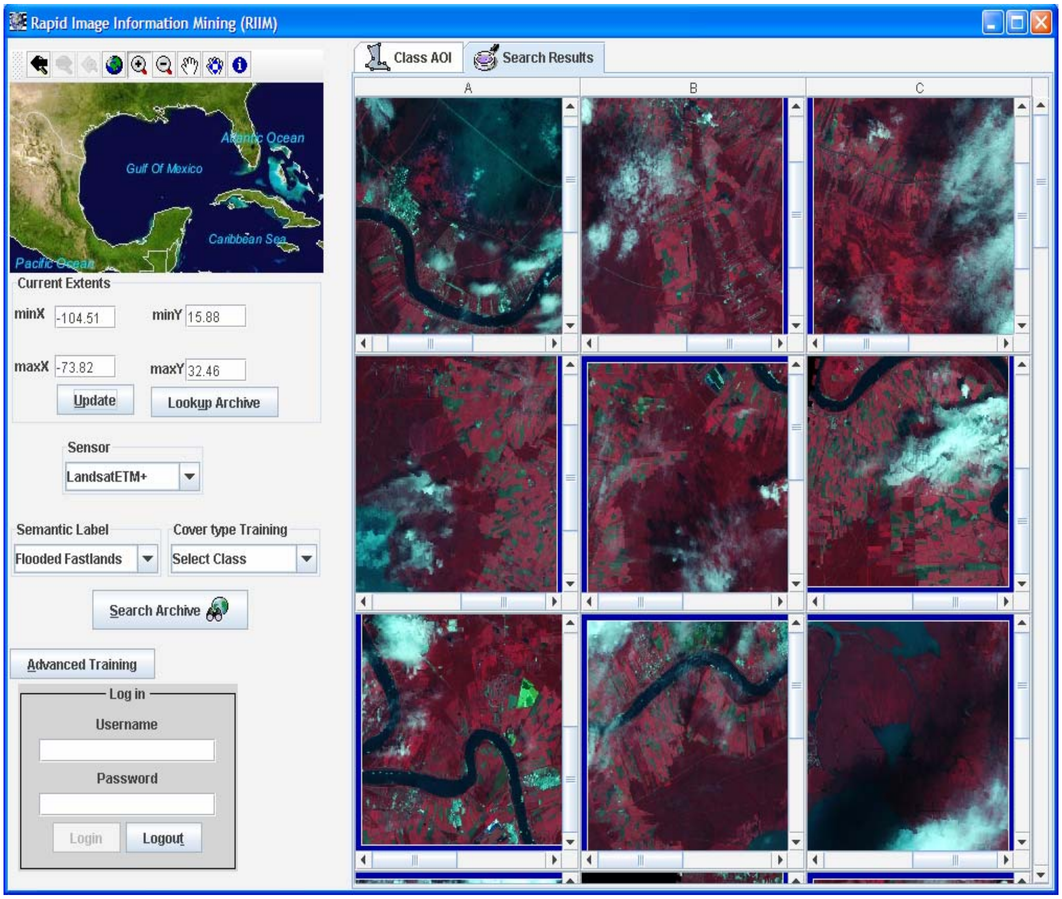
High precision indicates that most of the items you retrieve are relevant**.** High recall indicates that you have retrieved most of the available relevant regions in a repository. The F-measure is the weighted harmonic mean of precision and recall given by

$$\text{F-measure}=\frac{2 \times precision \times recall}{precision + recall} \qquad (20)$$

In the first set of experiments the GA-based wrapper approach was used only for feature selection for the six semantic classes. Table 5.1 depicts the feature selection results; precision, recall and, F-measure values using the reduced feature set as obtained from the GA. The accuracy of the retrieval is measured in terms of the relative number of correctly classified examples. From an initial set of 15 features, the GA algorithm produced an optimal feature subset specific to each semantic class (Table 5.2). The number of features reduced is more than 50 % in most of the cases, while maintaining good accuracy. These selected features were then used to generate a semantic model for each class. This helps to rapidly extract a limited number of features that are highly relevant to a semantic class from images in a hurricane disaster event, and begin the process of knowledge discovery.

In the second set of experiments the ability of the GA for feature selection and generation is tested, wherein the feature dependencies are explicitly revealed. In the experiments simple arithmetic operators were used; however, the methodology could be extended to the generation of complex features that exhibit nonlinear relationships. This allows recursive feature generation and thus, the construction of arbitrarily complex features [41]. Table 5.3 shows the results from the combined feature selection and generation approach. The accuracy has improved and also compound features were generated. For example, three features were generated for Agriculture; Gen1, Gen2, Gen 3. Gen 1 is obtained by the product of the features (Table 5.4), geometric moment and maximum probability. The retrieval from a semantic query relevant to the hurricane event is depicted in Figure 5.7. Several tiles images have been retrieved by the system that matches the query. The user then has the option to further look into the details of the system-derived knowledge by clicking on an image, which brings up a window that provides a detailed view of the actual regions that matched the user’s semantic query (Figure 5.8). It is also possible to know the confidence level of each of the retrieved regions; this is helpful to understand how probable the region of a particular type. The RIIM system automatically calculates these confidence values (Figure 5.9) based on the posterior probabilities for all the regions and store them in the database. This also facilitates the retrieval of regions which are above a particular confidence threshold that satisfies the user’s needs. The RIIM system provides capabilities for the user to select area of interest (AOI) regions (Figure 5.10) on several example images. This enables the primitive features for the AOI to be automatically extracted and stored in the database for further processing, and for developing custom semantic models.

Further, once the knowledge has been discovered about a particular region, it is possible to send queries to archived data from the same region but, belonging to different sensors through the Web Coverage Service (WCS) integrated into the system. This provides capabilities for extracting only a limited amount of distributed data that meets the requirement, thus saving on the network bandwidth.



Current Extents

minX -104.51 minY 15.88

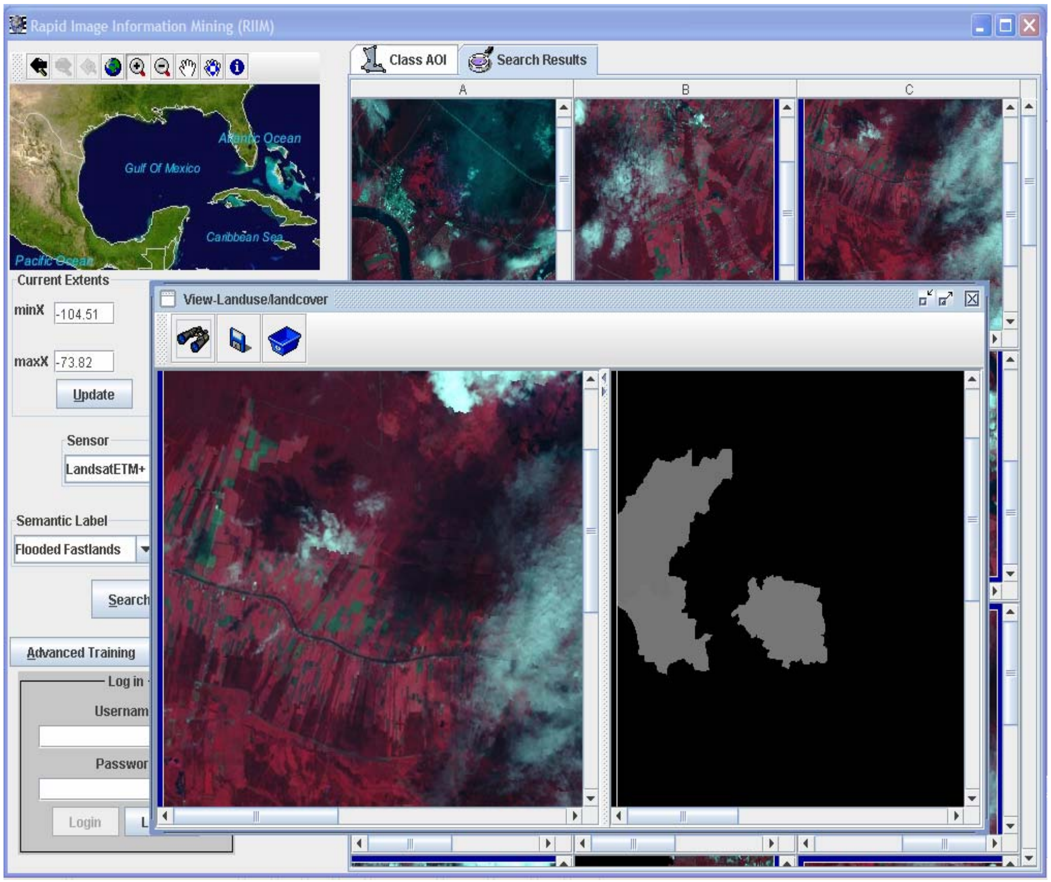
maxX -73.82 maxY 32.46

Sensor: LandsatETM+

Semantic Label: Flooded Fastlands

Cover type Training: Select Class

Figure 5.7 Results of a semantic query (flooded fastlands)



