Knowledge Representation and Exchange of Visual Patterns using Semantic Abstractions

---

A Dissertation
Presented to
The Academic Faculty of Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

---

by

Adrian S. Barb
Dr. Chi-Ren Shyu, Disertation Supervisor

Computer Science Department
University of Missouri
August 2008

The undersigned, appointed by the Dean of Graduate School, have examined the dissertation entitled:

Knowledge Representation and Exchange of Visual Patterns using Semantic Abstractions

presented by Adrian S. Barb
a candidate for the degree of Doctor of Philosophy
and hereby certify that in their opinion it is worthy of acceptance.

Approved by:

Dr. Chi-Ren Shyu, Adviser

Dr. Mary Schaeffer
(Division of Plant Science)

Dr. Dong Xu
(Computer Science Department)

Dr. Dmitry Korkin
(Computer Science Department)

Dr. Curt H. Davis
(Electrical & Computer Engineering)

Dr. David Jonassen
(School of Information Science and Learning Technologies)

Date Approved:

to Andreea, Simona, my Mom, and Carmen.

ACKNOWLEDGEMENTS

I want to express my gratitude to my adviser, Dr. Chi-Ren Shyu, for being flexible and allowing me the time and support that I needed for research. With his enthusiasm, inspiration, and effort, he helped make research fun and interesting. Throughout my dissertation-writing period, he provided encouragement, sound advice, good teaching, and lots of good ideas. I would also like to thank my committee members Dr. Dong Xu, Dr. Curt H. Davis, Dr. Dmitry Korkin, Dr. Mary Schaeffer, and Dr. David Jonassen for their for their valuable and constructive criticism during my PhD training.

I wish to extend my warmest thanks to all professors, colleagues, collaborators, and friends who have helped me with my research. I want to thank present and past members of the Medical and Biological Digital Library Lab at the University of Missouri for providing a productive and fun working atmosphere, and for sharing their knowledge with me. I want to give special thanks to Dr. Yash Sethi for a fruitful collaboration. I am also grateful to Dr. Charles Franz, Dr. Ronald Ebert, and Dr. Gregg Martin for getting me interested in research. I also owe my most sincere gratitude to Letha and Ken Albright who gave me untiring help during my difficult moments.

I cannot end without thanking my family, on whose constant encouragement and love I have relied throughout my time at University of Missouri. I am also grateful to the support of my sister and my mother. Their unflinching support will always inspire me, and I hope to continue, in my own small way, the noble mission to which they gave their lives. It is to them that I dedicate this work.

This work was supported in part by the National Geospatial-Intelligence Agency University Research Initiatives (NURI) under grant number HM1582-04-1-2028 and by the National Science Foundation under grant number DBI-0447794.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Modern technology enables organizations to build large-scale data repositories. The utility of such repositories, however, is limited if those repositories do not support flexible methods of extracting knowledge, especially for repositories of visual artifacts. Existing content-based visual media retrieval systems create models that often are optimized to the domain knowledge provided by experts during training processes. However, most of these systems lack the flexibility to address the gap between computer and human representations of visual patterns.

The scope of this dissertation is to research methods of knowledge exchange in a loosely integrated environment, while preserving the individual characteristics of knowledge representation. For this, we have developed a knowledge repository and exchange framework for large-scale image databases called *Essence*. This framework facilitates domain knowledge representation by mapping commonly agreed on semantics into low-level features using flexible association rules. It also provides novel and efficient methods of exchanging of both tacit and explicit knowledge using semantics. This research was applied to modeling the phenotype-genotype correlations of maize mutants (bioinformatics), studying patterns of pulmonary diseases found in high-resolution computed tomography images of lungs (medical informatics), and discovering relevant knowledge from satellite images (geospatial intelligence). Over the past four years, this research has empirically proven valuable in assisting domain experts in their decision-making processes. The *Essence* framework can be applied for training and decision making and could be the foundation of building a novel and flexible model for visual media retrieval that uses expert-defined semantics. With appropriate extensions, this approach can be adapted to other domain-specific visual media databases.

CHAPTER I

INTRODUCTION

Humans are characterized by the desire to extract meaning from the environment. Every day, we evaluate signs found in our interaction with the surroundings. These signs are not limited to words, images, sounds, or objects, but also to things that have no intrinsic meaning and become signs only when we invest them with meaning. According to Charles Sanders Peirce, "Nothing is a sign unless it is interpreted as a sign" [101]. When interpreting such signs, we associate them with things or events from previous experience using mental models-unique personal systems of conventions [137].

Individuals develop customized mental models that are affected by many factors. These factors may include knowledge, beliefs, goals, preferences, interests, misconceptions, plans, tasks, abilities, or work settings. Mental models are developed using intuitive or naïve knowledge acquired from interaction with the environment, as well as formal knowledge acquired from teachers who share their interpretation of the subject [142]. Mental models are dynamic and subject to a continuous adaptation to new knowledge that is acquired to explain new problems at hand [66]. The fact that we interpret surroundings according to our experience helps us understand that there is no unique and objective version of reality that applies to everyone, but rather a set of human interpretations of it, according to each individual set of values. When knowledge exchange is required, there is a need to bridge these interpretations to arrive at a common factor that will make the knowledge transparent to others.

Studying methods of deriving meanings from signs or patterns is called Semiotics [39]. The study of semiotics is not limited to facilitating communication but also includes the construction and understanding of reality models. According to Peirce [101], extracting meaning from patterns has a triadic model: (1) Sign or a perceived pattern; (2) Interpretant or sign sense; and (3) Object or concept to which the sign refers. This model is shown in

Figure 1 (a). For example, when evaluating an image, we use past experience to convert perceived visual patterns into more complex models through an iterative interpretative process. The result of this iteration is a concept to which the initial sign refers and which is in accordance with the individual's mental model.

The main process in attaching meaning to a pattern is the conceptualization of the connection between new patterns and concepts encountered in previous experience. An example of this model from the radiology domain, shown in Figure 1 (b), is as follows: While evaluating high resolution computer tomograph (HRCT) images of a lung, the radiologist may observe some dark, round, well-circumscribed areas of lobes–the sign. Using formal education and past experience, the radiologist will associate these patterns with "thin-walled, well-defined and well-circumscribed air-containing lesion"-the interpretant, which stands for a *cyst of lung*-the object [88]. Note that in this case *emphysema* has a similar description except that it shows very thin and less-defined walls [88]. Figures 1 (b), (c), and (d) show examples of extracting meaning from images in the medical, bioinformatics, and geospatial domains. Note that humans may evaluate successive interpretants of the same sign until they reach the best model [43], [66]. This means that an interpretation can be re-interpreted to adjust to the new experience, and that knowledge discovery is a dynamic process.

Modern information technology, which produces evermore powerful computers each year, enables us to collect and store large amounts of information at very low costs. Many organizations are able to compile vast archives of data repositories using electronic formats. However, collecting and storing data is not sufficient if there are no methods of extracting the meaning hidden in such collections. For example, the latest generation of high-resolution satellites, including IKONOS and QuickBird, are generating terabytes of imagery daily. Also, in the medical domain, a typical examination will take 20 to 100 images that correspond to a series of cross-sectional slices through the patient's body [93]. Each of these images ranges from less than one megabyte to around 30 megabytes for different modalities [146]. Considering the size of the problem at hand, manual evaluation by experts (expert-in-the-loop) of all of the images becomes expensive. In such cases, computer algorithms are

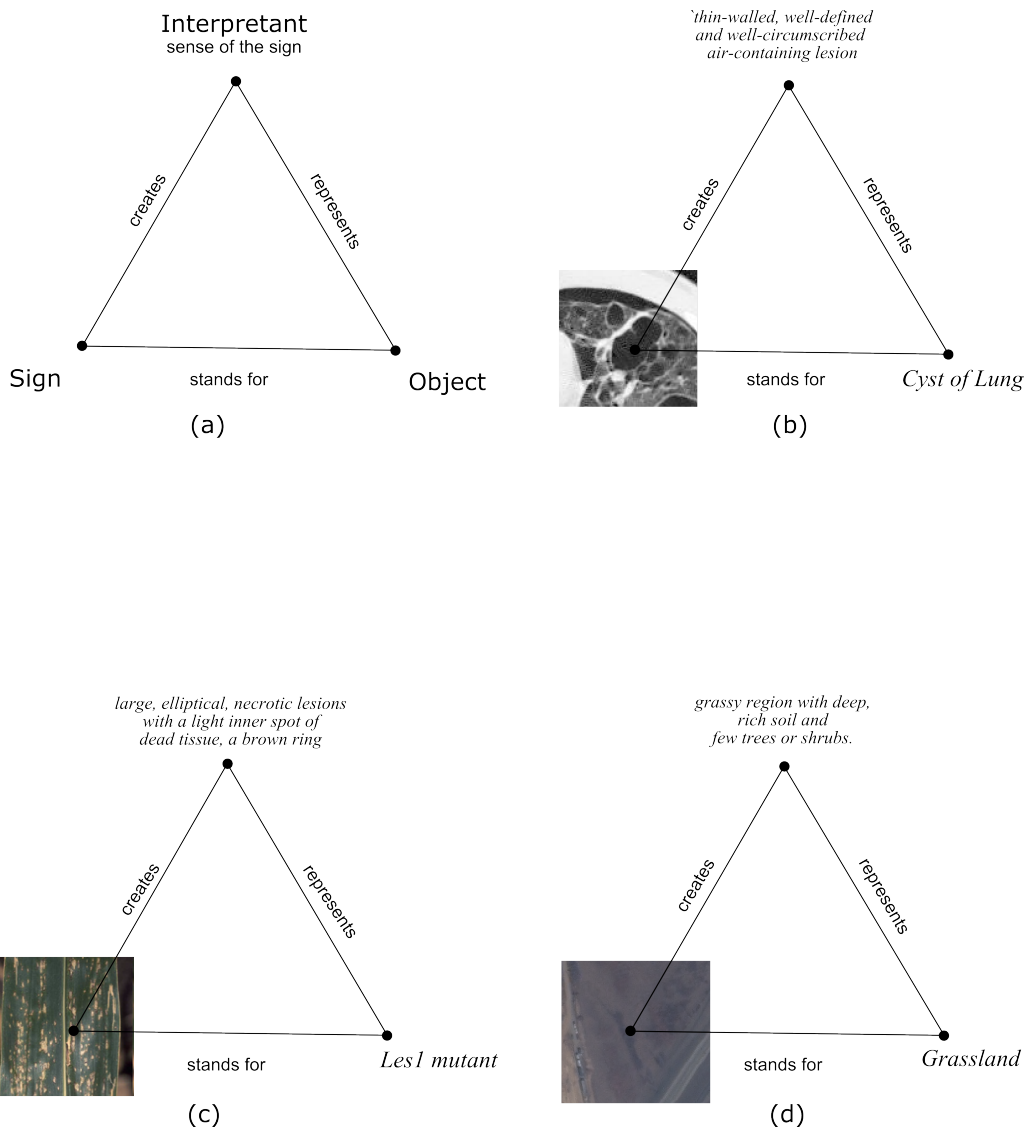


Figure 1: Model of extracting meaning from signs: (a) the semiotic triad described by Pierce [101]; Examples of deriving (b) *Cyst of lung* from a high resolution computed tomography image; (c) *Les1 mutant* from an image of maize leaf; and (d) *Grassland* from a satellite image.

expected to detect valuable visual patterns that can easily be used to filter the data that is presented to the human analysts. For instance, if a radiologist wants to search images containing *cyst of lung*, a computer-based system can reduce the burden on the expert by filtering out all the images that do not contain this visual pattern.

Images are human artifacts created with the purpose of communication. However, extracting meaning from visual information found in images is not a trivial task [115]. The computer stores image information as an array of pixel values that has no intrinsic meaning to humans. Rather than looking at raw pixel values, humans group sets of pixels together and identify familiar patterns. They then connect these patterns with models they acquired in past experience to extract meaningful knowledge. According to Santini [115], extracting meaning from an image is successful only in one of these scenarios: (1) image context is fully described by text and a discourse that links the image to the text; (2) the context is implicitly described in the social discourse, or (3) the context and consequently the linguistic discourse is provided by the user through iterative processes. In this dissertation, we will address the subject of extracting knowledge from domain-specific images. Applying our approach only to domain-specific images helps us describe their context using controlled vocabularies or ontologies. This also will reduce the ambiguity of semantic assignment, increase interoperability, and allow us to develop improved models for representing visual patterns.

1.1 From Data through Information to Knowledge

Knowledge is information organized to provide meaning and that has the ability to predict properties of the subject [131]. This representation is according to the "knowledge hierarchy" model [1], [151]. The knowledge hierarchy provides a model that structures data, information, and knowledge in different levels and defines processes that transform each of these entities into another entity at a higher level [110]. Data are situated at the basis of the pyramid and viewed as atomic, discrete properties of an object we observe. For example, in an image database, a pixel value of an image is considered to be a data entity. Through aggregation and quantification, we derive information by giving meaning to similar data

entities. In an image database, low-level features extracted from images constitute information that can be used to further derive knowledge. Knowledge is viewed as information assimilated in a personal and subjective manner that can be used to make decisions. It can be derived either by instruction or experience. In a medical image database, semantics such as *cyst of lung* represent knowledge. Knowledge is further used to gain wisdom, which is a unique human feature that helps us integrate tacit knowledge into decision-making processes. Continuing with the previous example, only through wisdom can a radiologist expert connect the *cyst of lung* found in a high resolution computed tomography image to a pattern of disease. According to the knowledge pyramid, knowledge is normally more valuable than information, which is more valuable than data. Also, it is easier to turn information into knowledge that makes things happen, than it is to make the "right" things happen by turning knowledge into wisdom [151].

In an attempt to cope with data flooding, humans turn to technology for filtering the immense quantity of data and finding relevant patterns of visual information. Typical computer-based algorithms perform better at the lower levels of the *knowledge hierarchy*. This is due to the fact that computers have limited ability to synthesize visual knowledge or to discern the best outcome for each individual case. While it is very difficult to encode wisdom in computational ways, there is a need for computer-based knowledge management systems to encode human knowledge that is specific to individual users. In such a setting, computers can help humans by suggesting alternative solutions and by sorting out irrelevant cases to present only the cases that contain relevant information for knowledge discovery. Computer methods that describe visual patterns using only text annotations may not accurately describe complex visual patterns. Therefore, instead of plain text, Content Based Image Retrieval (CBIR) methods have proven successful in extracting image contents and providing query methods using low-level image features.

Many approaches in the context of CBIR have borrowed concepts and techniques from the related field of information retrieval. However, image databases are different from typical databases and many information retrieval techniques may not be successful when applied to CBIR systems. In his paper "Image Databases are not Databases with Images,"

Santini [116] explains the main characteristics of image databases. Although images carry messages about reality, they do not specify the context to which the image belongs. Images, in general, need to undergo an external validation process to be fully understood. Only after such extra analysis can the message in an image be predicated and fully described [115]. As Jonassen states, "It's not enough to know that. In order to know how, you must know why" [65]. He proposes the concept of structural knowledge to provide the conceptual bases for describing means of connecting declarative knowledge. However, most CBIR systems lack the flexibility to address the gap between computer and human models in describing visual patterns, which is an intrinsic part of modeling knowledge.