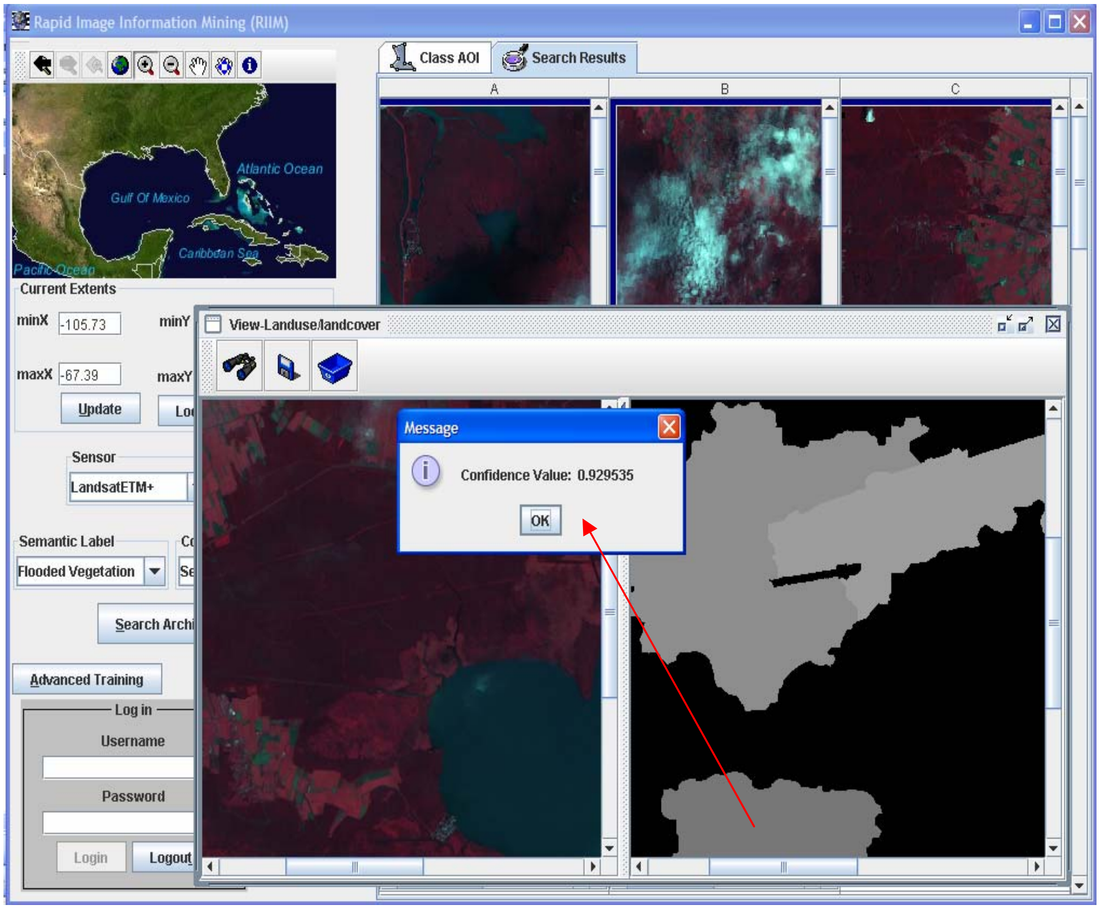
minX -104.51

maxX -73.82

Sensor: LandsatETM+

Semantic Label: Flooded Fastlands

Figure 5.8 Details of the semantic class (flooded fastlands). The regions on the right depict the retrieved regions matching the users query.



minX -105.73

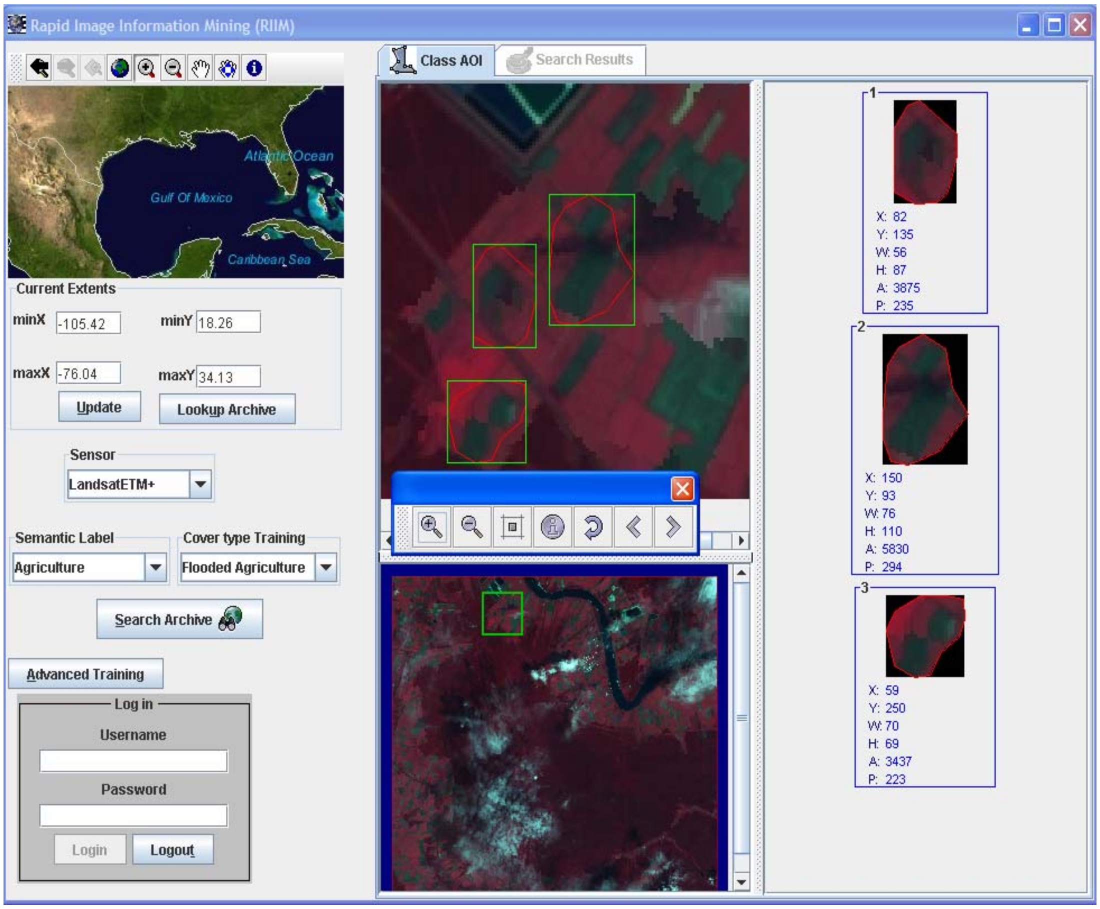
maxX -67.39

Sensor: LandsatETM+

Semantic Label: Flooded Vegetation

Confidence Value: 0.929535

Figure 5.9 Results of a semantic query (flooded vegetation). By clicking on any retrieved region on the right, the confidence value of that region is displayed.



minX -105.42 minY 18.26

maxX -76.04 maxY 34.13

Sensor: LandsatETM+

Semantic Label: Agriculture

Cover type Training: Flooded Agriculture

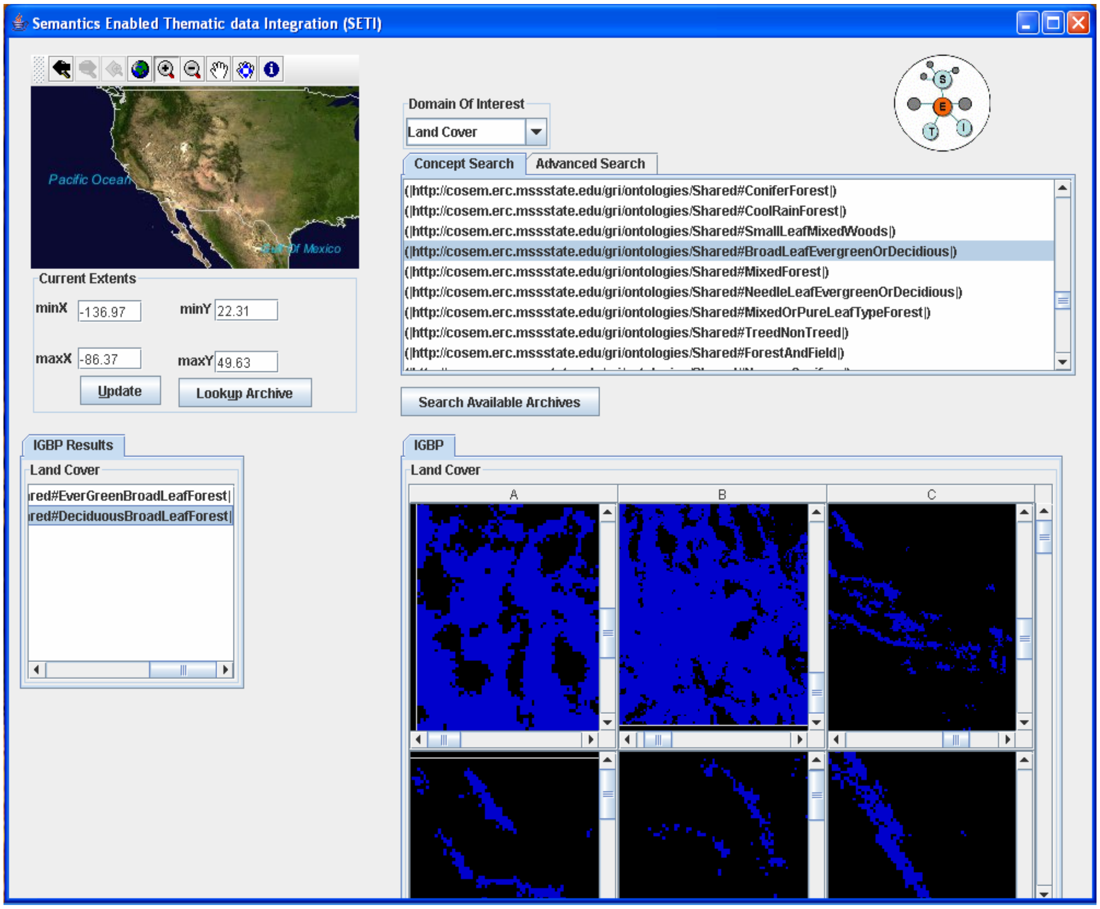
	X	Y	W	H	A	P
1	82	135	56	87	3875	235
2	150	93	76	110	5830	294
3	59	250	70	69	3437	223

Figure 5.10 Area of Interest (AOI) can be defined on example images which are later used to build custom semantic models.

5.3 Results from Semantics-Enabled Thematic Data Integration System

The SETI system's Graphical User Interface (GUI) provides the user with an integrated environment that provides functionalities to query based on the semantics across the thematic information repositories. For example, the user can first select a geographical extent he/she is interested in and then by selecting a domain of interest (e.g.

land cover), the concepts from the shared ontology belonging to land cover domain are automatically retrieved.



minX -136.97 minY 22.31

maxX -86.37 maxY 49.63

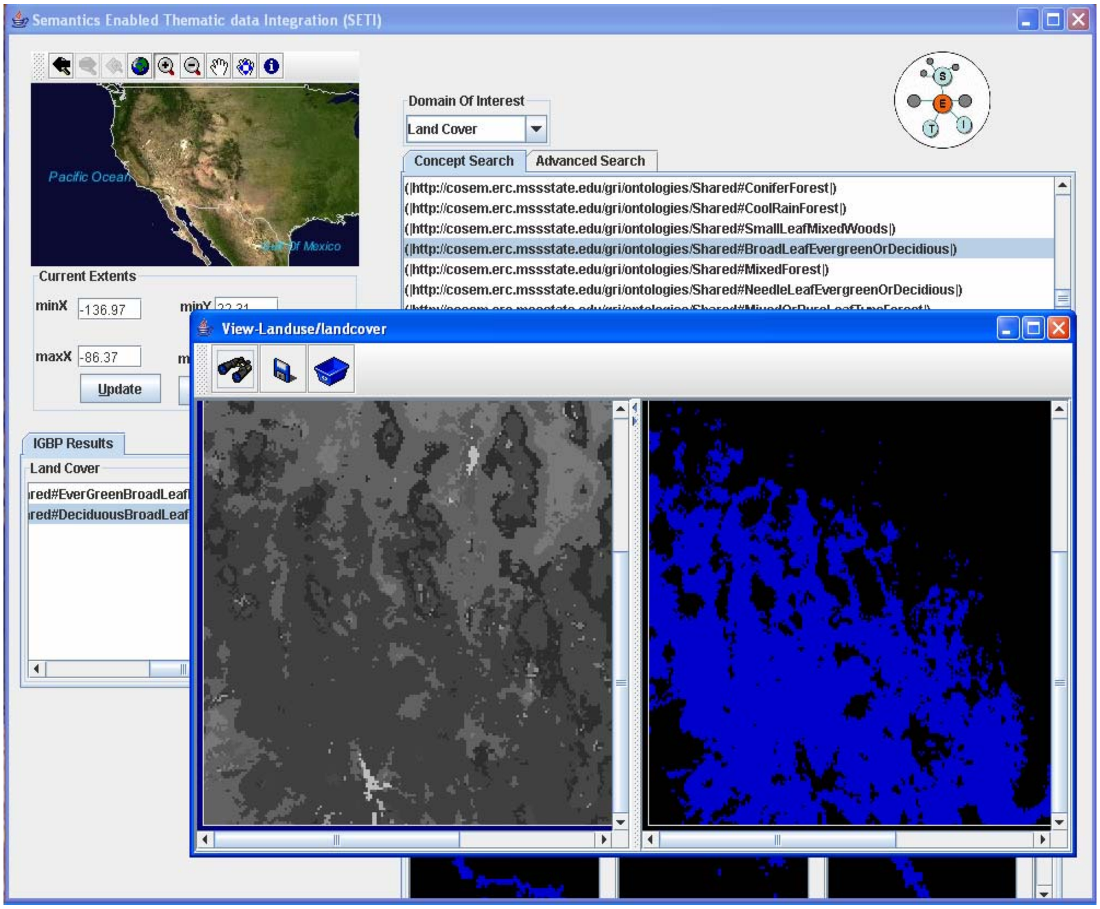
Domain Of Interest: Land Cover

(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolRainForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#SmallLeafMixedWoods|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#BroadLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NeedleLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedOrPureLeafTypeForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#TreedNonTreed|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ForestAndField|)

red#EverGreenBroadLeafForest|
red#DeciduousBroadLeafForest|

Figure 5.11 Prototype system for **S**emantics **E**nabled **T**hematic data **I**ntegration (**SETI**)

This method of providing predefined concepts in a domain of discourse is significant in two ways – first, it enables a user who is not very familiar with the terminology in a specific domain the ability to explore and select the concepts that approximately match his /her requirement.



minX -136.97

maxX -86.37

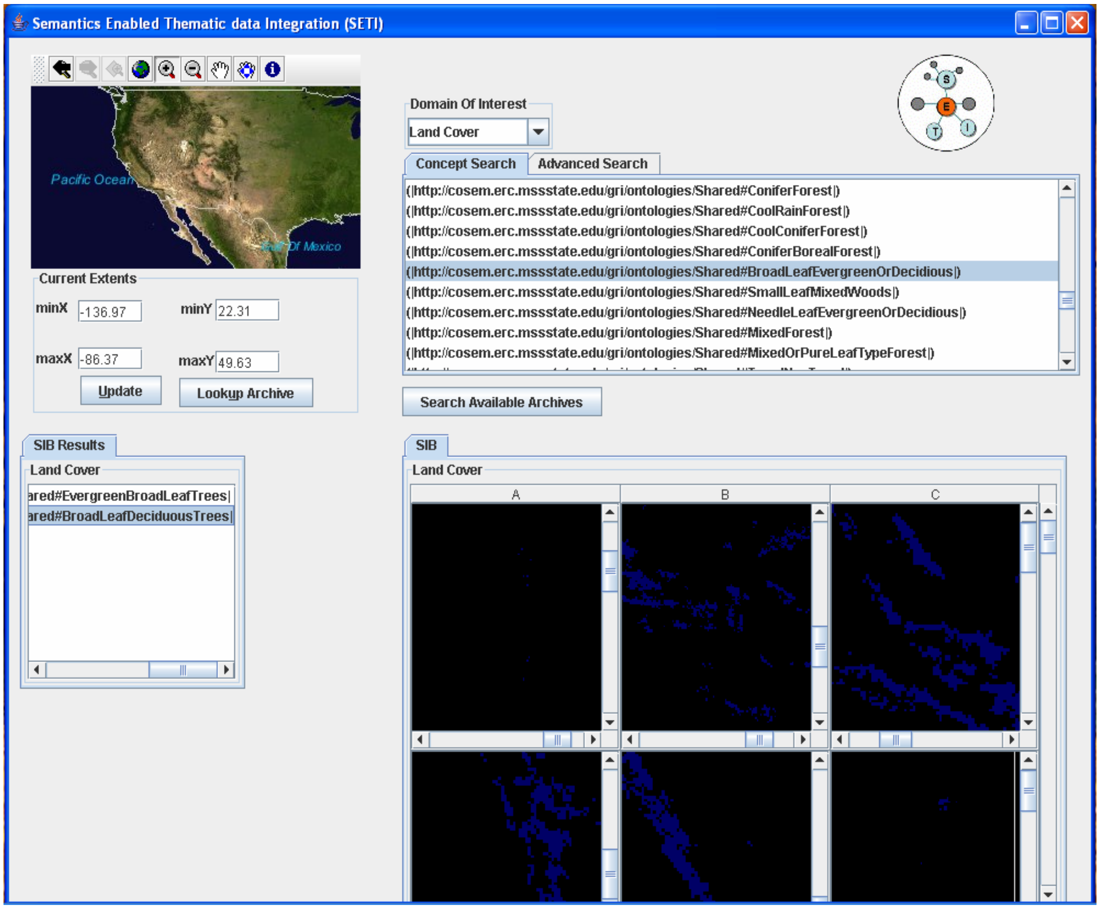
Domain Of Interest: Land Cover

(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolRainForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#SmallLeafMixedWoods|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#BroadLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NeedleLeafEvergreenOrDecidious|)

red#EverGreenBroadLeaf
red#DeciduousBroadLeaf

Figure 5.12 Details of deciduous broadleaf forest

This is because the shared ontology has been developed from the application ontologies and contains comprehensive terminology. The other purpose is that it will prevent the user from giving some wild keywords that the system may not recognize and also may not belong to the domain of interest.