Figure 2 shows the knowledge pyramid for visual knowledge representation extended from [8] and using computer-based representation methods. There is a significant difference between the ways humans and computers model knowledge related to visual patterns. On one hand, humans evaluate information found in visual patterns to create successive mental models of the concepts encapsulated in the visual pattern. Then, by evaluating what these models stand for, according to personal experience, humans are able to extract the knowledge hidden behind the observed sign and make judgments about the newly acquired knowledge. On the other hand, many computers use algorithms to mimic human behavior. For example, data is transformed into information using feature extraction algorithms. The result of this process is usually a set of low-level features that can be used for machine learning or data mining. Unsupervised learning and data mining create high-level information that is returned to the expert for further evaluation. In such cases, the expert is responsible for extracting knowledge and reaching conclusions. Supervised learning uses previously ground truth information to extract knowledge in the form of associations between visual patterns (image features) and ground truth information (classes). However, as shown in Figure 2, the area between information and knowledge is blurred when using computer techniques due to the fact that most supervised learning techniques represent shared knowledge and do not dynamically customize it to the preferences of each user.

The most complex issue faced by current knowledge representation technology is the gap between the answer constructed by the computer and the mental model of the answer in the

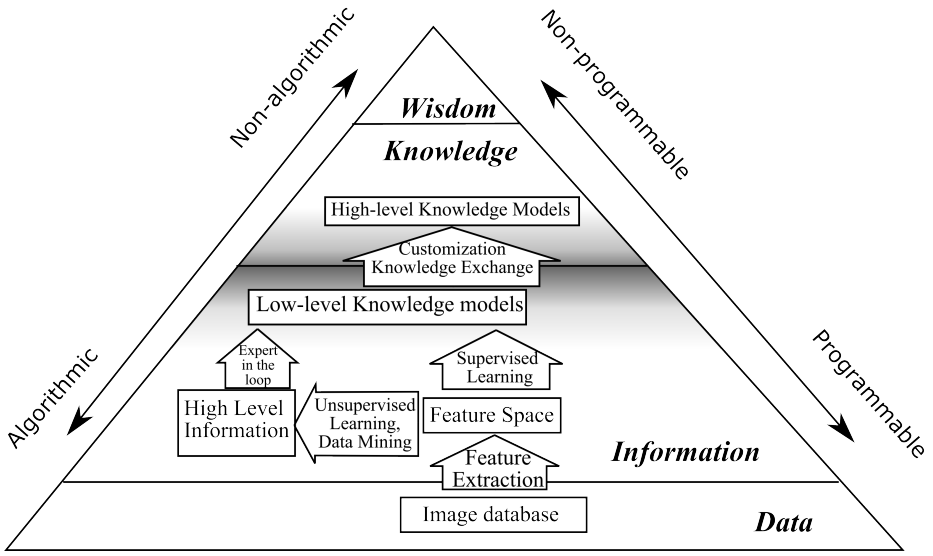


Figure 2: Knowledge pyramid adapted from E.M Awad [8] and applied to the computer-based knowledge representation.

analyst's mind. If this gap is wide, the user often becomes dissatisfied with the system and will eventually stop using it. The dissatisfaction appears even when the answer is relevant to other users of the system and is most likely attributed to the individual's resistance to change [28],[41]. To trigger the desire to change in computer-based systems, the analyst should be able to informally exchange knowledge with other peers who use the system. Such knowledge exchange may develop in situations where existing models are made explicit and challenged, and may lead to dissatisfaction with existing models. Dissatisfaction favors gradual conceptual change in users' mental representations of knowledge that cascades to other knowledge constructs [40]. When conceptual change occurs [66], it is more likely that users' acceptance of the system will increase.

In order to achieve a sufficiently accurate level of communication, a computer-based system should accommodate two basic conditions: (1) individual users should be able to express their representations of knowledge associated with the data stored by the system; (2) the system should provide methods for the users involved in communication to understand the subject of the exchange. If the system represents the knowledge using only a group or stereotype models, the users cannot articulate and challenge a particular representation of knowledge. Also, informal communication among peers should take place across possibly

heterogeneous information systems; this can be a challenge for computer algorithms.

1.2 Problem Statement

Extracting knowledge from large-scale image databases is challenging and requires a deep understanding of domain-specific knowledge. Image analysts' individualized models of visual patterns may not coincide with the models created by computer algorithms. To be successful, computer systems should adapt to the subjective views of image analysts. The overall objective of this work is to research methods of knowledge representation and exchange for large-scale image databases. More specifically we use semantic methods to model, represent, and exchange both explicit and tacit knowledge.

1.3 Our Approach

To achieve our research goals, we created a framework for a collaborative environment to share domain-specific image knowledge called Essence (Evolutionary System for Semantic Exchange of iNformation in Collaborative Environments). Our framework supports methods of individualized modeling of knowledge and peer-to-peer knowledge exchange methods. The foundation for knowledge representation and exchange is the use of domain-specific, user-defined semantics that describe the possible context of the knowledge hidden in the image. These semantics can be used to query other systems that use heterogeneous information models for decision support. In this environment, experts are not required to share identical semantic abstractions, although similarity is expected by the use of commonly agreed on vocabularies such as domain ontologies [55].

Although it is tempting to develop a single model of the group knowledge, we believe that it is important to maintain customized opinions of the various "views" of the information as seen by different users who participate in the knowledge discovery process. Our approach assumes that users in a local setting develop their unique conceptual models based on a stereotypical conceptualization model of the group knowledge. Even within a local setting, individuals may have a different vocabulary to express their perception of the domain knowledge. We believe that it is desirable to preserve the knowledge as possessed

by each user and to express it in a user-specific model. For example, in the radiology domain, physicians may use different descriptions for the same pathology due to differences in their training and geographical locations. The *tree-in-bud* (TIB) pattern is a direct CT scan finding of bronchiolar disease. The same pattern could also be called *Finger-in-glove* [126]. Our framework provides methods that ensure that all such individual preferences are preserved.

Our research objective is composed of five core major research goals, which determine the development of this work. They are as follows:

- Creating semantic models: These models are based on commonly defined vocabularies that map domain-specific semantics into low-level features extracted from images. Each user of the system should be able to create a customized model based on individual experience.

- Providing query by semantic methods: Users should be able to search similar images by domain-specific semantics or by example. When a user provides an example image, the system evaluates the relevant semantic abstractions to the image and retrieves similar images by semantics.

- Analyzing methods for semantic model customization: When users are dissatisfied with their semantic model, they should be able to customize it using user feedback methods.

- Providing semantic methods for knowledge exchange: In some difficult cases, multimedia users may need to reach for alternate opinions beyond the boundaries of their local organization. In such knowledge exchange processes, the system should be able to overcome the differences in knowledge representations.

- Investigating methods for semantic synchronization between two users or local settings: When knowledge exchange between two parties is not successful, the framework should provide an automated method to synchronize the semantic abstractions used in peer-to-peer knowledge exchange.

1.4 Main Contributions

The main contributions of this dissertation are the specifications and development of a semantic representation and exchange framework for domain-specific visual patterns. It offers methods for experts to refine their semantic settings on top of a shared ontology. This framework will be valuable for training and knowledge exchange. It can also be the foundation of building a novel and flexible model for an image retrieval system that uses expert-defined semantics. It accomplishes these tasks by assigning customized mappings of feature spaces to semantics and by adding new semantics to the knowledge represented by our computer model. Although the expert's decision-making process relies upon precise, scientific tests and measurements, it also incorporates subjective evaluations of visual patterns and relationships among human perception and semantic terms in a fuzzy and intuitive manner. The framework also facilitates knowledge exchange through peer-to-peer and centralized channels. There are four key components that make our work unique:

- Custom-defined linguistic variables closely related to known visual patterns: Our framework facilitates the management of domain knowledge in individual local settings, and offers an approach for modeling and integrating user-specific information into the shared model;

- Knowledge exchange and semantic setting customization: Our framework provides a novel method to support knowledge exchange among collaborative distributed systems that model information using domain-specific semantics;

- A generic implementation of the framework: The specified method will result in the implementation of a knowledge repository system that maps domain semantics into low-level features. A necessary and integral part of the environment is the development of methods of semantic customization and fusing individual models into the common conceptualization of the problem at hand;

- More desirable results: The customization methods which are provided by our framework provide for defining semantics is expected to better reflect the subjective view

of each image analyst.

Currently, there is no truly successful system for knowledge exchange among experts for differential opinion in image databases. We applied our framework to HRCT lung, geospatial, and plant domains and believe this approach is likely to be accepted by domain experts. With appropriate extensions, our framework can also be adapted to other domains that heavily utilize visual media.

1.5 Structure of the Dissertation

The dissertation is divided into six chapters. This introduction has outlined the objectives, problem definition, approach and contributions. Chapter 2 provides the background and context for the work presented in this dissertation by reviewing relevant state-of-the-art research in the domain of knowledge representation. Chapter 3 describes our proposed framework for knowledge representation using domain-specific semantics. Chapter 4 introduces methods of knowledge elicitation by proposing semantic ranking and query methods. It also addresses methods for improving the efficiency. Chapter 5 explains methods of knowledge exchange using peer-to-peer and centralized channels. Chapter 6 outlines experimental results. Finally, we conclude the dissertation and discuss future research directions.

CHAPTER II

LITERATURE REVIEW

Accurate knowledge elicitation is vital in decision making processes. In this process, experts do not use rigorous, well-defined formulas, rules, or laws, but rather they make inferences about the new case based on their previously acquired knowledge. This reasoning process makes it difficult for an expert to gain expertise that covers all the existing domain knowledge. In such cases, literature review or informal knowledge exchange is used to reach a decision. Computer models cannot represent such patterns in decision making due to the complexity of the human inference processes and of knowledge exchange. However, there is a need to develop systems to represent and exchange domain knowledge and to find computational ways to mimic experts' reasoning processes. Prominent researchers proposed solutions for knowledge representation and exchange. In this chapter, we overview the relevant work that has been done by other research groups. First we review research methods for knowledge representation and exchange. We then describe some associative learning algorithms for classification. In the end, we review domain ontology and known indexing techniques that are valuable resources for increasing the accuracy and time efficiency of our framework.

2.1 Review of Systems for Knowledge Representation and Exchange

In the past decade, researchers have been developing several prominent content-based image retrieval (CBIR) systems [27],[32],[69],[72],[95],[120],[31],[46],[78]. These CBIR systems mimic the domain knowledge to extract image contents and provide query methods for matching visual patterns using low-level image features. The prototype by Cai et al. [27] retrieved positron emission tomography images based on their specific physiological kinetic

features and developed a methodology of image compression that supports fast content-based image retrieval. Chu et al. [32] developed a semantic model for content based image retrieval for capturing the hierarchical, spatial, temporal, and evolutionary semantics of neural images in image databases. The system by Kelly et al. [69] associated each medical image with a signature for capturing textures and histograms of pathologies and retrieved images using query-by-example techniques. Fast query results for nearest neighbor search was addressed by Korn et al. [72] who used multidimensional indexing of medical tumors with similar shapes using an R-tree [57]. The system proposed by Nah and Sheu [95] used operational semantics to ensure the meaningfulness of content-based retrieval of neuroscience images. Robinson et al. [109] indexed shapes of cardiac boundary curves using a KD tree [14]. In the ASSERT system [120], Shyu et al. designed a suite of computer vision algorithms to extract visual abnormalities and used a multidimensional hashing approach to index pathologies of lung HRCT images. Chawla et al. [31] proposed an efficient approach for supervised spatial data mining that incorporates spatial properties of objects to discover interesting patterns embedded in spatial databases. Eklund et al. [46] used inductive learning techniques and artificial neural networks to classify and map soil types extracted from satellite images. Lees and Ritman [78] used decision tree induction methods for mapping vegetation types in areas where terrain and unusual disturbances confound traditional remote sensing classification methods.

The ultimate goal of these CBIR systems is to assist experts' evaluation of visual patterns. However, most of them use stand-alone knowledge and accomplish little to encourage knowledge elicitation and exchange among groups of experts. Knowledge, described as information with a productive component, is a very important aspect of the value generation process [131] in any organization. Tacit knowledge is an important part of human reasoning that evolves through human interactions with the surrounding environment. It was described by Mynatt in [108] as "the glue, texture, and backdrop for our interaction with people, places, and things." It can help experts reach conclusions when explicit knowledge fails to capture full explanations of a phenomenon but is very difficult to share, due to the human tendency to protect information that can give him or her a competitive advantage

over other members of the organization.

A major drawback of a system that tries to mimic experts' reasoning processes is the subjective assignment of the mapping between semantic terms and image features. If there is a significant discrepancy between the similarity, as assigned by the system, and the notion of similarity in the experts' minds, the results are destined to be unsatisfactory. Domain ontology can be used as a common framework for knowledge representation and exchange because it can connect image information to models of concepts used by experts to identify visual patterns. Leroy et al. [79] developed a tool (Medical Concept Mapper) for facilitating access to online medical information that uses human-created ontologies such as UMLS [19] and WordNet [89] to improve document retrieval performance. In the geospatial domain, Fonseca et al. [48] proposed an ontology-driven aerial information system for classifying geospatial images. Similarly in the bioinformatics domain, Stevens et al. [127] recognized the role of ontology in knowledge driven research domains and propose methods of building ontology that can be used within bioinformatics. The approach in [124] proposed a protein ontology to structure the representation of proteomics data with benefits in studying relationships among proteins and cellular functions of proteins. Wang et al. [138] proposed a multi-modality ontology model that integrates both the low-level image features and the high-level text information to represent image contents for retrieval for the canine animal domain. For a more comprehensive survey of the use of ontology in knowledge representation, readers are encouraged to read [23]. The use of ontology requires consensus on the ontological definitions among the community members to reduce ambiguities in communication. However, such consensus may limit the individual user's ability to view the knowledge according to his or her specific expertise. For this reason, experts should be able to customize their individual semantic terms in order to create a human-friendly environment for decision support.

Models used by experts for evaluating visual patterns are not always binary: existing or non-existing. In practice, there is no hard boundary that separates two visually similar semantics, such as many and few nodular opacities. If a crisp threshold of a low-level feature is set to distinguish two semantics, the threshold is always subjective and may not calibrate

what is in an expert's mind [42]. Fuzzy logic could be a good tool to handle this subjectivity of semantic assignments. Some approaches in the domain of general image retrieval, such as [3],[87],[92],[113] tried to implement fuzzy logic concepts to increase the meaningfulness of the retrieval results. Aguilera et al. [3] developed a model for fuzzy image retrieval by expressing image features and user queries in terms of fuzzy sets. Madasani et al. [87] represented image regions and queries as fuzzy attributed relational graphs and used an efficient fuzzy algorithm for matching them. Mouaddib et al. [92] developed a fuzzy relational schema that assigns to each tuple a degree of compatibility with the fuzzy constraints defined on the relationships. Saint-Paul et al. [113] applied fuzzy semantic hierarchies and relationships among terms. Techniques proposed by the fuzzy logic researchers will be valuable for modeling the knowledge representation environment to be able to integrate with experts' individual preferences.