Current Extents

minX -136.97 minY 22.31

maxX -86.37 maxY 49.63

Domain Of Interest: Land Cover

(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolRainForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferBorealForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#BroadLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#SmallLeafMixedWoods|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NeedleLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedOrPureLeafTypeForest|)

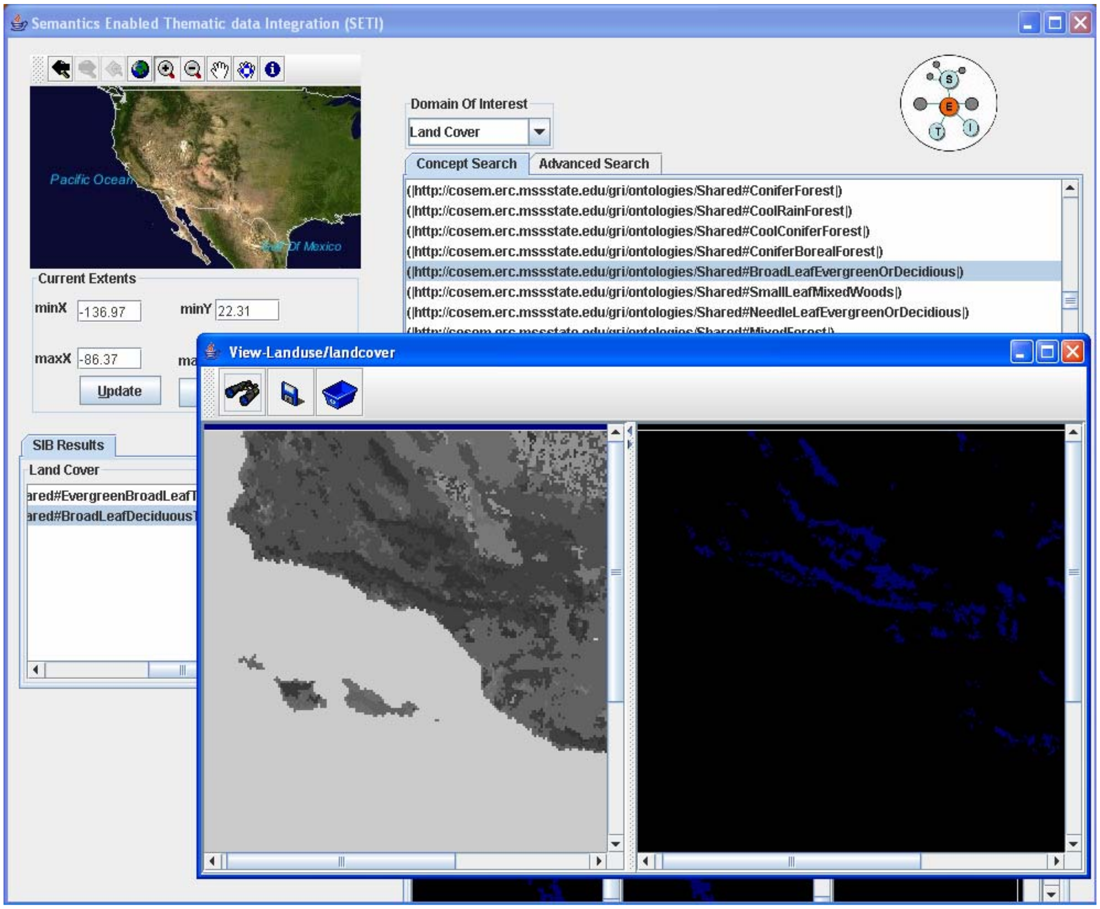
Land Cover

ared#EvergreenBroadLeafTrees|
ared#BroadLeafDeciduousTrees|

Land Cover

Figure 5.13 Results retrieved from SIB classified data repository

Figure 5.11 depicts the results of a concept query where the system searches for a Broadleaf Evergreen or Deciduous type of vegetation. The returned results contain the actual concepts that have been searched in application ontologies related to each of the thematic data repositories (e.g. IGBP, USGS, SiB, OGE etc). The DL-based reasoner uses the definition of concepts from the shared ontology which essentially are the subsumed concepts of the query concept and then searches for all the concepts that satisfy the criteria in each of the application ontologies.



Domain Of Interest: Land Cover

minX -136.97 minY 22.31

maxX -86.37

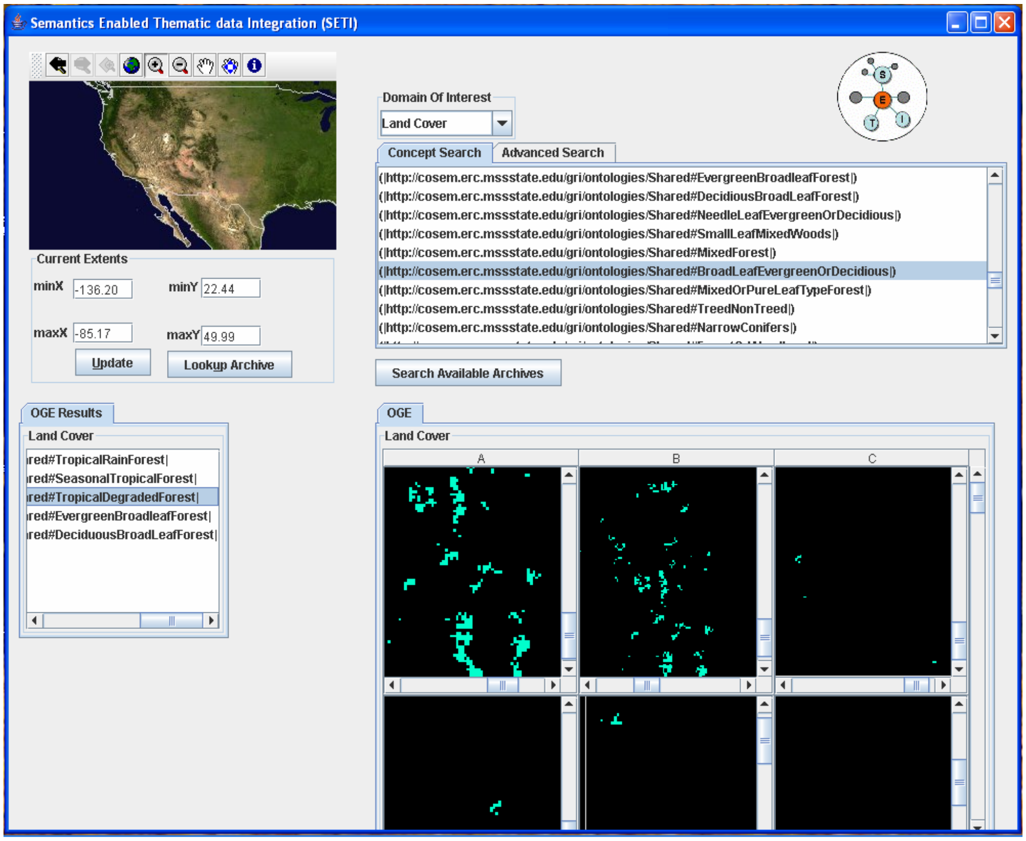
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolRainForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#CoolConiferForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#ConiferBorealForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#BroadLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#SmallLeafMixedWoods|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NeedleLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedForest|)

Land Cover
ared#EvergreenBroadLeafT
ared#BroadLeafDeciduousT

Figure 5.14 Details of result retrieved from SIB classified data repository

The intersection of the concept definitions that match with the subsumed concept definitions of the shared ontology forms the resulting query concepts. The instances of these concepts in the knowledge base form the results of the query. As shown in Figure 5.11, evergreen broadleaf forest and deciduous broadleaf forest are the concepts that have matched in the IGBP ontology for the given user query.

Selecting one of these concepts (Figure 5.11) the images corresponding to the 1992-1993, North America land cover characterization data in IGBP classification scheme are retrieved. The user can further explore and see greater details (Figure 5.12) in the retrieved images by clicking on the retrieved images; a new window opens depicting the original classified image and the corresponding image that contains the queried result. Similarly, Figure 5.13 shows the results returned from searching the thematic repository corresponding to SIB classification scheme.


Semantics Enabled Thematic data Integration (SETI)
S
E
T
I
Update
Lookup Archive
Concept Search
Advanced Search
Search Available Archives
OGE Results
OGE
A
B
C

Current Extents

minX -136.20 minY 22.44

maxX -85.17 maxY 49.99

Domain Of Interest: Land Cover

(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#EvergreenBroadleafForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#DecidiousBroadLeafForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NeedleLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#SmallLeafMixedWoods|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#BroadLeafEvergreenOrDecidious|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#MixedOrPureLeafTypeForest|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#TreedNonTreed|)
(|http://cosem.erc.msstate.edu/gri/ontologies/Shared#NarrowConifers|)

Land Cover

red#TropicalRainForest|
red#SeasonalTropicalForest|
red#TropicalDegradedForest|
red#EvergreenBroadleafForest|
red#DeciduousBroadLeafForest|

Land Cover

Figure 5.15 Results retrieved from Olson Global Ecosystems (OGE) classified data repository

The concepts retrieved from SIB ontology contain *evergreen broad leaf trees* and *broadleaf deciduous trees*, which differ semantically from the concepts in the IGBP ontology, but since the reasoner works at the conceptual level, the correct concepts are retrieved. Suppose a keyword-based search is conducted in a similar scenario these concepts might not have been discovered. Figure 5.14 depicts more details from a SIB classified image resulting from the user query.

In Figure 5.15, results from OGE classified data are shown, it can be seen that there are more concepts that have satisfied the query concept compared to the previous two results (IGBP and SiB) due to the granularity of the OGE classification scheme being finer (94 classes). Since the classes have been modeled as defined concepts and in concept hierarchy, the subsumed concepts in the shared ontology that matched the concepts in the OGE ontology have been retrieved. Since our primary interest is in image information mining, the retrieval of images from the database has been shown. However, the retrieval of the relevant textual data (e.g. mean NDVI values, biome, structure etc.) is trivial as it also forms an instance of the retrieved concepts.

In addition to the above, functionalities for Boolean querying are provided in the SETI system by the advanced search interface. This allows the combination of concepts with Boolean operators. Once the user discovers a particular information entity from the above semantic based querying, he/she can use the OGC WMS service to extract only the necessary data that meets their requirements. This also enables search of distributed archives and helps alleviate data transfer bottlenecks over the network. Once the knowledge has been discovered by mining through the thematic archives, WMS can also be used to facilitate decision-making by analyzing data (change detection studies etc) from multiple sensors (e.g., MODIS, Landsat) at different resolutions of the same region by separate requests to distributed archives (e.g., NASA, NOAA).

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

This research has resulted in the development of an image information mining system I^{3}KR; which is semantics-enabled image knowledge retrieval system for exploration of distributed remote sensing image archives. The process of image segmentation and primitive feature extraction followed by unsupervised learning via a KPCA approach has been developed. The SVM learning method has been described for the classification of the unsupervised content and subsequent model generation. A middleware that provides support for ontology storage, retrieval, and conceptual querying by means of DL reasoning enables the proposed system to provide enhanced knowledge discovery, query processing, and searching in a way that is not possible with ordinary keyword-based searches.

It has also been shown that the concept assignment of the model predicted objects could be achieved by classification via a DL reasoner through subsumption and equality, which enables classification from one context to another. The practical applications of the I^{3}KR system were demonstrated by executing semantic querying on archives of two sensors (MODIS and Landsat). The Graphical User Interface (GUI) developed for this research provides flexible access to the modules and in a coherent form. Currently the interactivity is restricted to the semantic cover type selection, knowledge base browsing and concepts exploration through the shared ontologies; however, future work should incorporate relevance feedback mechanisms.

The rapid image information mining (RIIM) prototype developed in this research is reliable and fast and is focused on image exploration for hurricane affected regions in near real-time scenarios. The computationally intensive tasks of feature extraction and model generation are considerably reduced by the wrapper-based approach for feature selection and generation shown in this research. This is vital for emergency response activities. The RIIM system provides capabilities for a first assessment of the disaster situation through the querying of the actual content in the remote sensing images, which is currently limited by queries only at the image metadata level. The developed RIIM system currently uses imagery from only one sensor, but can be easily scaled up to be used with a variety of sensors.

This research also presented the SETI system, which enables the retrieval of information from thematic data archives via semantics-driven searches. The need for such a system has been described and the paucity of such applications in Earth observations domain is highlighted. The components of the proposed system have been described in detail, including the ontology development process and the requirement for a shared ontology is presented along with the steps necessary to develop it. The shared ontology approach has been implemented by pursuing a motivating example, describing the semantic heterogeneities in the land cover classification schemes and the North

America land cover characterization dataset has been used as the source to demonstrate the proof of concept. The DL-based querying uses the semantic relations between the concepts (objects) hence it provides more expressiveness, and hence enables the proposed system to provide enhanced knowledge discovery, query processing, and searching in a way that is not possible with ordinary keyword-based searches. Results from the system corresponding to IGBP, SIB and OGE show that semantic reconciliation can be achieved by the proposed SETI system, and has been able to retrieve correct information from disparate thematic data repositories. Currently, semantics-based conjunctive queries are not handled by the system and it is proposed to enhance it with such querying capability in the next version of SETI.

6.2 Potential Topics for Future Work

Finally this section describes some useful directions and potential areas in which the current research could be advanced.

6.2.1 Parallel implementation of image information mining modules

The image information mining system developed in this research could be augmented with parallel implementations of some of its modules. Such an implementation would provide the following advantages:

- Better scalability of the system for processing archived imagery within reasonable time
- Improved savings in computation and resources

Real-time image segmentation is a well known problem as it is a computationally expensive operation with a high degree of uniformity for the operations applied to all pixels in an image. Hence, the segmentation algorithm that is used in this research to generate the regions in the images is a prime candidate for parallelization. Depending upon the segmentation algorithm used, three principal ways of doing segmentation are

- Detection of discontinuities (e.g. edge-based)
- Thresholding (e.g. based on pixel intensities)
- Region processing (e.g. group similar pixels)

Implementations on different parallel modes need to be investigated and compared for the above methods.

The other module of the proposed architecture that could be parallelized is the feature extraction component, where several algorithms based on color, texture, and shapes have been used to extract primitive features. The feature extraction task is computationally intensive; hence each algorithm could be run in parallel and also parallelized on each region in an image.

The searching for a semantic region within images is also an area for parallel implementation where several images could be searched in parallel with the generated semantic models obtained by machine learning methods. This would enable to search huge archives and produce useful results in reasonable time.

6.2.2 Development of methods for qualitative spatial reasoning on image data

In addition to the proposed research there is a need for greater advances in remote sensing imagery understanding in a number of costal zone scenarios. One of the objectives of such a research endeavor could be to understand the qualitative spatial relations between different land cover classes in an image.

“Retrieve all wetlands from LandsatETM+ archive that are in the southeastern part of state X and are near to a surface water body”

Proximity to surface water is an indicator of the likelihood that polluted runoff entering a wetland would otherwise enter surface water. Similar queries would also help in the evaluation of a wetland in relation to its significance to a watershed, habitat etc. It could also function as a rapid assessment technique by aggregating basic information on wetlands and landscape conditions - a necessary first step for detailed data analysis. The potential for the existence of runoff into a wetland may be assessed according to its spatial relation with respect to the surrounding land cover classes. If the wetland is surrounded by agricultural fields or surrounded by developed areas from which pollutants are likely to enter surface runoff, the wetland’s potential for removing non-point source pollutants is high. If, on the other hand, the wetland is mostly surrounded by natural vegetation from which runoff is likely to be largely unpolluted, it’s potential for removing significant pollutants is low [119-120]. Further, it is assumed that the higher in its watershed a wetland is located, the higher is its significance in non-point source removal.

As can be seen from the above, several qualitative spatial relationships (Figure 6.1) could be used to describe the potential of a wetland, such as:

- Wetlands *near* a water body
- Wetlands *close* to intermittent streams.

- Wetlands *surrounded* by natural vegetation
- Wetlands *surrounded* by agriculture.
- Wetlands *higher* in the watershed.
- Wetlands *adjacent* to significant source of polluted runoff.
- All wetlands that are *adjacent* to streams or rivers are considered to be riverine wetlands