Knowledge exchange in domain-specific environments is difficult, especially due to the autonomy of local organizations and the importance of the tacit component of the knowledge [86]. Domain experts, who usually carry this knowledge, have close concordance with their local setting, in which both previous experience and colleagues' opinion have a major influence. However, local knowledge is often limited and insufficient to deal with difficult, new cases that have not been previously evaluated [63]. The tradeoff between knowledge value and elicitation effort becomes very important because experts typically have a limited amount of time to respond to a case or to share expertise with peers. Looking for knowledge beyond the local setting is necessary but difficult due to the differences in group cultures and locally defined methods of encoding information into semantics.

Several systems that focused on knowledge exchange have been developed [49],[50],[70]. Fox and Thompson [49] proposed a unified technology for clinical decision support and disease management that emphasizes integrated methodologies for developing clinical applications. Gardner et al. [50] designed a framework, using XML-derived schemas, that defines an interoperability standard for neuroscience informatics resources. The knowledge exchange framework developed by Kindberg et al. [70] addressed the issue of communicating through peer-to-peer networks as well as methods of facilitating data and knowledge

exchange. While it is true that knowledge-based systems cannot perform better than human experts [9], they are capable of filtering the information to be presented to experts for diagnoses. The approach by Economou et al. [44] proposed a computer-aided medical system that allows a human-in-the-loop, step-by-step procedure for approximating the final diagnoses in various fields of medicine. In the radiology community, knowledge sharing is more complex than other medical domains because it is very difficult to accurately describe visual patterns using plain text annotations. Therefore, instead of plain text, systems require a common base to share and exchange knowledge related to visual content of the abnormality present in diagnostic medical images. That is, no matter what annotations are associated with the images, if two medical images share similar visual abnormalities, identified by experts, they should also share similar visual contents detected by computer algorithms.

Peer-to-peer networks have proven to be successful in tacit knowledge elicitation and exchange by facilitating alternative opinions and revisions [108]. Knowledge exchange through these networks has a very high knowledge creation potential due to its capabilities of creating strong temporary connections. These connections are derived from existing weak ones to maximize knowledge generation by connecting disparate groups of users in a common environment. It is rare to see a CBIR system that encourages experts to define their own semantics to the database and to adapt individual preferences of semantics to the common knowledge base. These aspects become important in domain-specific applications because concepts, with their empirical characteristics, are subject to a continuous semantic and conceptual adaptation [26].

In this dissertation, we propose a knowledge discovery procedure to determine complex association rules among visual semantics and image features. This approach, which easily adapts to users' preferences, emphasizes the use of flexible semantics to analyze visual patterns found in images. It enables fast retrieval of images using semantic queries that are more relevant to users through the use of content-based image retrieval methods and domain ontologies. It also enables automatic knowledge exchange inside and across distinct local settings. The system will be valuable for computationally interpreting visual patterns

in domain-specific environments that use semantics to model visual patterns.

2.2 Supervised Learning Based on Association Rules

Association rule mining is a widely-used approach in data mining. It is capable of revealing interesting relationships in large databases. These relationships can be used both for describing the relationships in the database and for classifying database instances. Classification Based on Association Rules Methods (CARM) help solve the comprehension, interestingness and scalability problems that plague existing classification methods. The main issue of CARMs is that the size of the generated classification association rules (CARs) is 2^d for a d-class problem, which can make the problem intractable even for moderately sized databases. To address this issue, existent approaches define measures of relevance to guide the search and reduce the search space.

2.2.1 Classification Based on Associations

The Classification Based on Associations (CBA) algorithm integrates association rule mining and classification by generating association rules having the consequent class - Class Association Rules (CARs). A classifier is then built based on the generated set of CARs. The algorithm first generates all of the CARs with a user-specified minimum support and minimum confidence. An optional pruning phase reduces the size of the set of CARs.

To build a classifier based on the generated CARs, CBA uses two techniques. The first technique ranks the generated CARs based on their confidence and then applies a database coverage technique for pruning. The first step in pruning by database coverage is to rank the generated rules by a quality measure such as confidence. Then it counts the number of training instances that are covered by a newly generated association rule that were not previously covered by higher ranked association rules. Only high quality rules that cover at least one training data instance are retained for classification. Pruning by database coverage introduces a bias toward CARs with high confidence but low support, which may lead to an overfitted model. A second technique generates a count of the total errors made by each possible classifier. It then finds the first rule in the set of ordered CARs that has the lowest error count and discards all the more specific rules. When an unknown case is presented

to the classifier, the first rule that satisfies the case will be used. If there is no rule that applies, the default class will be taken.

The drawback of the CBA algorithm is that it generates a huge amount of rules (especially before pruning) and that it requires $|R|$ passes through the database, where $|R|$ is the number of rules generated.

2.2.2 Apriori Total-from-Partial

Apriori Total from Partial (Apriori-TFP) [34] is an "apriori" style algorithm designed to process a binary valued input dataset so as to identify frequent itemsets. Instead of operating with the raw input data directly, the input data is first preprocessed and placed in a set enumeration tree that we call the P-tree (Partial support tree). Then, the algorithm stores the resulting frequent itemset information in another form of set enumeration tree called a T-tree (Total support tree). This latter tree can then be processed to identify association rules relevant for classification. Experimental work shows that Apriori-TFP is advantageous when operating with data that features many duplicate records and with records with duplicate leading sub-strings. However, it does not address all the other issues that are characteristic to apriori-like methods.

2.2.3 Classification Based on Multiple Association Rules

Classification based on Multiple Association Rules (CMAR) uses a weighted chi-square analysis for classification using multiple association rules. Given a new data object, CMAR collects the subset of rules matching the new object from the set of rules for classification. These rules may not be consistent with the class labels. CMAR first groups those according to class labels. Then, a "combined effect" is accounted for each group by adopting a weighted chi-square as the measure to determine the final class membership of the object.

Although CMAR reduces the number of database scans by using the FP-tree data structure [60], it does not address the size of the generated association rules and their relevance in classification problems.

2.2.4 Classification Based on Predictive Association Rules

Classification based on Predictive Association Rules (CPAR) is an extension of the Predictive Rule Mining (PRM) algorithm [149]. It is a hybrid technique that uses both associative classifiers and rule-based classifiers. It uses a greedy algorithm to search the space of attributes. The main difference from the rule-based approaches is that it keeps all close-to-the-best attributes in rule generation, unlike rule-based methods that only use the best attribute. Rules are built by assessing the gain in positive examples of additional literals to a current rule, with literals giving the highest gain used to create new rules. The major distinction between PRM and CPAR is that CPAR is able to choose a number of attributes if they have similar best gains, whereas PRM and FOIL [81] choose only the attribute that displays the best gain on each iteration. After the set of rules is created, each rule is evaluated by calculating its expected accuracy. For classification, the best rules for each class are determined by the expected accuracy. These values are then averaged, and the class with the highest average accuracy is used for the classification.

2.3 Domain Ontologies

Every human has a unique view of the surrounding environment. This view is encapsulated in mental models that are triggered by signs found in day-to-day life and is called the conceptualization of reality. These mental models are dynamic and evolve with experience or training. However, to facilitate communication of knowledge, domain experts use a common understanding in terms of the language. These terms are assumed to be a shared vocabulary that rely on commonly agreed meanings of concepts that have little variety among experts. The use of such shared conceptualizations of reality provide terminologies that can be used for communication. A shared conceptualization is never universally accepted, but is only valid for a limited number of persons committing to it.

An ontology is a controlled vocabulary that describes concepts from a knowledge domain and their relations in a formal way. Gruber defines an ontology as a "specification of a conceptualization" of a knowledge domain [56]. As a branch of philosophy, ontology answers the question *what is* and addresses the kinds and structures of objects, properties, events,

processes, and relations in every area of reality. Information systems consider ontology to be a kind of agreement on a domain representation. This representation includes a set of non-exhaustive concepts, their definitions, and their inter-relationships. Ontologies are designed to overcome the problem of implicit and hidden knowledge by making the conceptualization of a domain explicit. There are three main characteristics of domain-specific ontologies: (1) concepts and their relations should be explicitly defined, (2) ontology should be machine understandable, and (3) ontology should capture consensual knowledge of a group [56].

Knowledge can be represented in ontologies using five knowledge facets : concepts, relations, functions, axioms, and instances. Ontological concepts represent the main constructs and define the conceptualization of a domain knowledge. However, an ontology is more than a list of concepts. Although ontologies may differ vastly, they define relations between concepts. For example, taxonomical ontologies are hierarchical models of a domain knowledge in which concepts are organized using sub- and super-relations. Ontological representations of knowledge may also be used to differentiate between subtypes and their super-types using axioms.

A great number of domain-specific ontologies were developed for describing domain-specific knowledge. The selected list of existing ontologies that we present below covers a wide range of domains and shows the relevance of using ontologies in representing domain-specific knowledge.

- The Agricultural Ontology Service [80] attempts to standardize agricultural terminology in multiple languages related to the agricultural community.

- Galen [132] is a medical terminology server for supporting the development of clinical coding schemes. The ontology used by Galen is intended to serve a wide variety of medical applications such as surgical procedures, electronic health-care records, clinical user interfaces, decision support systems, knowledge access systems, and natural language processing.

- The Gene Ontology [6] is a dynamic controlled vocabulary that describes the molecular function, biological process, and cellular location of gene products.

- The Ontogeo Geospatial Ontology [67] is a repository for research and teaching activities targeted primarily to researchers.
- The Open Biological Ontologies [94] project manages a set of well-structured controlled vocabularies for shared use across different biological domains. Among these ontologies are Arabidopsis, Cereal, Plant, and Maize ontologies.
- The Phenotype and Trait Ontology (PATO) [54] is an ontology of phenotypic qualities, intended for use in a number of applications, primarily defining composite phenotypes and phenotype annotation.
- The Sequence Ontology [45] describes both raw and interpretations of features on nucleotide or protein sequences.
- The Trait Ontology [148] is a controlled vocabulary that describes each distinguishable feature, characteristic, quality, or phenotypic feature of developing or mature individuals.
- The Unified Medical Language System (UMLS) [19] provides a biomedical vocabulary from disparate sources such as clinical terminologies, drug sources, vocabularies in different languages, and clinical terminologies.

2.4 Indexing Large Databases

With the advancement of technology, experts are flooded with information to the point that they frequently cannot mentally process it in a timely manner. They have to rely on advanced information systems to computationally filter information. In an ideal case, a computer-based system should be flexible enough to provide query methods that are meaningful to the user. In such a multimedia retrieval process, the user provides the desired query criteria to the system and expects the system to provide the most relevant media from the database.

It is well recognized that content-based approaches are promising for multimedia retrieval. These approaches apply computer vision and image processing algorithms to extract characteristic features from database media. The features are then mapped into a

high-dimensional feature space that is used for similarity search. When dealing with high-dimensional and high-cardinality datasets, indexing becomes an important factor in overall performance. In such cases, brute-force algorithms perform a large number of evaluations to assess the similarity of all of the media to a query. This makes a brute-force search computationally infeasible for real-time retrieval systems even when using state-of-the-art computational power.

Searching high-dimensional feature spaces using indexing structures is affected by some qualitative effects [20] known as the "curse of dimensionality". These effects have severe implications for the performance of the indexing structures. First, the volume of the space increases exponentially with the number of dimensions. Also, the number of selected data points decreases exponentially with the size of the feature space [140]. The most important consequence is that a query may search parts of indexing pages that do not contain relevant points, an issue known as "dead space" [20].

The most widely used approaches to the high-dimensional indexing problem employ tree approaches [15],[33],[57],[114],[143]. They use hierarchical clustering of data in which similar vectors are stored in the same or neighboring nodes. However, these approaches are not capable of performing range queries that filter only a subset of features. Range queries are important for systems that search by semantics [122]. To address this issue, the pyramid strategy [16] maps high-dimensional points into a one-dimensional space and uses a B^+-tree to index the embedded space. Weber [140] proposes the vector approximation files technique that divides the space into hyper-cubical subspaces with a unique key. In a similar fashion, space-filling curves can be used to index media data by mapping them from a high-dimensional space into a one-dimensional space. The feature space is divided into hyper-cubical subspaces, and each subspace is given a unique index on a space-filling sequence. Data points are then approximated to the center of the hypercube in which they reside and indexed in a B^+-tree. A drawback of using space-filling curves for indexing high-dimensional data is that the number of mapping points may exceed the number of data points. In such cases, query efficiency is reduced by searching many empty mapping points.

In this chapter, we have presented related research that is relevant to knowledge representation and to exchange of visual patterns found in images. As shown above, the huge amounts of images captured on a daily basis by research groups makes this an area of much interest. To cope with this flood of data, researchers have focused on methods of mimicking the domain knowledge and providing query methods by image content. Also, research has been conducted in formalizing the domain-specific knowledge and incorporating it into computer models of knowledge representation. However, there is a need to provide methods for knowledge elicitation and exchange among groups of experts. To accomplish this, systems need to provide methods for representing knowledge according to the unique view of each expert and for sharing this unique view through peer-to-peer and centralized channels.

CHAPTER III

KNOWLEDGE REPRESENTATION

In this chapter, we address methods of knowledge representation that are suited for knowledge exchange. First, we introduce a model of knowledge representation that mimics experts' models about visual patterns found in domain-specific images. This representation model uses semantics organized in a hierarchical structure. Then, we provide two procedures of knowledge discovery that map these high-level semantics into features extracted from images. The first procedure uses high-level, complex features to build semantic models; the second procedure applies data mining techniques to discover association rules between semantics and low-level features. These procedures are expected to minimize the semantic gap between human and computer-based models of visual patterns.

3.1 Knowledge Representation Framework

Most of the decisions in the expert domains are made by comparing the data at hand against existing domain knowledge. During the decision-making process, experts usually base their evaluations on a set of heuristics developed from different areas as a "multi-dimensional intuition" [107] in which tacit knowledge plays a very important role [22]. Web-based knowledge management systems have the unique feature of going beyond the typical boundaries of groups of experts [96]. They have to deal with different settings for users and with complicated information exchange procedures [74]. Such systems should effectively support most of the strategies used by experts so the decision process is not constrained [107] by system limitations. Brown et al. [24] described a knowledge-based approach to HRCT image segmentation by using anatomical structure and various domain-specific knowledge. The knowledge based approach developed by N. Tayar [130] focused on data consistency and incremental development by dividing the knowledge base into layers. The model developed by Wei et al. [141] focused on representing the complex heuristics and

data-intensive knowledge specific to the medical domain that facilitates interactions among heterogeneous and autonomous medical data sources. In the geospatial domain, Fonseca et al. [48] proposes an ontology-driven aerial information system for classifying geospatial images. In the bioinformatics domain, Kazic [68] proposed a method of knowledge exchange among disparate systems using semantics, called semiotes, that axiomatically define simple domain definitions. All of these approaches bring novel ideas to knowledge management, but do little to address means of customizing the settings to the users' preferences.

The goal of our knowledge representation and exchange framework-called Essence-is to provide a visual and cooperative environment that facilitates knowledge exchange in more intuitive ways. As depicted in Figure 3, the main components in Essence are (1) Semantic domain, (2) Image space, (3) Feature extraction algorithms, (4) Feature domain, (5) Preference domain, (6) Query system, and (7) Information exchange module. In this figure, knowledge components are represented in rectangles, and knowledge-driven actions, such as search and discovery, are represented in oval shapes.

The Semantic domain is organized as a local-as-view data integration subsystem [36]. This system lets users build, refine, and further decompose their semantics independently, with minimum effort, on top of the shared ontology [13],[11],[52]. The shared ontology is to be exploited by all users who access the system. Along with the Preference domain, the Semantic domain represents the expert's knowledge in an XML format. Using a similar format, the framework represents the knowledge of a specific case, a medical image, in the Feature domain. Each element in the Feature domain is a signature of a domain-specific image in the Image space. The signature is computed by executing the feature extraction algorithms designed by computer vision and image processing researchers [120],[122],[121]. The Query system searches the knowledge base, selects relevant images, and translates the result into a human-readable format. It provides two mechanisms to access the knowledge base: 1) query by semantics; and 2) synchronization of semantic terms. The Information exchange module facilitates knowledge exchange among users through peer-to-peer and centralized channels.

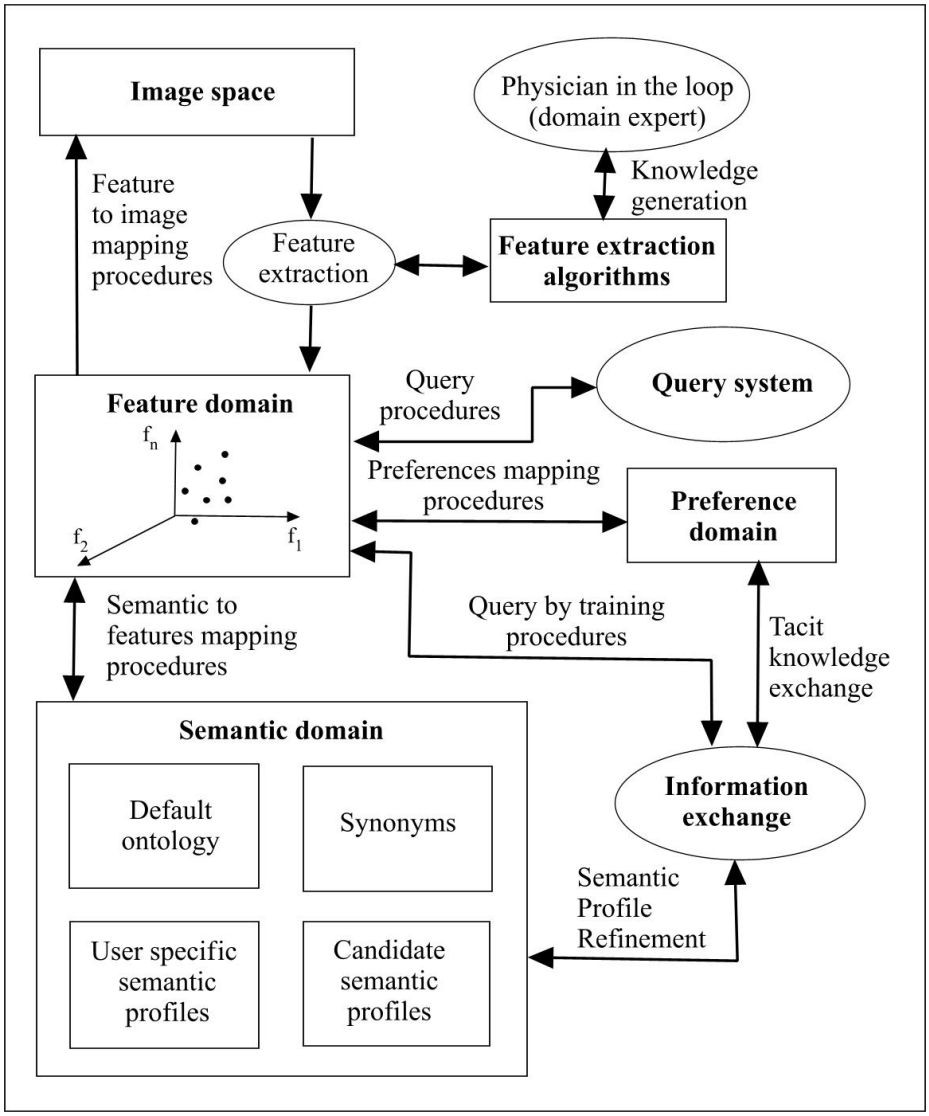


Figure 3: Architecture of the knowledge representation and exchange framework in Essence.

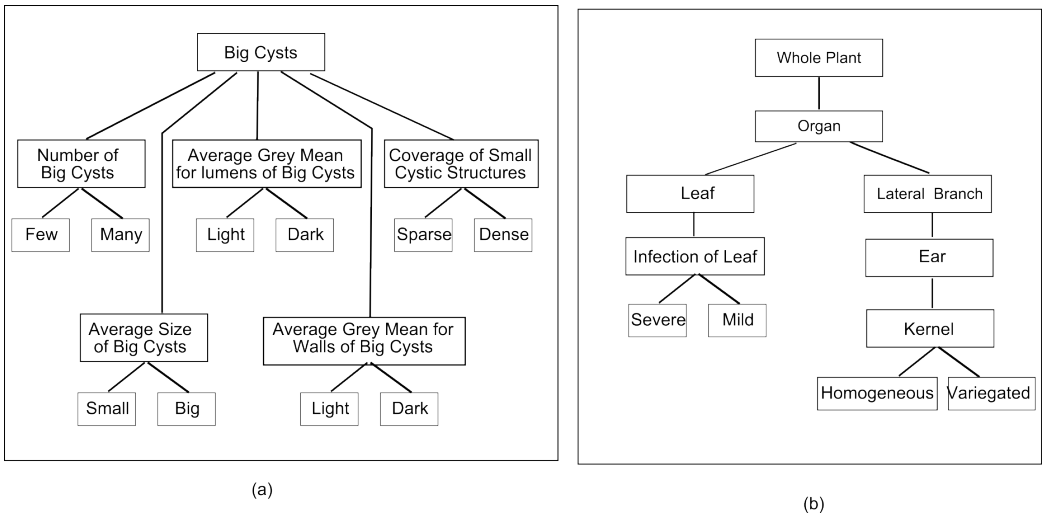


Figure 4: Examples of hierarchical structure of domain-specific linguistic variables and semantic terms from the (a) radiology domain and (b) plant domain

3.1.1 Semantic Domain

Experts use several perceptual categories for recognizing associated visual patterns in domain-specific images. We define linguistic variables to model those perceptual categories used by experts. Each of these linguistic variables is assigned a set of semantic terms that represents a semantic assignment for a visual pattern. The linguistic variables and their semantic terms are arranged in a hierarchical structure as depicted in Figure 4. This figure shows two examples of hierarchical organization of linguistic variables from the radiology domain (Figure 4(a)), and plant domain (Figure 4(b)). According to the structure in Figure 4(a), the linguistic variable *Number of Big Cysts* has been assigned a semantic term set {*Few*, *Many*}. Similarly, for the plant domain shown in Figure 4(b), the linguistic variable *Kernel* is assigned the semantic term set {*Homogeneous*, *Variegated*}. The union of all hierarchical structures of linguistic variables constitutes a semantic profile that is used to query the image space. To simplify the explanation of the concepts, we use examples from the radiology domain for the rest of this chapter.

3.1.2 Linguistic Variables

The linguistic variables, as defined by experts, are tuples l in the form $< u, c, c1, \delta >$, where u is the user that defined the linguistic variable, c is an indexing code, $c1$ is the indexing

code of the parent linguistic variable if any, and δ is a description of the linguistic variable. For example, the instance of the knowledge base in Table 1 shows some linguistic variables defined by user *adrian* for the radiology domain. The linguistic variable *cysb* is described by the tuple $< adrian, cys, lngs, \text{Cysts} >$, and has the meaning:"User *adrian* describes the *cysb* linguistic variable as *Cysts*". This linguistic variable is defined in the semantic tree as a descendant of the linguistic variable *lngs* (*Lung pathologies*).

Table 1: An Instance of Knowledge Base For Storing Linguistic Variables

User	Code	Parent code	Description
adrian	lngs		Lungs pathologies
adrian	cys	lngs	Cysts
adrian	cysb	cys	Big cysts
adrian	cysbn	cysb	Number of big cysts
adrian	cysbs	cysb	Average size of big cysts

3.1.3 Semantic Terms

The semantic terms associated to a linguistic variable are defined as tuples ς with the form $< u, c, l, \delta, G >$ where u is the user that defined the semantic, c is an indexing code, l is the linguistic variable to which the semantic term is attached, δ is a description of the semantic term, and $G = \{g_1, g_2, \ldots, g_r\}$ is the description of a set of functions that define the semantic assignment for lung pathologies. The assignment of lung pathologies is done by specifying a matching degree to all linguistic variable measurements in relation to the semantic term used.

When adding new user semantic terms, our system follows the principles for designing ontology: 1) parsimony - semantic terms are added only if strictly necessary, 2) clarity - semantic terms should effectively communicate the intended meaning, and 3) coherence - all new terms should be locally consistent. Also, each semantic term should be mapped to a normalized possibility distribution. That is, there should exist at least one image that fully matches the semantic term [38].

For example, the first row of Table 2 lists a semantic term with the following attributes: *cysbnf* - indexing code and *Few big cysts* - description. This term is a child node of the

linguistic variable *Number of big cysts* (indexing code *cysbn*). In this case, the semantic assignment is described by a function that will be explained in the next section. This function has the type *RB*–right-bounded, maps the semantic *cysbnf* into the *fea1* feature, and is described by a series of coefficients λ.

3.1.4 Semantic Profiles

To adapt itself to users' preferences, our system creates four semantic profile types: *default*, *candidate*, *user-specific*, and *working*. The first two are designed for all users; the last two for an individual user. In this semantic profile, each non-leaf node holds a linguistic variable as described in Section 3.1.2, while a leaf-node holds a semantic term as described in Section 3.1.3.

For a new user or knowledge depositor, the *user-specific* profile is initially empty and the user inherits the semantic setting from the *default* one which contains all of the linguistic variables and semantic terms commonly agreed to by the existing users. Also, the new users have access to all of the other semantic terms added by other users by using the *candidate* profile which is updated only when new linguistic variables or semantic terms are added to the system. When a user customizes his or her personal settings, the new semantic assignment function is saved in the *user-specific* profile. To retrieve database images by semantics, a *working* semantic profile is created on the fly. This profile inherits all of the linguistic variables and semantic terms from the *default* profile and appends all new variables and terms from the *candidate* profile. However, settings in the *user-specific* profile are mandatory to overwrite those in both *default* and *candidate* profiles. Figure 5

Table 2: An Instance of Knowledge Base For Storing Semantic Terms. Shown is an example of domain concepts from the domain of high resolution computed tomography of lung.

Code	Description	Parent variable	Semantic Assignment Function
cysbnf	Few big cysts	Number of big cysts	$< RB, \{(\text{fea1}, \lambda_R^1 = 4, \lambda_R^2 = 4, \lambda_R^3 = 2.1)\} >$
cysbnm	Many big cysts	Number of big cysts	$< LB, \{(\text{fea3}, \lambda_L^1 = 12, \lambda_L^2 = 5, \lambda_L^3 = 1.8)\} >$
cysbsav	Big cysts	Size of big cysts	$< RB, \{(\text{fea3}, \lambda_R^1 = 11, \lambda_R^2 = 3, \lambda_R^3 = 0.9)\} >$
cysbsv	Very big cysts	Size of big cysts	$< LB, \{(\text{fea3}, \lambda_L^1 = 17, \lambda_L^2 = 3.5, \lambda_L^3 = 2)\} >$

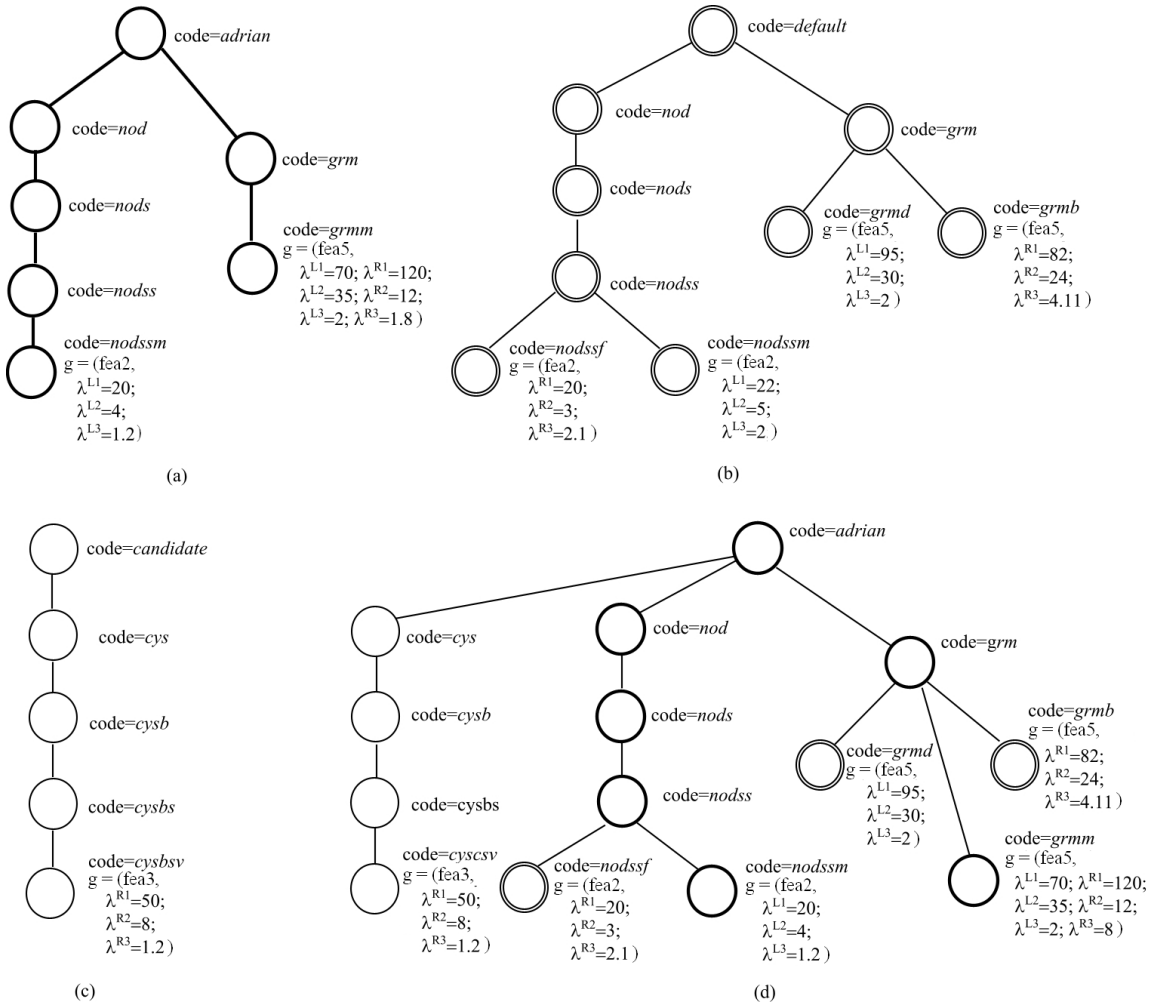


Figure 5: Example of semantic profiles. The working semantic profile shown in (d) is the result of combining the user-specific profile (a), the default profile (b), and the candidate profile (c).

shows the process of building a working profile for the user *adrian*. The *working* profile shown in Figure 5(d) inherits the default profile with double-circle nodes and appends the *candidate* profile with a thin-circle node. The bold single-circle nodes are from the *user-specific* profile.