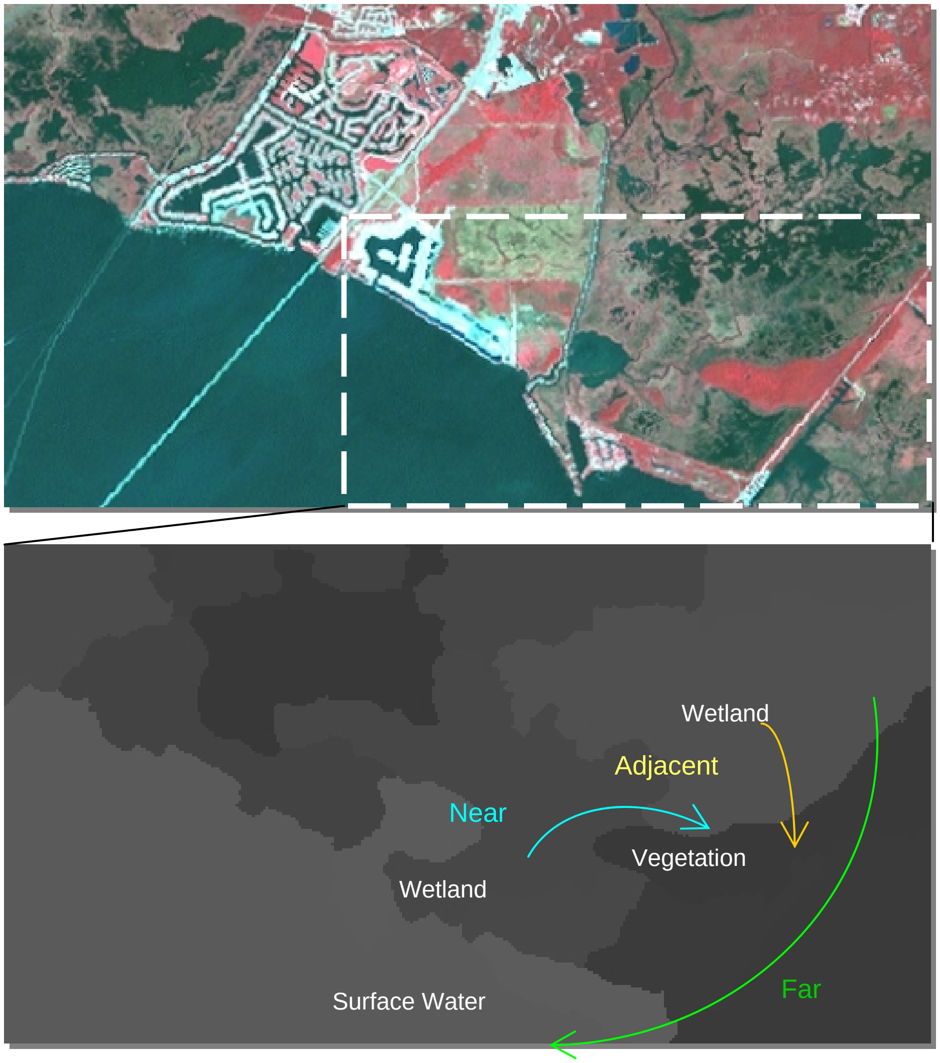


Figure 6.1 Some spatial relationships in wetlands domain

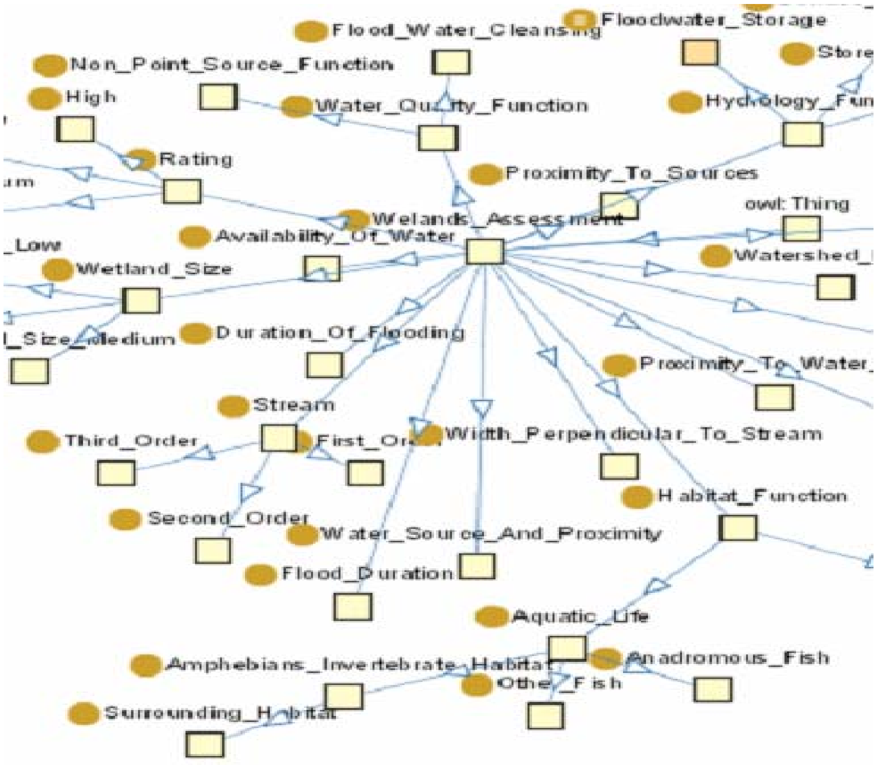


Figure 6.2 Wetlands assessment ontology

All bottomland hardwood wetlands must be *adjacent* to a river where they receive seasonal floodwaters from the channel.

The knowledge about the spatial relationships could be encoded as concepts that formalize the spatial arrangement that is unique for a type of wetland rating criteria. As shown in Figure 6.2, the ontology depicts the higher-level conceptualization of the terminology involved in wetlands assessment. The following provides the restriction on the class Riverine: *IsNear {Streams}*. Similarly for a Depressional/wetflat: *IsFar {SurfaceWater}*

Future research could work on the development of Spatial Arrangement Templates (SATs) that depict the instantiation of the relative arrangement of objects in a domain of discourse (e.g. Wetlands).

6.2.3 Fuzzy semantic metadata for spatial relations

A widely used method for modeling spatial relations has been proposed by Freeman [121], he also put forward the idea of fuzzy degree of truth to be associated with each spatial relation (topological and proximity). Each spatial relation thus defined gives a distinct semantic meaning. Yang et al proposed a method based on fuzzy K-NN classifier for the automatic generation of semantic metadata that describes the spatial relations [122]. They define the Semantic metadata as the fuzzy degree of truth with it associated spatial relation. Earlier studies have taken basically two approaches; the first one consists of algorithms that are designed for specific purposes and do not consider the human perception [122] and have not been very successful. The second approach draws upon the machine learning algorithms. It could be argued that the metadata generated by fuzzy K-means by the above method, although very useful, is not strictly semantic in the sense that

- The metadata generated is not machine understandable.
- Does not have enough semantic relationships built-in to enable reasoning by inferencing engines.
- Does not link semantic concepts for various degrees of fuzzy membership.

Thus future work could look into these aspects of generating semantic metadata for qualitative spatial relations.

6.2.4 Tools for ontology development

The development of applications for resolving semantic heterogeneities requires automated/semi-automated tools. There is a need for development of tools for learning ontologies and extracting metadata which is currently a key research area. Tools are also required for merging, aligning and storage of ontologies.

Existing tools are still in early stages of development and lack across-the-board integration. This is one of the next challenges in getting more widespread acceptance of the semantic web.

6.2.5 Enabling community participation in the ontology development

The ontologies that will furnish the semantics for the Semantic Web must be developed, managed, and endorsed as a community effort and focused towards the domain specific needs. The challenge is to bring together ontology engineers and domain experts and provide a platform for the shared understanding of the domain. Meta-standards in terms of upper ontologies for Earth science domains should be the next major focus of the international standardization organizations (e.g. ISO, FGDC, OGC, etc). The upper ontologies consist of the basic abstract categories and the major relations that link them. An upper ontology can help cut the time and effort to build domain-specific application ontologies and avoid simple mistakes. It also enables to share an ontology and make it more likely to be re-usable.

REFERENCES

[1] AmeriFlux network. [Online]. Available: http://public.ornl.gov/ameriflux/

[2] Fluxnet. [Online]. Available: http://daac.ornl.gov/FLUXNET

[3] S. Hankin and the DMAC Steering Committee. “Interoperable Data Discovery, Access, and Archive,” Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I, Ocean, US, Arlington, VA, pp. 292, and 2004. Available: http://www.dmac.ocean.us/dacsc/imp_plan.jsp.

[4] S. S. Durbha and R. L. King, “Semantics enabled framework for knowledge discovery from earth observation data archives”, *IEEE Transactions on Geosciences and Remote Sensing* Journal Vol.43, No.11, November 2005.

[5] G. Asrar, and J Dozier., “EOS: Science Strategy for the earth observing system”, Woodbury, NY: American Institute of Physics,1994

[6] A.S. Dennings, G. J. Collatz, C. Zhang, D. A. Randall, J.A Berry, P. J. Sellers, G. D. Colello, and D. A. Dazlich, “Simulation of terrestrial carbon metabolism and atmospheric CO_2 in a general circulation model, Part1: surface carbon fluxes”, Tellus B, 48, 521-542, 1996.

[7] B Hatton, (1997). Distributed Data Sharing. Paper presented at the AURISA 97, Christchurch, New Zealand.

[8] Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata, Version 2, FGDC-STD-001-1998. 2004. [Online]. Available: http://www.fgdc.gov/metadata/contstan.html

[9] I. B. Arpinar, A. Sheth and C. Ramakrishnan, “Geospatial Ontology Development and Semantic Analytics”, Handbook of Geographic Information Science, Eds: J. P. Wilson and A. S. Fotheringham, Blackwell Publishing (in print 2004)

[10] Y. Bishr, “Overcoming the semantic and other barriers to GIS interoperability”.

Int. J. Geographical Information Science, 12 (4), pp. 299-314.,1998.

[11] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. G. Machetti, and S D'Elia, "Information mining in remote sensing image archives: system concepts," IEEE Trans. on Geosci. Remote Sensing, vol. 41, no.12, pp. 2923-2936, 2003.

[12] G. H. Cheng, "Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources". PhD, MIT, 1997.

[13] A. S. Belward, (editor), "The IGBP-DIS Global 1 km Land Cover Data Set (DISCover): proposal and implementation plans", IGBP-DIS Working Paper 13, International Geosphere-Biosphere Programme Data and Information System, Toulouse, France. 1996.

[14] J. R. Anderson, E. E. Hardy, J. T. Roach, and R. E. Witmer, "A land use and land cover classification system for use with remote sensor data", U.S. Geological Survey Professional Paper 964, U.S. Geological Survey, Reston, VA, USA, 1976.

[15] J. S. Olson, "Global Ecosystems Framework: Definitions", Internal Report, USGS EROS Data Center, Sioux Falls, SD, USA, 1994.

[16] P. J. Sellers, S. O. Los, C. J. Tucker, C, O. Justice, D. A. Dazlich, G. J. Collatz, and D. A. Randall, "A revised land surface parameterization (SiB2) for atmospheric GCMS-part II: The Generation of Global Fields of Terrestrial Biophysical Parameters from Satellite Data". Journal of Climate, 9, pp. 706-737, 1996.

[17] P. J. Sellers, Y. Mintz, Y. C. Sud, and A. Dalcher, "A simple biosphere model (SiB) for Use within General Circulation Models", Journal of Atmospheric Science, 43, pp. 505-31. 1986.

[18] R. E. Dickinson, A. Henderson-Sellers, P. J. Kennedy, and M. F. Wilson, "Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model". NCAR Technical Note NCAR/TN-275+STR, Boulder, CO, USA. 1986.

[19] G. Pass and R. Zabih, "Histogram refinement for content-based image," Proc. IEEE workshop on Applications of Computer Vision, Sarasota, Florida, pp. 96-102, Dec. 1996.

[20] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC System," IEEE Computer, vol. 28, no. 9, pp. 23-32, 1995.

[21] A. Pentland, R.W. Picard, and S. Sclaroff, “Photobook: content-based manipulation for image databases,” Int’l J. Comput. Vis., vol. 18, no. 3, pp. 233-254, 1996.

[22] A. Gupta and R. Jain, “Visual information retrieval,” Commun. ACM, vol. 40, no. 5, pp. 70-79, 1997.

[23] C. Carson, S. Belongie, H. Greenspan, and J. Malik, “BlobWorld: Image Segmentation Using Expectation- Maximization and its Application to Image Querying,” IEEE Trans. Pattern Anal. Machine Intell., vol. 24, no. 8, pp. 1026-1038, 2002

[24] W. Y. Ma and B. S. Manjunath, “NeTra: A toolbox for navigating large image databases,” Proc. IEEE Int’l Conf. Image Processing, pp. 568-571, 1997.

[25] J. R. Smith and S. F. Chang, “VisualSEEK: a fully automated content-based query system,” Proc. 4th ACM Int’l Conf. on Multimedia, pp. 87-98, 1996.

[26] J. C. Tilton, G. Marchisio and M. Datcu, "Knowledge Discovery and Data Mining Based on Hierarchical Segmentation of Image Data,” research proposal submitted October 23, 2000 in response to NRA2-37143 from NASA's Information Systems Program. [Online]. Available: http://is.arc.nasa.gov/IDU/tasks/HierSeg.html.

[27] T. R. Gruber, “A translation approach to portable ontology specifications,” in Knowledge Acquisition, vol. 5, pp. 199-220, 1993.