3.1.5 Image Space and Feature Domain

The raw information processed by our system is a collection of domain-specific images. For each new image in the database, feature extraction algorithms [123] are applied and an image feature profile is created. This profile has a hierarchical structure similar to the combined structure of the *default* and *candidate* profiles, which were discussed in Section 3.1.4, except that the semantic terms are replaced by feature measurements. The knowledge base describes image features as tuples f with the form $< \iota, f, m >$, where ι is an image, f is the feature extracted by a feature extraction algorithm, and m is the measurement assignment for visual feature. The example in Table 3 shows an instance of a knowledge base that stores information of image features. In this example, the value for the linguistic variable *Number of small nodules* of image id Essence-01 is 25.

Table 3: An instance of a knowledge base for storing features of high resolution computed tomography images of lung.

Image	Feature	Measurement
Essence-01	Number of small nodules	25
Essence-01	Average size of small nodules	1.55
Essence-02	Number of small nodules	44
Essence-02	Average size of small nodules	0.77

3.2 Mapping High-level Features into Semantics

In this section, we discuss our approach of mapping the complex image features to semantic terms. We start by giving some examples of such complex features for HRCT images of lungs. Then we explain the mapping procedure.

3.2.1 Algorithms for Extracting High-level Features

To extract high-level semantics from the raw images, a suite of computer vision and image processing algorithms are designed to identify visual patterns in domain-specific images. One of the most important abilities of humans is to perceive and interpret spatial object in the surrounding environment such as shape, size and orientation of objects [98]. The perception of these spatial structures extend also to evaluating two-dimensional images of the surroundings.

To have a compact presentation of the main theme of this dissertation in knowledge sharing and semantic modeling, we only briefly discuss the algorithms that were designed to extract two perceptual categories (out of 24): small nodular opacities and cystic structures.

1. Algorithms to extract nodular opacities: An example lung disease resulting in nodular opacities on HRCT images is sarcoid [88]. Important features to describe nodule opacities include: 1. the gray values associated with nodules since the values carry important information with regard to whether the tissue is benign or malignant; 2. the size and spatial distributions associated with the nodular opacities; and 3. the roundness of high grey-level objects. To extract image features related to nodular perceptual category, we have implemented the following procedure:

   - Extract the lung regions [118] and apply Otsu thresholding [97] on them.
   - Apply labeling to high pixels.
   - Compute the roundness of labeled objects by $roundness = 4*Area/(\pi*Diameter^2)$
   - Group labeled objects into small nodules and big nodules based on two measurements: roundnesses and sizes of labeled objects. Both thresholds were learned from the training data.

   Effective feature measurements for images with this type of pathology include: 1. Number of small nodules, 2. Roundness mean of small nodules, 3. Grey mean of small nodules, 4. Average nearest neighboring distances (NNDs), 5. Standard deviation of NNDs, and 6. Histogram of NNDs partitioned into six bins. In this dissertation, we

show the semantic term derived from the first feature. Other semantic terms, such as "uniformly distributed small nodules" and "skew distributed small nodules" are modeled by using features 4 to 6.

2. Algorithms to extract cystic structures: To identify the presence and absence of cystic structures, we have applied the following procedure:

   - Extract the lung regions and apply a dual-thresholding on the regions to highlight potential cyst walls from high pixels and possible lumens from low pixels.
   - Set all pixels outside the lung regions to 0. These outside pixels and pixels of the lumens have the same gray value.
   - Apply watershed algorithm [111] to repair the broken cystic walls.
   - Apply component labeling to high pixels.
   - Compute the sizes of the labeled objects and the average grey-scale mean difference between labeled objects (potential walls) and pixels bounded by the labeled objects (possible lumens).
   - Prune out those labeled objects that fall at least one standard deviation away from the means of the sizes and grey mean differences of training cystic structures.

   Effective attribute measurements for images with this type of pathology include: 1. Number of cystic cells, 2. Average size of cells, and 3. Coverage of cystic structures within the lung regions.

A more comprehensive discussion for all perceptual categories used in Essence can be found in [118]. Each perceptual category studied in this dissertation has a set of relevant image features which were tested by Multivariate Analysis of Variance (MANOVA) and empirically proven to be efficient [120] to distinguish categories from each other. A multi-dimensional feature vector is then formed for each raw image. Whenever a new linguistic variable is defined, Essence either reuses the existing algorithms or asks the computer vision/image processing researchers to develop a new feature extraction algorithm that is dedicated to this new variable.

3.2.2 Mapping Procedure

The mapping process uses three types of information: (1) semantic information, (2) image feature information, and (3) user preferences. The possibility distribution that maps semantic terms to image features is expected to capture users' preferences in a computational way. Mitiam et al. [90] analyzed different types of shapes in fuzzy set theory by testing how these shapes can approximate different testing functions. Although the best shape is subjective and data/application dependent, this research concludes that there are set functions that could approximate better than the triangular or trapezoid ones.

For the purpose of our model, we extended Mitiam's research by adding an asymmetric property to the possibility distributions of semantic terms for perceptual categories. This property is believed to be better in fitting users' semantic preference than commonly used symmetric functions. In depth evaluation for using such asymmetric property is described in Chapter 6. There are three parameters that control the shape of the possibility distribution: (1) the center of the function (λ^1), (2) the width (λ^2), and (3) the exponential factor (λ^3). For example, in Figure 6, the sigmoid part of the possibility function noted A, has the parameters, $\lambda^1 = 10$, $\lambda^2 = 6$, and $\lambda^3 = 2$.

Each possibility distribution is used to model a semantic term for a perceptual category, which is presented by a linguistic variable. Let L be the set of linguistic variables assigned to a database image, $\varsigma =< u, c, l, \delta, \upsilon >$ be a primitive semantic term defined by a user for the linguistic variable $l \in L$, and $NOT(\varsigma) =< u, c, l, \delta, 1 - \upsilon >$ be the Boolean function that is true when semantic term ς is absent from a query. The semantic term ς associates to the linguistic variable l a possibility distribution function as: $a_\varsigma : R^l \rightarrow [0, 1]$ defined by a user over the universe R^l of the linguistic variable l. For example, we can define the semantic term *Average number of big cysts* (indexing code *cysbsa*) for the linguistic variable *Size of Big Cysts* (indexing code *cysbs*) $l =< adrian, cysbs, lngs, Cyst >$ as $\varsigma =< cysbsa,$ Average number of big cysts, $cysbs, g_{bcs}(m) >$. Figure 7 shows an example of possibility distribution for the semantic term ς and the degree of significance (53.79%) of the measurement $m = 28$ to ς (*Average number of big cysts*).

We define three types of possibility distributions to model semantic terms shown in

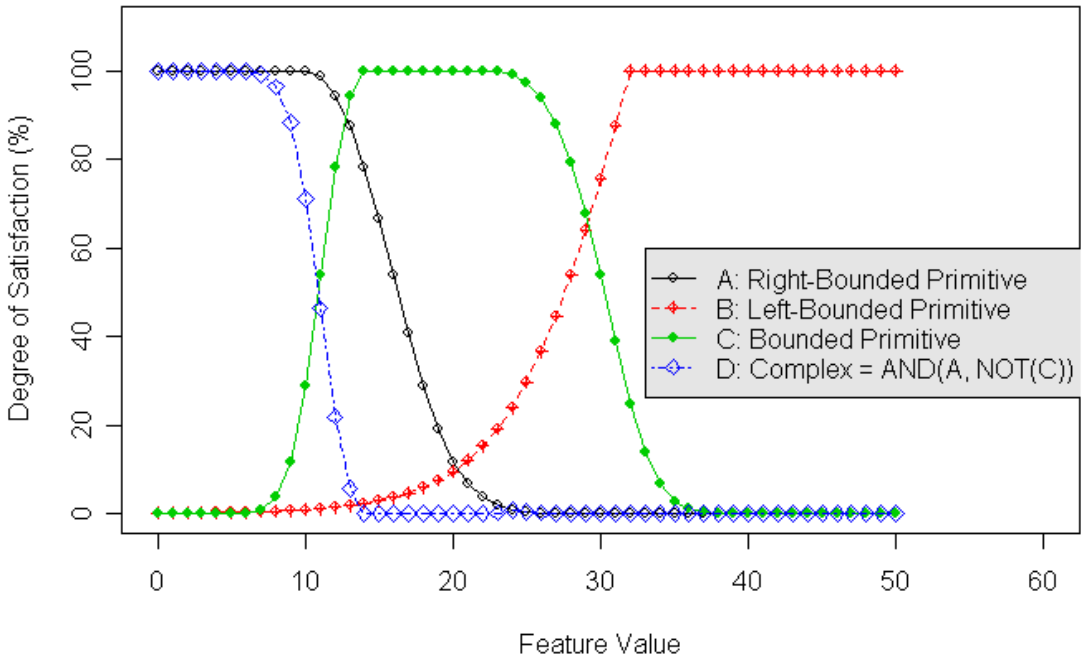


Figure 6: Example of possibility functions.

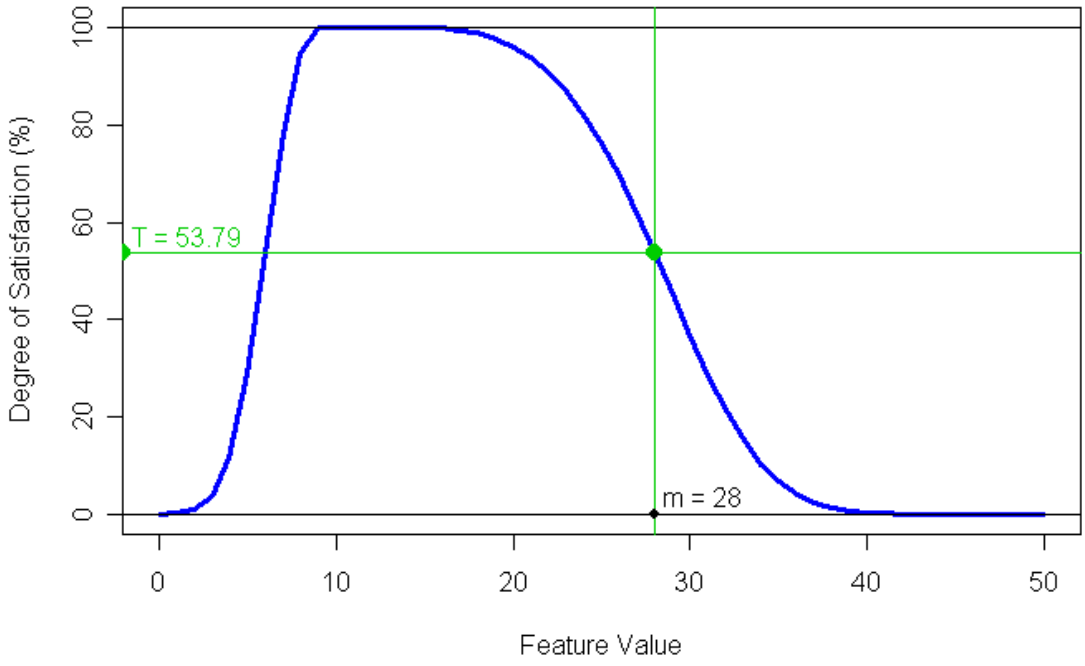


Figure 7: Computation of the degree of satisfaction.

Figure 6: Left-bounded primitive, Right-bounded primitive and Bounded primitive. We also define a complex semantic term that is composed of multiple primitive terms concatenated by logical operations.

3.2.3 Left-bounded Primitive Semantic Term

The left-bounded primitive semantic terms assign a full degree of significance to all measurements that are greater than a specified value. Semantic terms such as *big*, *many*, and *huge* fall into this category. The following equation is used to model this type of primitive semantic term:

$$g(m) = \begin{cases} \frac{2}{1+e^{\left((\lambda_L^1 - m)/\lambda_L^2\right)^{\lambda_L^3}}} & for \quad m < \lambda_L^1, \\ 1 & for \quad m \geqslant \lambda_L^1. \end{cases} \tag{1}$$

It is defined as the union of a constant function and a sigmoid function. The sigmoid function is centered at λ_L^1, has width λ_L^2, and an exponential factor λ_L^3. The degree of

significance of the left bounded primitive semantic term equals 1 for any measurement $m \geqslant \lambda_L^1$.

3.2.4 Right-bounded Primitive Semantic Term

The right-bounded primitive semantic terms assign a full degree of significance to all measurements that are smaller than a specified value. Semantic terms such as *small*, *few*, and *little* fall in this category. The following equation is used to model this type of primitive semantic term:

$$g(m) = \begin{cases} 1 & for \quad m \leqslant \lambda_R^1, \\ \frac{2}{1+e^{\left(\left(m-\lambda_R^1\right)/\lambda_R^2\right)^{\lambda_R^3}}} & for \quad m > \lambda_R^1. \end{cases} \tag{2}$$

It is defined as the union of a constant function and one sigmoid function. The sigmoid function is centered at λ_R^1, has width λ_R^2, and an exponential factor λ_R^3. The degree of significance of the right-bounded semantic term equals 1 for any measurement $m \leqslant \lambda_R^1$.

3.2.5 Bounded Primitive Semantic Term

The bounded primitive semantic terms combine the characteristics of the previously defined semantic terms. It assigns a full degree of to all measurements in a specified interval. Semantic terms such as *average*, *medium*, and *median* fall in this category. The following equation is used to model this type of primitive semantic term:

$$g(m) = \begin{cases} \frac{2}{1+e^{\left(\left(\lambda_L^1-m\right)/\lambda_L^2\right)^{\lambda_L^3}}} & for \quad m < \lambda_L^1, \\ 1 & for \quad m \in \left[\lambda_L^1, \lambda_R^1\right], \\ \frac{2}{1+e^{\left(\left(m-\lambda_R^1\right)/\lambda_R^2\right)^{\lambda_R^3}}} & for \quad m > \lambda_R^1. \end{cases} \tag{3}$$

It is defined as the union of a constant function and two sigmoid functions. The sigmoid functions are centered at λ_L^1 and λ_R^1 with widths λ_L^2 and λ_R^2 , and the exponential factors λ_L^3 and λ_R^3 . The degree of significance of the bounded semantic term equals 1 for any measurement $m \in [\lambda_L^1, \lambda_R^1]$.

3.2.6 Complex Semantic Term

Let ς_1 and ς_2 be two semantic terms and g^{s_1}, g^{s_2} be the associating possibility distributions. We define a set of logic operators for these functions $OP = \{AND, OR, NOT\}$ where $AND(s_1, s_2) = min(g^{s_1}, g^{s_2})$, $OR(s_1, s_2) = max(g^{s_1}, g^{s_2})$, and $NOT(s_1, s_2) = 1 - g^{s_1}$. A complex semantic term is defined as $\varsigma =< u, c, c_l, \delta, OP(g_1^{\varsigma}, g_2^{\varsigma}, \ldots, g_n^{\varsigma}) >$, where $OP(g_1^{\varsigma}, g_2^{\varsigma}, \ldots, g_n^{\varsigma})$ defines the rules to compose multiple primitive semantic terms or other complex terms using logic operators in OP and all other variables are defined in Section 3.1.

For example, we can construct a complex semantic term ς - *Many, above average size, with sparse coverage calcified regions* by combining the possibility distributions of two primitive semantic terms and one complex semantic term. ς_1–*Many calcified regions* and ς_2– *Sparse coverage of calcified regions*. The semantic term *Above average size calcified region* is defined by a user who wanted to find images with calcified regions that are either *big* or *average* size. Such term is not in the collection of the primitive semantic terms defined in the Semantic domain. Therefore, an intermediate complex term ς_3– *Above average size calcified regions*, is constructed by applying OR logic to another two primitive terms: ς_4 - *Average size calcified regions* and ς_5 - *Big calcified regions*. The possibility distribution for ς_3 is:

$$g^{\varsigma_3} = OR(\varsigma_4, \varsigma_5) = max(g^{\varsigma_4}, g^{\varsigma_5}) \tag{4}$$

where g^{ς_4}, g^{ς_5} are the possibility distributions for ς_4 and ς_5, respectively. Subsequently, the possibility distribution for ς is expressed by:

$$\begin{aligned} g^{\varsigma} &= AND(\varsigma_1, \varsigma_2, \varsigma_3) \\ &= AND(\varsigma_1, OR(\varsigma_4, \varsigma_5), \varsigma_3) \\ &= min(g^{\varsigma_1}, g^{\varsigma_2}, max(g^{\varsigma_4}, g^{\varsigma_5})) \end{aligned} \tag{5}$$

3.3 Mapping Low-level Features into Semantic Terms

The procedure discussed in Section 3.2 uses high-level feature that map semantics using one-to-one relations. In such approach relevant information is extracted by image analysts and it is difficult to evolve in time. In this section, we introduce an approach for knowledge discovery that uses association rules to map high-level semantics into low-level feature subspaces [119]. Our approach is composed of five modules: (1) Preprocessing and Feature Extraction: domain-specific images are processed by a suite of image processing and computer vision algorithms to obtain a high-dimensional feature vector. (2) Feature Selection: a set of unsupervised and/or supervised feature selection algorithms is applied to the feature vector resulted from the feature extraction algorithm to eliminate non-relevant features. (3) Transformation: a decision tree-based algorithm is implemented to partition a continuous space into multiple discrete ranges for individual features. (4) Data Mining: association rules [2] that map feature intervals into semantic categories are discovered. (5) Semantic modeling: the degree of significance of feature measurements to association rules is modeled using a possibility function.

3.3.1 Preprocessing and Feature Extraction

In this section, we will look into the the merits of several features that can be extracted from images. There is a wide variety of features that can be extracted from digital images. However, this task may be computationally expensive to process because it would create a high-dimensional feature vector and consequently increase the complexity of the data mining process.

When extracting features from an image, it is important to understand how humans extract meaning from images. One of the most important theories of vision is theat of constructivism [136]. According to constructivism, humans construct global models by combining local information from the surroundings with some internal mechanisms of perceptions in a series of unconscious inference processes [136]. A set of mental models–or interpretations – is constructed and, using a "heuristic interpretation process" [98], the most highly probable model is selected. According to Palmer [98], computer "vision is

extremely difficult," although the same task is accomplished by humans very quickly and accurately. Therefore, developing computer algorithms to discover visual patterns is a very complex task. Important steps were accomplished in computer vision such as edge detection, image segmentation, and 3D reconstruction. Next, we will present some algorithms for feature extraction. They are separated into color and texture methods.

3.3.1.1 Color Features

Color features are among the most important and extensively used low-level features in image database retrieval. They are usually less sensitive to noise, resolution, orientation, and resizing. Colors are represented using color spaces in terms of their intensity values. The most used color space is RGB– this represents each pixel as a triplet: (red, green, blue). The intensity of each color in this triplet takes values between 0 and 255. Another color model is HSV, which represents color using hue, saturation, and value. The hue is what is normally thought of as color, saturation is the amount of gray that is mixed into the color, and the value component is a measure of its lightness. One of the most important characteristics of human color perception is that, in normal conditions, the human eye can simultaneously perceive and conceptualize only a relatively small number of colors [98]. In addition, humans' notion of color is highly dependent on the surrounding colors, and varies with lighting conditions and colors of the display device.

The most frequent representation of color content is the color histogram. It represents the joint probability of the intensities of the three color channels and captures the global color distribution in an image. A distance measure, such as Minkowsky distance, is widely used to estimate the similarity/dissimilarity between two histograms [129]. Color histograms have several shortcomings such as: (a) its quantization of the color space is finely compared to the human perception of color, (b) it does not incorporate any spatial information and so it cannot discriminate between rapidly varying color patterns and solid colors that appear as small or large blobs around the image, and (c) it does not take into account the global color composition of the image.

Several color descriptors have been proposed with the effort to include spatial information, including the compact color moments [128], binary color sets [125], the color coherence vector [99], and the color correlogram [64]. To reduce the dimensionality of the feature space, methods such as singular value decomposition (SVD) [59] and Hilbert curve fitting [29] were applied.

3.3.1.2 Texture

Texture analysis is important in computer image analysis for classification or segmentation of images. Although there is no generally accepted formal definition, texture is a measure of smoothness, coarseness and regularity of an image region. In a more formal way, texture analysis refers to a class of mathematical procedures and models that characterize the spatial variations within imagery as a means of extracting information. Texture procedures fall into four general categories: statistical, structural, model-based methods and signal processing methods [134].

Statistical methods for texture extraction collect image signal statistics from the spatial domain as feature descriptors. First-order statistics evaluate the properties of individual pixels, while ignoring the spatial relation with other pixels. Examples of such statistical features are mean, median, skewness, standard deviation, and kurtosis. Second-order statistics evaluate the spatial inter-dependency between two pixels at specific relative positions. An example of such statistics is the co-occurrence matrices proposed by Haralick et al [61] that evaluates the spatial relationship of each pixel pair in the image. From these matrices, a number of statistical quantities can be measured, such as mean, variance, entropy, energy, contrast, and correlation. Auto-correlation is another statistical measure of texture and is used as a measure of the coarseness of textures in different directions. The discrete Fourier transform is another example of techniques that can be applied to an image. The Fourier transform represents the image in the frequency domain where each point represents a particular frequency contained in the spatial domain image.

Geometrical methods try to describe texture measures using texture elements or primitives. An example of such texture measures are the Voronoi tessellation features [133].

Voronoi tessellation has been proposed because of its desirable properties in describing local spatial neighborhoods. First, texture tokens are extracted and then the tessellation is constructed using texture primitive features [134]. Examples of primitive elements are image edges or contours.

Model-based methods construct a parametric model that describe and synthesize texture features. They are able to capture the local, contextual information in an image. These models assume that the intensity at each pixel in the image depends on the intensities of only the neighboring pixels. An example of model-based methods are Markov random field models that consider the conditional probability of the intensity of a given pixel dependent only on the intensities of the pixels in its neighborhood [85]. The Gaussian Markov random fields consider the intensity of a pixel as a linear combination of the values in its neighborhood plus a correlated noise term.

Signal processing methods analyze the frequency content of the image. Texture description with these methods is done by filtering the image with a bank of filters, using specific frequencies for each filter. Texture features are extracted from the filtered images. Spatial domain filters, such as Roberts and Laplace operators [111], are examples of such methods. Signal processing methods can also use spatial moments, which correspond to filtering the image with a set of spatial masks. Frequency analysis can be performed in the Fourier domain. Spatial dependency is incorporated into the Fourier transform by a window function. The squared magnitude of the two-dimensional version of the window function can be used in analysis of shape from texture. The wavelet transform can perform multi-resolution texture analysis by using a variable window function, whose width changes as the frequency changes. An example of wavelet transforms is the Gabor transform that uses the Gaussian function [17].

3.3.2 Feature Selection

Searching the space of features becomes computationally intensive, especially if it involves a large number of features [83]. In such cases, feature selection algorithms may be helpful in reducing the computational load while maintaining a similar level of prediction performance

and improving the understanding of underlying process feature extraction [58]. The task of selecting the relevant features can be described as an important problem in machine learning and it is subjective to the particularities of the model to be built [18]. Feature selection algorithms search for an optimal subset of the original features based on certain criteria defined to maximize the precision of the output of the model. The criteria specify the details of measuring the relevance of feature subsets as well as the relevance of each feature.

In this dissertation we evaluate various feature selection strategies and investigate their performance. There are two types of feature selection that we use: (1) unsupervised and (2) supervised.

3.3.2.1 Unsupervised Feature Selection

The goal of unsupervised feature selection is to remove the irrelevant features by testing the data in absence of semantic labels [150]. Our model maps semantics into low-level features using sets of ranges to capture the relevance of an image to semantic terms. Basd on this model, features that have all of the image measures grouped in a very small interval are irrelevant since our algorithm will not be able to define necessary feature subspaces. Also, the fact that the image measurements are grouped toward the minimum or maximum values of the feature intervals should decrease their relevance. Based on this goal, we define an unsupervised feature filter that uses the mean and standard deviation to compute feature relevance as shown in the following equation:

$$relevance = Stdev \cdot \sqrt{min(mean, |1 - mean|)} \tag{6}$$

The relevance of all features are computed, and the features with relevance less than a threshold θ are filtered out.

3.3.2.2 Supervised Feature Selection

Supervised feature selection algorithms use a specific set of labels, and a set of example objects with the appropriate labels. The goal of supervised feature selection is to generalize the relevance of each feature from the training objects and to extrapolate it to a set

of unlabeled images for classification purposes [83]. In this dissertation, we evaluate the performance of supervised feature selection using the following subset evaluation criteria included in the Weka package [51]: greedy-stepwise, best-first and genetic.

The greedy-stepwise feature selection algorithm [145] adds or removes one feature at each decision point. Within each of these steps, the search algorithm considers all possible local changes to the current subset and then selects the best one. Once a change is accepted, it is never reconsidered. The best-first algorithm [147] starts with an empty set of features and generates all possible single feature expansions. The subset with the highest evaluation is chosen and expanded in the same manner by adding single features. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues in a similar manner. Genetic feature selection algorithms [62] are an approach to computation inspired by biological evolution. Each training example is represented by a chromosome having feature measurements represented by genes. The population of chromosomes evolves according to specific rules of selection and through operators such as crossover and mutation. The fitness of each chromosome in the environment is evaluated by training and testing a classifier. Feature selection is done by selecting high-fitness individuals.

Images in the training dataset that have multiple labels are entered multiple times; that is, one time for each label. In Figure 8, image *Essence-1048*, marked with italicized characters, contains both the *Ateletasis* and *Subpleural Line* semantics. This fact is represented by two lines in the training dataset.

3.3.3 Data Transformation

After the most relevant features are selected, we evaluate the relevance of individual feature subspaces to semantic assignments. In an ideal case, a relevant feature subspace should contain data labeled with one semantic or it should contain the majority of data points from one class. This means that such a feature subspace should minimize the Shannon entropy [117]. On the other hand, a feature subspace containing data that are evenly distributed over all classes should be considered irrelevant to semantic assignments. In our approach, we are interested in finding both relevant and irrelevant feature subspaces

...

Essence-1048, adrian, Atelectasis

Essence-1048, adrian, Subpleural Line

Essence-10314, adrian, Cyst

Essence-9813, adrian, Fibrosis

Essence-1273, adrian, Honeycombing

Essence-3312, adrian, Peribronchiolar Fibrosis

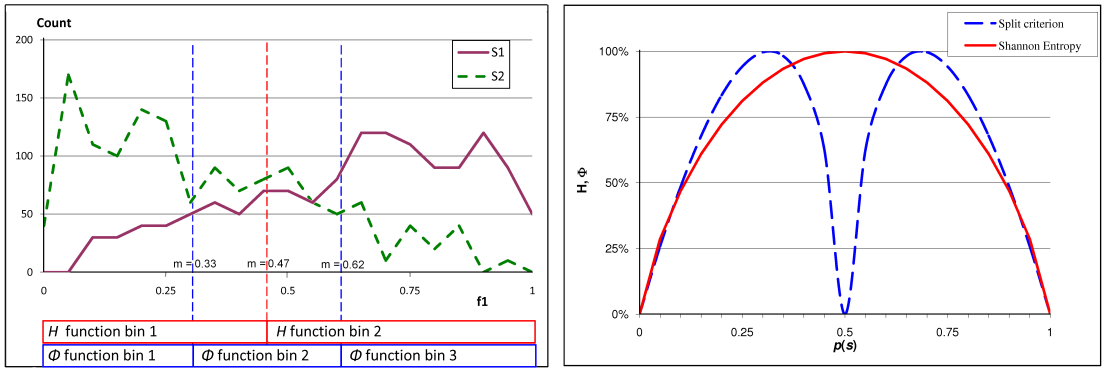
...

Figure 8: Example of an instance of the training dataset from the radiology domain. Images with multiple labels are entered multiple times. Each line contains the image code and the semantic separated by commas.

to semantic assignments. The example in Figure 9(a) shows the histogram of a two-class distribution. In this example, the majority of data points in the subspace $[0, 0.33]$ is labeled with the semantic ς_1 while the majority of data points in the interval $[0.62, 1]$ is labeled with the semantic ς_2. In the same example, the data points in the interval $(0.33, 0.62)$ are evenly distributed over the two classes. Our approach would consider the first two sub-spaces as relevant to semantic assignments and the third interval as irrelevant to semantic assignments.

Decision trees is a good way to determine feature sub-spaces relevant to labels in a training set. In a decision tree, each leaf-node represents a unique classification of a given instance, and each non-leaf node represents a measurement test [105]. Decision trees are widely used in classification problems, although they are prone to errors in problems with multiple classes and relatively small numbers of training examples [91]. It is interesting to note that decision trees are greedy algorithms that make only positive assertions about the assignment of data to a class. For example, in Figure 9(a), a decision tree that uses the Shannon entropy as split criterion will divide the data into two equal subspaces separated at 0.47. This means that such a tree will also classify the interval $(0.33, 0.62)$–that we previously considered irrelevant.