[28] EOSDIS Core System Information for Scientists. [Online]. Available: http://observer.gsfc.nasa.gov/

[29] A. S. Islam, B. Beran, V. Yargici and M. Piasecki, “Ontology for Content Standard for Digital Geospatial Metadata (CSDGM) of Federal Geographic Data Committee (FGDC),” [Online]. Available:

http://loki.cae.drexel.edu/~wbs/ontology/fgdc-csdgm.htm

[30] M. Uschold, and M. Gruninger, “Ontologies: Principles, methods and applications”, Knowledge engineering review, 11(2): pp. 93-155. 1996

[31] R. Jasper, and M. Uschold, “A framework for understanding and classifying ontology applications”, proceedings of the 12th Banff knowledge acquisition for knowledge-based systems workshop. Technical report, University of Calgary/Stanford., 1999.

[32] H. Stuckenschmidt, F. V. Harmelen, “Information sharing on the semantic web”, Springer, 2005

[33] KAON -The Karlruhe Ontology and Semantic Web tool suite, 2004. [Online]. Available: http://kaon.semanticweb.org (last visited on Sept 30, 2005).

[34] On-to-knowledge, 2002. [Online]. Available: http://www.ontoknowledge.org (last visited on Sept 30, 2005).

[35] WebODE Ontology Engineering Platform. 2003. [Online]. Available: http://delicias.dia.fi.upm.es/webODE/ (last visited on Sept 30, 2005).

[36] S. Decker, M. Erdmann, D. Fensel, and R. Studer, "Ontobroker: Ontology based access to distributed and semi-structured information". In Meersman et al. editor, Semantic Issues in multimedia systems, Proc. of DS-8, pp. 351-369. Kluwer, Boston, MA, 1999.

[37] E. Mena, and A. Illarramendi, V. Kashyap, and A. Sheth, "OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies", Int'l. Journal of Distributed and Parallel Databases (DAPD), 8(2):pp. 223-272, 2002.

[38] U. Visser, H. Stuckenschmidt, G. Schuster, and T. Voegele, "Ontologies for geographic information processing". Computers in Geosciences, 28: pp. 103-117, 2002.

[39] A. Rodriguez, and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", IEEE Trans. on Knowledge and Data Engineering, vol.15, No. 2, March/April, pp. 442-456, 2003.

[40] S. Ram, J. Park, "Semantic Conflict resolution Ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts", IEEE Trans. on Knowledge and Data Engineering, Vol.16 no.2, Feb., 2004.

[41] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", Scientific American, May 2001.

[42] G. Klyne and J. J. Carroll, Resource description framework (RDF): Concepts and abstract syntax. W3C Rec., 10 Feb. 2004. http://www.w3.org/TR/rdf-concepts.

[43] http://www.w3.org/TR/daml+oil-reference (last visited on Sept 30, 2005)

[44] S. Bechhofer, F.V. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, "OWL Web Ontology Language Reference," Feb. 2004. [Online]. Available: http://www.w3.org/TR/owl-ref/ (last visited on Sept 30, 2005)

[45] M. Paul and J. Golbeck, "Visualization of Semantic Metadata and Ontologies," Proc. of Information Visualization, July 16-18, 2003, London, UK, 2003.

[46] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-based Integration of Information - A Survey of Existing Approaches," Proc. of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, WA, pp.108-117, 2001

[47] Protégé. "The Protégé Ontology Editor and Knowledge Acquisition System", 2006.[Online].Available: http://protege.stanford.edu/

[48] V. Haarslev and R. Möller, "Description of the RACER System and its Applications," Int. Workshop on Description Logics (DL-2001), Stanford, USA, Aug 1-3, 2001.

[49] Beck and H.S. Pinto (2002). Overview of Approaches, Methodologies, Standards and Tools for Ontologies, unpublished. [Online]. Available: http://www.fao.org/agris/aos/Documents/BackgroundAOS.html

[50] Open Geospatial Consortium (OGC). 2004. [Online]. Available: http://www.opengeospatial.org/

[51] Y. Deng, and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," IEEE Transactions on Pattern Analysis and Machine Intell., vol. 23, no. 8, pp. 800-810, Aug. 2001.

[52] S. Haykin, Neural Network-A Comprehensive Foundation, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[53] B. Schölkopf, A. Smola, K.R. Müller, "Nonlinear component analysis as a kernel Eigenvalue problem," Neural Computation, vol. 10, pp. 1299-1319, 1998.

[54] B. Schölkopf, A. Smola, K. R. Müller: "Kernel principal component analysis," Advances in Kernel Methods-Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, pp. 327-352, 1999.

[55] V. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[56] C. J. C. Burges, B. Schölkopf, "Improving the accuracy and speed of support vector learning machine," Advances in neural information processing systems, Cambridge, MA: MIT Press, pp. 375-381, 1997.

[57] V. Blanz, B. Schölkopf, H. Bulthoff, C. Burges, V. Vapnik, T. Vetter, "Comparison of view–based object recognition algorithms using realistic 3d models," proc. ICANN'96, , Springer lecture notes in computer science, Berlin ,vol. 112, pp.251-256, 1996.

[58] E. Osuna, R. Freund, F. Girosi, 'Training support vector machines: an application to face detection.' Proc. Computer vision and pattern recognition, pp. 130-136, 1997.

[59] R. Courant and D. Hilbert, Methods of Mathematical Physics. New York: Wiley, 1953.

[60] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers", Proc. the fifth annual workshop on computational learning theory, Pittsburgh, vol. 5, pp. 144-152, 1992.

[61] C-C Chang, C-J Lin, "LIBSVM: a library for support vector machines," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[62] J.C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, Eds. pp. 61-74, MIT Press, 1999.

[63] D. Calvanese, G. D. Giacomo, D. Nardi, M. Lenezerini, "Reasoning in expressive Description Logics," Hand Book of Automated Reasoning, B.V.: Elsevier Publishers, pp. 3-12, 2001.

[64] G. Schuster and H. Stuckenschmidt. "Building shared ontologies for terminology integration," In proc. KI-01 Workshop on Ontologies, G. In Stumme, A. Maedche, and S. Staab, Eds. vol. 48. 2001.

[65] M. Stefik. Introduction to knowledge systems. San Francisco, CA: Morgan Kaufmann Publishers, 1995.

[66] M. Datcu, K. Seidel, and G. Schwarz, "Elaboration of advanced tools for information retrieval and the design of a new generation of remote sensing ground segment systems," in Machine Vision in Remote Sensing, I. Kanellopoulos, Ed. Berlin, Germany: Springer-Verlag, pp. 199–212, 1999.

[67] J. Loughery, P. Cunningham,"Overfitting in wrapper-Based Feature subsets Selection: The harder you try the worse it gets" In Proceedings of the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2004.

[68] L. Bruzzone, "An approach to feature selection and classification of remote sensing images based on the Bayes rule for minimum cost. IEEE Trans. on Geosci. Remote Sensing, vol.38, no.1, January 2000.

[69] J. A Richards, Remote sensing Digital image analysis, 2nd Ed. New York: Springer-Verlag, 1993.

[70] P. H. Swain and S. M. Davis, Remote Sensing: The Quantitative Approach. New York: McGraw-Hill, 1978.

[71] I. L. Thomas, N. P. Ching, “A review of multi-channel indices of class separability,” *Int. J. Remote Sens.*, vol. 8, pp. 331–350, 1987.

[72] P.W. Mausel, W. J. Kramber, and J. K. Lee, “Optimum band selection for supervised classification of multispectral data”, Photogram. Eng. Remote Sens., vol. 56, pp.55-60, Jan.1990.

[73] L. Bruzzone, F. Roli, and S. B. Serpico, “An extension of the Jeffreys-Matusita distance to multiclass cases of feature selection”, IEEE Trans. on Geosci. Remote Sensing, vol. 33, pp. 1318-1321, Nov.1995.

[74] M. Pei, E.D.GoodNam, W.F. Punch “ feature extraction using genetic algorithms”, Proceeding of International Symposium on Intelligent Data Engineering and Learning’98 (IDEAL’98), Hong Kong, Oct. 1998.

[75] P.A. Deviver and J. Kittler, “Pattern recognition: A statistical approach”. New York: Prentice-Hall. 1982.

[76] R. Kohavi, G. H. John, “Wrappers for Feature subset selection”, Artificial intelligence journal special issue on relevance, 1997.

[77] E. Cantu-Paz, “Feature subset selection, class separability and genetic algorithms”, submitted to Genetic and evolutionary computing conference, WA, 2007.

[78] D. Aha, R. Bankert, “A Comparative evaluation of sequential feature selection algorithms”, Artificial intelligence and Statistics, D. Fisher and J.H. Lenz, New York 1996.

[79] S.K. Pal and P.P. Wang Eds., “Genetic algorithms for Pattern Recognition”, Boca Raton, FL: CRC, 1996.

[80] F.Z. Brill, D.E. Brown, W.N. Martin, “Genetic algorithms for feature selection for counter propagation networks”. Tech. Rep. No. IPC-TR-90-004, University of Virginia, Institute of Parallel Computation, Charlottesville, 1990.

[81] T.W. Brotherton, and P.K. Simpson, “Dynamic feature set training of neural nets for classification”, In McDonnell, J.R., Reynolds, R.G., Fogel, D.B., eds.: Evolutionary Programming IV, Cambridge, MA, MIT Press, 83(94), 1995

[82] J. Bala, J. De, K. J. Huang, H. Vafaie, H. Wechsler, “Using learning to facilitate the evolution of features for recognizing visual concepts”, Evolutionary Computation, 4, pp. 297-311, 1996.

[83] J.D. Kelly, L. Davis, “Hybridizing the genetic algorithm and the K-nearest neighbors classification algorithm”. In R.K. Belew, L.B. Booker, eds.: Proceedings of the Fourth International Conference on Genetic Algorithms, San Mateo, CA, Morgan. pp. 377-383, 1991

[84] W.F. Punch, E.D. Goodman, M. Pei, L. Chia-Shun, P. Hovland, “Further research on feature selection and classification using genetic algorithms”, Proceedings of the Fifth International Conference on Genetic Algorithms, San Mateo, CA, Morgan Kaufmann, pp. 557-564, 1993.

[85] M.L. Raymer, W.F. Punch, E.D. Goodman, P.C. Sanschagrin, L.A. Kuhn, “Simultaneous feature scaling and selection using a genetic algorithm”, Proceedings of the Seventh International Conference on Genetic Algorithms, San Francisco, Morgan Kaufmann, pp. 561-567, 1997.

[86] M. Kudo, K. Sklansky, “Comparison of algorithms that select features for pattern classifiers. Pattern Recognition, vol. 33, pp. 25-41, 2000.

[87] I. Inza, P. Larra~naga, R. Etxeberria, B. Sierra, “Feature subset selection by Bayesian networks based optimization”, Artificial Intelligence, Vol. 123, pp. 157-184, 1999.

[88] E. Cantu-Paz, “Feature subset selection by estimation of distribution algorithms” In Langdon, W.B., Cantu-Paz, E., Mathias, K., Roy, R., Davis, D., Poli, R., Balakrishnan, K., Honavar, V., Rudolph, G., Wegener, J., Bull, L., Potter, M.A., Schultz, A.C., Miller, J.F., Burke, E., Jonoska, N., eds., Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, CA, Morgan Kaufmann Publishers, pp. 303-310, 2002.

[89] A. Jain, D. Zongker, “Feature selection: evaluation, application and small sample performance”. IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 19, pp. 153-158, 1997.

[90] M. M. Rizki, M. A. Zmuda and L. A. Tamburino, “Evolving pattern recognition systems”, IEEE trans. on evolutionary computation, vol. 6, no. 6. Dec, 2002.

[91] M. Sasikala and N. Kumaravel, “Comparison of feature Selection techniques for detection of malignant tumor in brain images”, IEEE Indicon conference, Chennai, India, Dec 2005.

[92] A. Frietas, “ Understanding the crucial role of attribute interaction in data mining”, 16(30), pp 177-199, 2001

[93] Oliver Ritthoff, Ralf Klinkenberg, Simon Fischer, Ingo Mierswa, "A Hybrid Approach to Feature Selection and Generation Using an Evolutionary Algorithm**",** Proceedings of the 2002 U.K. Workshop on Computational Intelligence (UKCI-02), 2002.

[94] A.S. Fraser," Simulation of genetic systems by automatic digital computers-I: Introduction," Australian J. Bio.Sc*i*, vol.10, pp.484-491, 1957.

[95] J.L. Crosby, "Computers in the study of evolution", Sci.Prog.Oxf., vol.55, pp.279-292, 1967.

[96] H. J. Bremermann, M. Rogson, and S. Salaff, "Global properties of evolution processes, "Natural automata and useful simulation, H. H. Pattee, E. E. Edlsack, L. Fein, and A.B. Callahan, Eds. Washington, DC: Spartan, pp.3-41, 1966.

[97] J. Reed, R. Toombs and N. A. Barracelli, "Simulation of biological evolution and machine learning: Selection of self-reproducing numeric patterns by data processing machines, effects of hereditary control, mutation type and crossing, " J. theorist. Biol., vol. 17, pp. 319-342, 1967.

[98] D. B. Fogel, Ed., Evolutionary Computing: The Fossil Record. New York: IEEE Press, 1998.

[99] J. H. Holland, Adaptation in natural and Artificial Systems. Ann Arbor, MI:Univ. Michigan Press, 1975.

[100] D. Goldberg, Genetic Algorithms in search and optimization, and Machine learning, Reasoning, MA: Addison-Wesley, 1989.

[101] R. Poli, J. E. Rowe and N. McPhee. Markov chain models for GP and variable-length gas with homologous crossover. In Lee Spector et al., editor, proceedings of the genetic and evolutionary computation conference (GECCO-2001), pp 112-119, San Francisco, CA, USA, 2001.

[102] H. Stuckenschmidt, F. V. Harmelen, "Information sharing on the semantic web", Springer, 2005.

[103] A. Rodriguez, and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", IEEE Trans. on Knowledge and Data Engineering, vol.15, No. 2, March/April, pp. 442-456, 2003.

[104] Y. Bishir, "Semantic Aspects of Interoperable GIS," Wageningen Agricultural Univ. and ITC, the Netherlands, 1997.

[105] M. Bright, A. Hurson, and S. Pakzad, "Automated Resolution of Semantic Heterogeneity in Multidatabases," ACM Trans. Database Systems, vol. 19, pp.

212-253, 1994.

[106] P. Fankhauser and E. Neuhold, “Knowledge Based Integration of Heterogeneous Databases”, Proc. Database Semantics Conference Interoperable Database Systems IFIP WG2.6, H. Hsiao, E. Neuhold, and R. Sacks-Davis, eds., pp. 155-175, 1992.

[107] C. Collet, M. Huns and W. Shen, “Resource Integration Using Large Knowledge Base in Carnot”, IEEE Computer, vol. 24, pp. 55-62, 1991.

[108] V. Kashyap and A. Sheth, “Semantic Heterogeneity in Global Information Systems: The Role of Metadata, context, and Ontologies”, Cooperative Information Systems: Trends and Directions, M. Papazoglou and G. Schlageter, eds., pp. 139-178, 1998.

[109] B. Bergamaschi, S. Castano, S. De Capitini di Vermercati, S. Montanari, and M. Vicini, “An Intelligent Approach to Information Integration,” Proc. First Int’l Conference. Formal Ontology in Information Systems, N. Guarino, eds., pp. 253-268, 1998.

[110] A. Gangemi, D. Pisaneli, andG. Steve, “Ontology Integration: Experiences with Medical Terminologies”, Formal Ontology in Information Systems, N. Guarino, ed., pp. 163-178, 1998.

[111] T. R.Loveland, Z. Zhu, D. O. Ohlen, J. F. Brown, B. C. Reed, and L. Yang, “An Analysis of the IGBP Global Land-Cover Characterization Process”. Photogrammetric Engineering and Remote Sensing, vol. 65, no. 9, pp. 1021-1032. 1999.

[112] T.R. Loveland, B.C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, J. W. Merchant, “ Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data”, Int. J. Remote Sensing, vol. 21, no. 6 & 7, pp. 1303-1330, 2000.

[113] http://edcsns17.cr.usgs.gov/glcc/nadoc2_0.html (last visited on Sept 30,2005)

[114] http://www.w3.org/TR/owl-guide/ (last visited on Sept 30,2005)

[115] T. Bittner, M. Donnelly; S. Winter, “Ontology and Semantic Interoperability”. In: Prosperi, D.; Zlatanova, S. (Eds.), Large-Scale 3D Data Integration, CRC Press, London. 2005.

[116] T. C. Pape, and K. Thoresen, “Evolutionary Prototyping in a Change Perspective”, Information Technology and Peopl*e*, vol. 6, pp. 2-4.1992

[117] B. Selman, and H. Kautz, “Knowledge compilation and theory approximation”.

Journal of ACM, 43(2): pp.193-224.1996.

[118] Java Technology, 2006. [Online]. Available: http://java.sun.com.

[119] NC-CREWS: North Carolina Coastal Region Evaluation of Wetland Significance. *A report of the strategic plan for improving coastal management in North Carolina*. 1999.

[120] [Online] http://dcm2.enr.state.nc.us/Wetlands/nccrews.htm

[121] J. Freeman, "The modelling of spatial relations. Computer Graphics and Image" Processing, 4(2):156171, 1975.

[122] Y. Wang, F. Makedon, J. Ford, L. Shen, and D. Goldin. "Generating fuzzy semantic metadata describing spatial relations from images using the R-histogram", *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, pp 202-211, 2004.

APPENDIX A

ACRONYM LIST

AVHRR Advanced Very High Resolution Radiometer

BATS Biosphere Atmosphere Transfer Scheme

CBIR Content-Based Image Retrieval

CS-W Web Catalog Service

CSDGMD Content Standards for Digital Geospatial Meta Data

DAAC Distributed Active Archive Centers

DL Description Logics

DAML DARPA Agent Markup Language

EO Earth observation

EOSDIS Earth Observing System Data and Information System

FGDC Federal Geographic Data Committee

FCC False Color Composite

GA Genetic Algorithm

GRI GeoResources Institute

IOOS Integrated Ocean Observation System

IGBP International Geosphere Biosphere Programme

ISO International Standards Organization

I^{3}KR Intelligent Interactive Image Knowledge Retrieval

KPCA Kernel Principle Component Analysis

KES Knowledge Enabled Services

MODIS Moderate Resolution Imaging Spectroradiometer

NASA National Aeronautics and Space Administration

NOAA National Oceanic and Atmospheric Administration

NSF National Science Foundation

OGE Olson Global Ecosystems

OWL Web Ontology Language

RDF Resource Description Framework

RIIM Rapid Image Information Mining

SiB Simple Biosphere model (SiB)

SiB2 Simple Biosphere model2(SiB2)

SQL structured query language

SETI Semantics Enabled Thematic Data Integration

SVM Support Vector Machines

USGS United States Geological Survey

USGS United States geological Survey

WMS Web Map service

WFS Web Feature Service

WCS Web Coverage Service