To determine relevant feature subspaces $\vartheta = \{\varphi, m_{min}, m_{max}\}$ for a feature φ, we use, for each feature, a decision tree approach based on the C4.5 algorithm [105]. This algorithm



(a) (b)

Figure 9: Data transformation. (a) Histogram of of a two class distribution. Feature subspaces discovered using the Φ and H functions. (b) We replaced the the Shannon entropy function H with a new entropy function Φ to split the tree.

uses a recursive splitting criterion to identify both relevant and irrelevant feature subspaces based on the training set for maximum information gain. Further, to avoid noise and over-fitting [104], we apply Chi-square methods to prune the decision tree to improve the efficiency without degrading the accuracy. To do this, we propose a new tree splitting function Φ. This function is shown in Equations 7 and 8 for a two-class problem in a one-dimensional space. This splitting function is different from the Shannon entropy–shown in Equation 9–in that it computes the probability p that an image $\iota_i \in I$ labeled with the semantic $\varsigma_s \in S$ are in the current bin. The value of the Φ function is small when: (a) the dataset contains only data points from one semantic ς or (b) the data points are evenly distributed over all semantics in $\varsigma \in S$. The information gain is computed using the formula shown in Equation 10. Figure 9(b) compares the Φ function with the Shannon entropy for a two-class problem in a one-dimensional space. Although the new approach tends to fragment intervals of high Shannon entropy, it does not affect the outcome of our method because such intervals–having the Shannon entropy H above a threshold θ_H–are considered irrelevant and are not included in the semantic assignments.

$$\Phi(M) = \sum_{i=1}^{|S|} \left(|1 - 2 \cdot p(\varsigma_i)| \cdot \log_{|S|} \frac{1}{|1 - 2 \cdot p(\varsigma_i)|} \right) \tag{7}$$

$$\Phi_{m_1}(M) = \frac{|M_{m<m_1}|}{|M|} \cdot \Phi\left(M_{m<m_1}\right) + \frac{|M_{m\geq m_1}|}{|M|} \cdot \Phi\left(M_{m\geq m_1}\right) \tag{8}$$

$$H(I) = \sum_{i=1}^{|S|} \left(p(\varsigma_i) \cdot \log \frac{1}{p(\varsigma_i)} \right) \tag{9}$$

$$\Delta\Phi_{m_1}(M) = H(M) - \Phi_{m_1}(M) \tag{10}$$

The result of splitting the distribution in Figure 9(a) using the Φ is shown at the bottom of the figure. It splits the data into three subspaces: $\vartheta_1 = [0, 0.33]$, $\vartheta_2 = (0.33, 0.62)$, and $\vartheta_3 = [0.62, 1]$. Notice that ϑ_1 and ϑ_3 are characterized by low Shannon entropy while ϑ_2 is characterized by high Shannon entropy. The relevance of these subspaces is then evaluated and only subspaces with $H(\iota|\varphi(\iota) \in \vartheta_i) \leq \theta_H$ will be retained for semantic assignments. The result of this step is a set of feature subspaces $V = \{\vartheta_v\}$ for all features $\varphi \in F$. If we consider each training image ι a transaction in a data mining process, the intervals ϑ determined by the data transformation step can approximate items of the transaction and can be used for discovering association rules for semantic assignments.

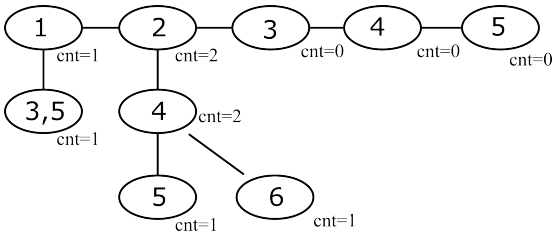


Figure 10: Example of partial-support tree. The partial-support tree is generated by reading the training data from the secondary storage.

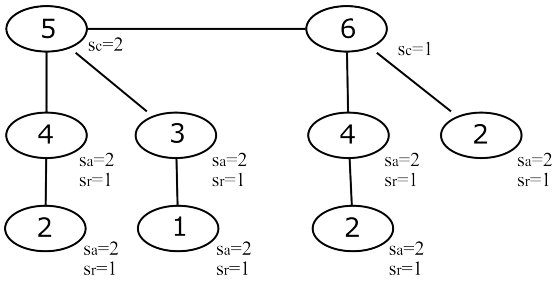


Figure 11: Example of total-support tree. The total-support tree is generated from partial support tree.

3.3.4 Mining Association Rules for Image Content

We use the set of feature subspaces V determined in the previous section to generate a complete set of decision rules with the form $r = \{a, c\}$ where $a = \{\vartheta \in V\}$ and $c = \varsigma$. Each rule has a set of feature subspaces V_r as antecedent and a semantic ς as consequent. For generating association rules we use the Total-From-Partial approach as described by

Coenen et al [35]. The advantages of this approach is that the training dataset is read only once from the secondary storage to build a partial-support tree. The example in Figures 10 and 11 show an example generating the partial-support and total-support trees. In this example we have used the following transactions: $\{1,3,5\},\{2,4,5\}$, and $\{2,4,6\}$ with 5 and 6 considered semantic classes. The partial-support tree is a data structure that holds a compressed copy of the training data. Its main advantages are that it merges duplicate records and allows the effective storage of partial support of item sets. For example, in Figure 10, by traversing the partiaaal-support tree we determine that the support of item set $\{2,4\}$ is two transactions or 66%. The partial-support tree is used to generate a total-support tree. The total-support tree is a "reverse" order prefix tree. The example in Figure 11 shows the corresponding total-support tree. In this tree, each node holds the following information: (1) support of the item set and (2) support of the antecedent of the item set (item set without the semantic class). Note that, to reduce the size, we have modified the total-support tree structure. In our implementation, at root level we keep only the nodes representing semantic classes. To find frequent item sets we scan the total-support tree starting from the second level. For example, in Figure 11 the item set $\{4,5\}$ has the support one transactions, the support of the antecedent (4) equal to two transactions, and the support of the consequent (5) two transactions. For more detailed information on partial-support and total-support trees, the reader is referred to [35].

Once generated, the association rules are evaluated for relevance to the semantic model using a nonparametric Wilcoxon signed-rank test [144]. The Wilcoxon signed rank test compares the median differences in paired data by sorting the absolute values of the differences from smallest to largest, assigning ranks to the absolute values and then finding the sum of the ranks of the positive differences. In our case the first set of observations is the initial relevance of the semantic model to the training dataset, while the second set of observations is the relevance of the semantic model to the training dataset after adding the newly discovered association rule. The newly discovered rule is added to the model only if the z-value calculated by the Wilcoxon test is greater than a threshold (in our experiments we have used 5% level of significance). To avoid generation sub-optimal models, we use a

Sequential Forward Floating Selection Algorithm (SFFS) [102]. This algorithm reevaluates previously added association rules and removes them if the Wilcoxon test returns a better z-value. The relevance of the semantic model to each transaction in the training set is calculated using Equation 11 for classification and Equation 12 for ranking. In these formulas, T_ι^ς is the relevance of image ι to the semantic ς and will be discussed in detail in section 4.1, $rank\,(T_\iota^\varsigma)$ is the rank of T_ι^ς over all the relevances, and w_T is a weight that is used to help include rules that increase T_ι^ς without changing the ranking order. The complexity of this algorith is $O(n \cdot r)$ where n is the number of training samples and r is the number of association rules evaluated with the Wilcoxon test.

$$\text{Relevance}_{\text{Classification}}(R_\varsigma, \iota) = \begin{cases} T_\iota^\varsigma - \underset{\varsigma_1 \neq \varsigma}{Average}(T_\iota^{\varsigma_1}) & \text{if correctly classified} \\ T_\iota^\varsigma - \underset{\varsigma_1 \neq \varsigma}{Max}(T_\iota^{\varsigma_1}) & \text{else} \end{cases} \tag{11}$$

$$\text{Relevance}_{\text{Rank}}(R_\varsigma, \iota) = \begin{cases} \frac{rank(T_\iota^\varsigma)}{|I|} + w_T T_\iota^\varsigma & \text{if } \iota.\varsigma{=}\varsigma \\ 1 - \frac{ranking(T_\iota^\varsigma)}{|I|} - w_T T_\iota^\varsigma & \text{else} \end{cases} \tag{12}$$

The selected decision rules are then used to model the semantic assignment of features to semantics. Figure 12 shows an example of the mapping of semantics to low-level features using association rules. 12(a) shows a partial example of the maize ontology. The semantic $\varsigma =$ "*Mild infection of leaf*" is mapped into the low-level feature space $F = \{\varphi_1, \varphi_2\}$ using two rules: $r_1 = \{\{\{\varphi_1, 0.5, 0.8\}, \{\varphi_2, 0.45, 0.8\}\}, \varsigma\}$ and $r_2 = \{\{\{\varphi_2, 0.15, 0.2\}\}, \varsigma\}$. The rule r_1 has a confidence $\chi = 70\%$ and a support $\sigma = 16\%$ while r_2 has a confidence of $\chi = 65\%$ and a support $\sigma = 16\%$.

The set $R = \{r | r_c = \varsigma\}$ of frequently co-occurring items from the selected rules will be used to model every visual semantic ς. Each item represents an interval of a specific feature, and is mathematically modeled by a flexible possibility function. The semantic assignments of the feature vectors are then stored in a shared taxonomical structure for semantic queries and retrievals.

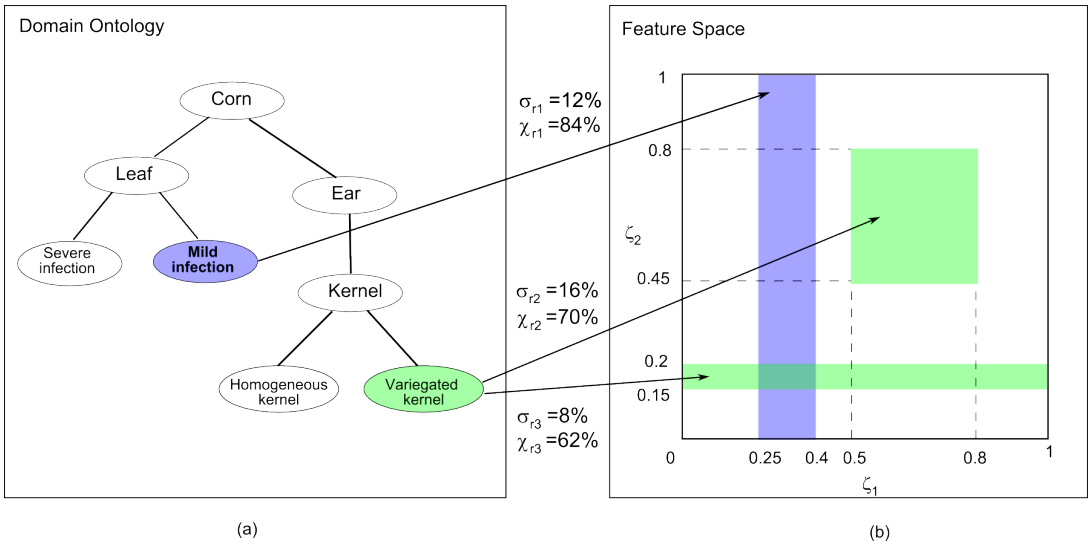


Figure 12: Example of mapping of domain semantics into a two dimensional feature space. Semantics are selected from the domain ontology and then mapped into the low-level feature space using association rules.

3.3.5 Semantic Modeling

After the association rules are discovered, we replace the crisp intervals in antecedents of all of the rules with possibility distributions as described in Section 3.2. This approach has the advantage of capturing users' preferences in a computational way using an asymmetric property modeled by two halves of sigmoid functions.

This possibility distribution is shaped using the information in the training dataset. First, we compute the normalized histogram of the distribution of data labeled with the target semantic ς over the feature interval φ. Using this information, the possibility function is computed using the algorithm described in [12] by first computing a non-parametric function, and then computing the parametric possibility function using a nonlinear least square fitting algorithm. Figure 13 shows an example of shaping a possibility function. The thin continuous line represents the initial subspace as determined by the data mining process. The normalized distribution histogram is displayed as a thick continuous curve, while the possibility function is displayed using a thick curve with square markers. In this example, the parameters of the possibility function are: $\lambda_1^L = 0.10$, $\lambda_1^R = 0.14$, $\lambda_2^L = 0.035$, $\lambda_2^R = 0.026$, $\lambda_3^L = 2$, and $\lambda_3^R = 1.0$. This function is then used to determine the degree

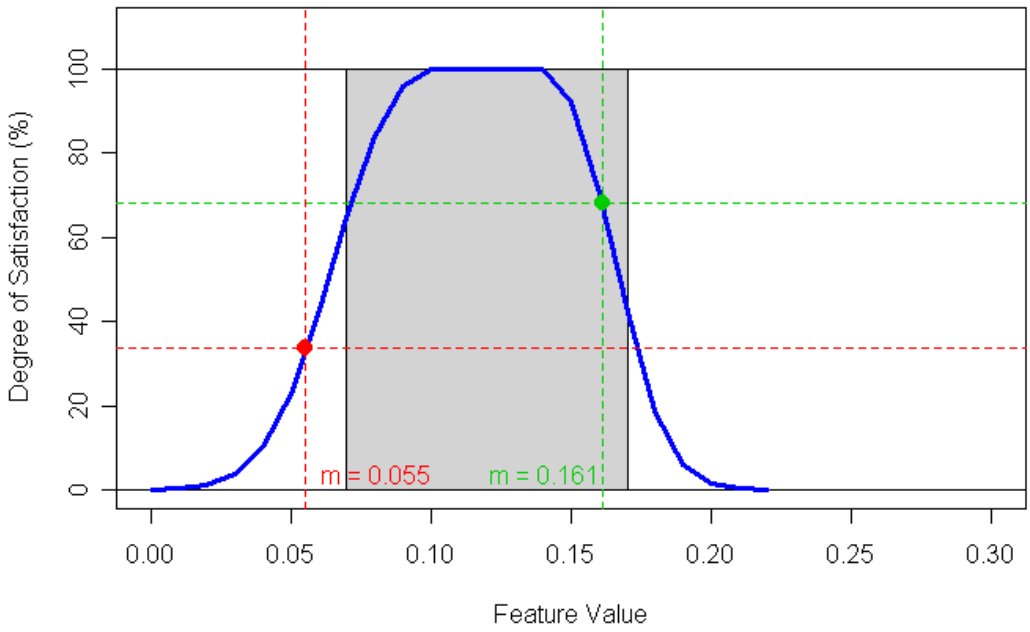


Figure 13: Example of semantic modeling. We replace each crisp feature subspace ϑ in the antecedents of a rule with a flexible parametric function.

of satisfaction of feature measurements in an association rule. For example, for a feature measure $m = 0.161$, the function determined in Figure 13 will return a degree of satisfaction of 0.68 or 68%, while for a feature measure $m = 0.055$, the degree of satisfaction of 0.336 (33.6%) will be returned. For the same measurements, using the crisp interval, the degree of satisfaction will be 1 and 0 respectively.

CHAPTER IV

QUERY METHODS USING KNOWLEDGE MODELS

The main type of interaction between experts and the *Essence* system is query. There are two types of queries provided to users: (1) query by semantics and (2) query by example. Query by semantics can be used to evaluate a new set of images that were added to the system or for training purposes. In this setting the user inputs a set of semantics to the system and retrieves a ranked set of images that are relevant to the input semantics. When querying by example, the user inputs a new image to the system and retrieves a set of similar images. For each of the retrieved images, the ranked set of system semantic relevance is also provided. In the next sections we will discuss these query methods in depth.

4.1 Query by Semantics

The main tasks performed by the semantic query are: (1) processing semantic query constraints from the user's input, (2) searching image databases by semantics, and (3) accumulating the query history for updating the user's preferences. For a given query, such as "retrieve lung images with *big cysts*," the semantic query first finds the semantic term *big cysts* from the semantic profile tree, as shown in Figure 14, and then forms a possibility distribution for this semantic term on-the-fly. The system ranks the qualified images based on the descending order of the degree of significance by substituting the measurement, in this example the size of cysts, into the possibility distribution function. These three tasks are implemented using the pseudocode shown in Figure 15.

4.1.1 Selecting Semantic Terms

To select a set of semantic terms, users access their working semantic profiles and select linguistic variables–*size of cysts* with semantic terms–ς = big, as shown in lines 05 to 13 of the Query_by_semantic function. In our system a query constraint is defined as the set of rules R_{ς} having the consequent $c = \varsigma$. The user query $Q_u = \{R_{\varsigma_1}, R_{\varsigma_2}, \ldots, R_{\varsigma_r}\}$ is formed

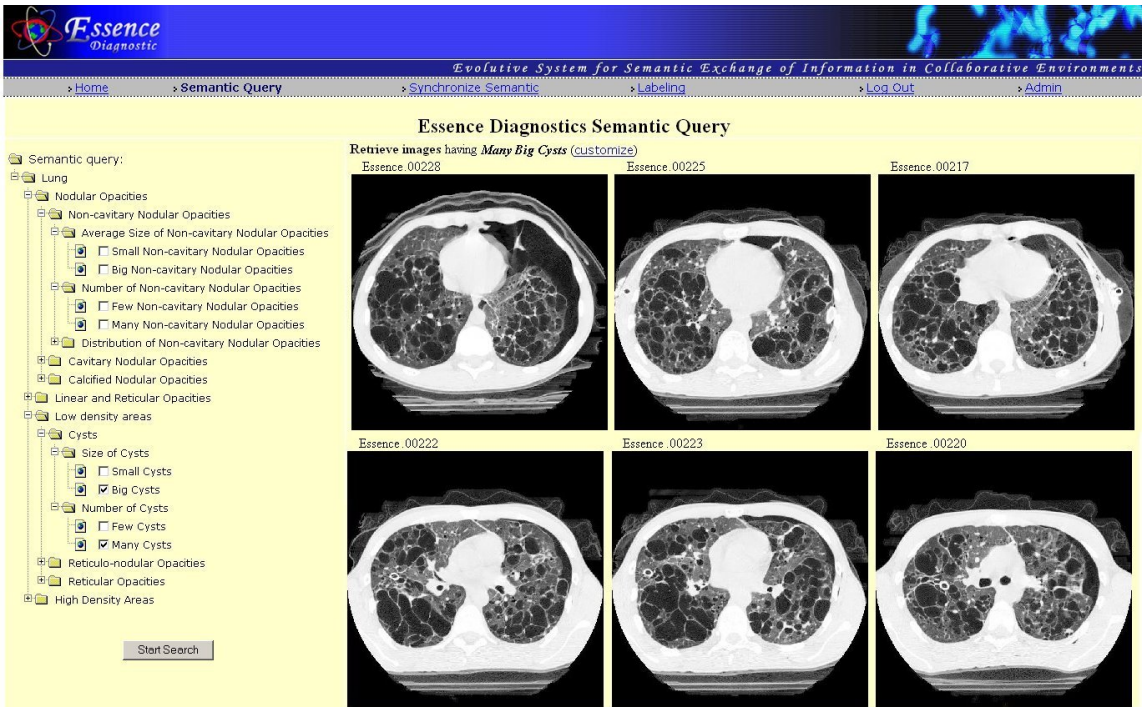


Figure 14: Set of images retrieved upon querying for the *average size of cysts* linguistic variable and *Big* semantic term.

by the set of all r querying constraints defined by the user in the semantic query.

4.1.2 Image Relevance to Semantics

The proposed query method searches the image databases by semantics using the association rules described previously. For a given query, such as "Retrieve images of lung with big cysts," the system first finds the set of decision rules $R_\varsigma = \{r | c_r = \varsigma\}$ that apply to the semantic $\varsigma =$ "*cyst*," and then determines the set of association rules that are relevant to each image $\iota \in I$: $R_{\varsigma,\iota} = \{r | c_r = \varsigma, T_\iota^\varsigma > 0\}$ where T_ι^ς is the relevance of image ι to rule r, given by the following equation:

$$T_\iota^{\varsigma,r} = \min_{all\ A} \left(g_\varsigma^A(m) \right) \cdot \frac{\chi(r) \cdot \sigma(r)}{(1 - \tau_\sigma) \cdot \chi(r) + \tau_\sigma \cdot \sigma(r)} \tag{13}$$

In Equation 13, χ is the rule confidence, σ is the rule support and τ_σ is a weighting factor for support. In our experiments we have used $\tau_\sigma = 0.995$ that was empirically shown to result in best accuracy.

```
01 Query_by_semantic(user)
02   QCS ← "rootSemantic"; ▷ set of query constraints
03   RIM ← array[1, 1]; ▷ ordered set of retrieved images
04   count ← 1; number of linguistic variables in QCS
05   while count > 0
06     count ← 0;
07     display the linguistic variables child terms
08     let user choose the desired terms
09     QC ←  set of terms selected by users
10     for i ← 1 to length[QC]
11       if QC[i] is linguistic variable
12         count = count + 1;
13         break;
14   for i ← 1 to length[image database]
15     a[i] = 1 ▷ overall degree of significance for image
16     AS = array[1, length[QCS]▷ set of degrees of significance
17     for j ← 1 to length[QCS]
18       compute AS[j]   ▷  the degree of significance image relative to j
19       if AS[j] < a[i] then
20         a[i] = AS[j]
21       if length[QCS] < QCS0 or a[i] > a[length[QCS]]
22         length[RIM] = length[RIM] + 1
23         add image to RIM
24   display RIM to the user
25   collect and save user preferences
26   return;
```

Figure 15: Semantic query pseudo code

To determine the rank of each image we use an approach derived from the Kruskal-Wallis statistical test [75]. First, we sort all $T_\iota^{\varsigma,r} \in R_\varsigma$. Let $rank(T_\iota^{\varsigma,r})$ be the rank of the relevance $T_\iota^{\varsigma,r}$ of rule r to the image ι. The final rank of image ι is given by the following equation:

$$T_\iota^\varsigma = \left(\frac{\sum\limits_{r \in R_{\varsigma,\iota}} rank\left(T_\iota^{\varsigma,r}\right)}{|R_{\varsigma,\iota}|} - \frac{|R_{\varsigma,\iota}| \cdot \sum\limits_{r \in R_\varsigma} rank\left(T_\iota^\varsigma\right)}{|R_\varsigma|} \right) \tag{14}$$

The system then ranks images that have the highest T_ι^ς values for the semantic of interest and displays them to the users. Figure 15 shows an example of the interface of the query by semantic system for the medical domain. Images in this example were retrieved upon querying for the *average size of cysts* linguistic variable and *big* semantic term and are ranked from the most relevant to the least relevant. Figure 16 shows an example of interface for searching visual patterns of maize leaves. Images were retrieve upon querying for *large brown lesions of maize leaf*.

4.2 Query by Example

4.2.1 Semantic Relevance to Images

If a new image is presented to the system we can use the generated association rules to classify it. For a given image ι with an unknown semantic assignment, the system first finds the set of decision rules $R_\varsigma = \{r | T_\iota^{\varsigma,r} > 0\}$ that are relevant to the image ι.

Then it determines the set of association rules that are relevant to each semantic $\varsigma \in S$: $R_{\iota,\varsigma} = \{r | c_r = \varsigma, T_\iota^\varsigma > 0\}$ where T_ι^ς is the relevance of image given by Equation 13: To determine the relevance of each semantic to the image ι we use an approach derived from the Kruskal-Wallis statistical test [75]. First, we sort all $T_\iota^\varsigma \in R_\varsigma$. Let $rank(T_\iota^{\varsigma,r})$ be the rank of the relevance $T_\iota^{\varsigma,r}$ of rule r to the image ι. The final rank of image ι is given by the following formula:

$$T_\iota^\varsigma = \left(\frac{\sum\limits_{r \in R_{\iota,\iota}} rank\left(T_\iota^{\varsigma,r}\right)}{|R_{\varsigma,\iota}|} - \frac{|R_{\iota,\varsigma}| \cdot \sum\limits_{r \in R_\iota} rank\left(T_\iota^\varsigma\right)}{|R_\iota|} \right) \tag{15}$$

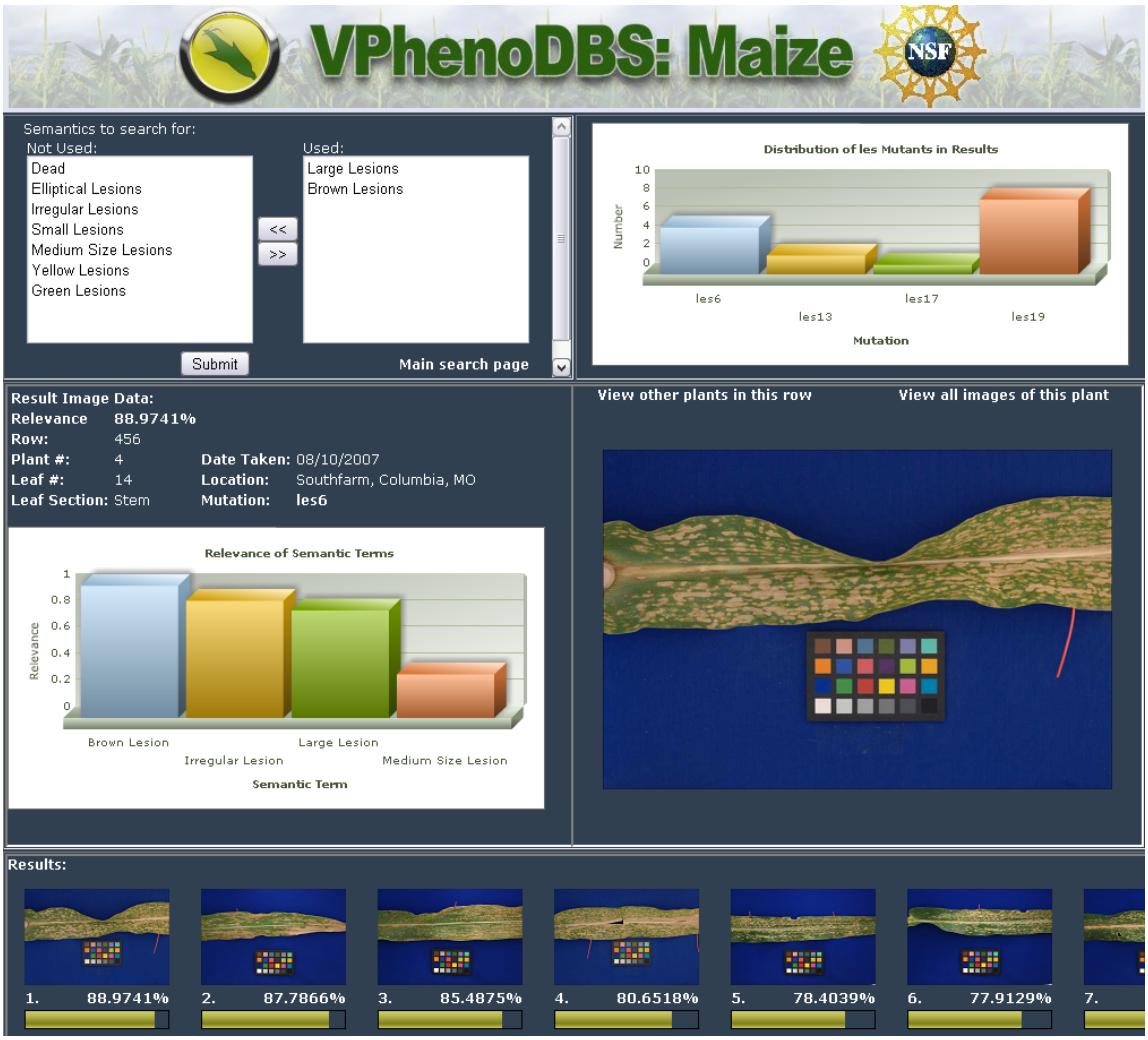


Figure 16: Set of images retrieved upon querying for *large lesions of maize leaf* and *brown lesions of maize leaf* semantics.

The system then selects the semantic ς that has the highest T_ι^ς values for the image of interest and displays them to the users. The complexity of ranking is $O(n \cdot log(n) \cdot r)$ where n is the number of images to rank, and r is the number of rules in a semantic model.

4.3 Improving the Retrieval Time

Space-filling curves are bijective functions that can traverse every point in a multi-dimensional space. Many types of space-filling curves were developed over the last century such as Peano, Hilbert, and Gray [112]. The order of point traversal can be used as a unique mapping of a high-dimensional space into a one-dimensional space. Using this mapping, an n-dimensional vector can be described by the length along the space-filling curve, measured from an initial point. The example in Figure 17(a) approximates the two-dimensional data space using a second order Hilbert space-filling curve ($c = 2$). In this example, both features f_1 and f_2 are divided into $2^c = 4$ subspaces for a total of 16 hypercubes. These hypercubes are traversed by the Hilbert curve in the order shown and noted with numbers between 1 and 16. All of the data points that reside in a hypercube can be approximated to this integer value of the point on the curve. For example, the point shown in Figure 17(a) that has measurement ($f_1 = 0.15, f_2 = 0.65$) can be approximated index as 12 on the Hilbert curve, which is used as an indexing key in a B^+-tree for fast retrieval.

The clustering properties of the space-filling curves makes them suitable for data indexing because they can reduce the number of disk accesses. The work in [47] proposed the use of the Hilbert curve, in order to design good distance-preserving mappings for multi-dimensional data indexing. Dafner et al. [37] used customized space-filling curves for representing pixel information of a predetermined group of images. Space-filling curves were also used in indexing structures by Lawder et al. [76]. They proposed a compact tree representation of the the n-dimensional Hilbert curve that is used to map data points into the corresponding Hilbert sequence number.

There are several drawbacks when using space-filling curves for multi-dimensional data

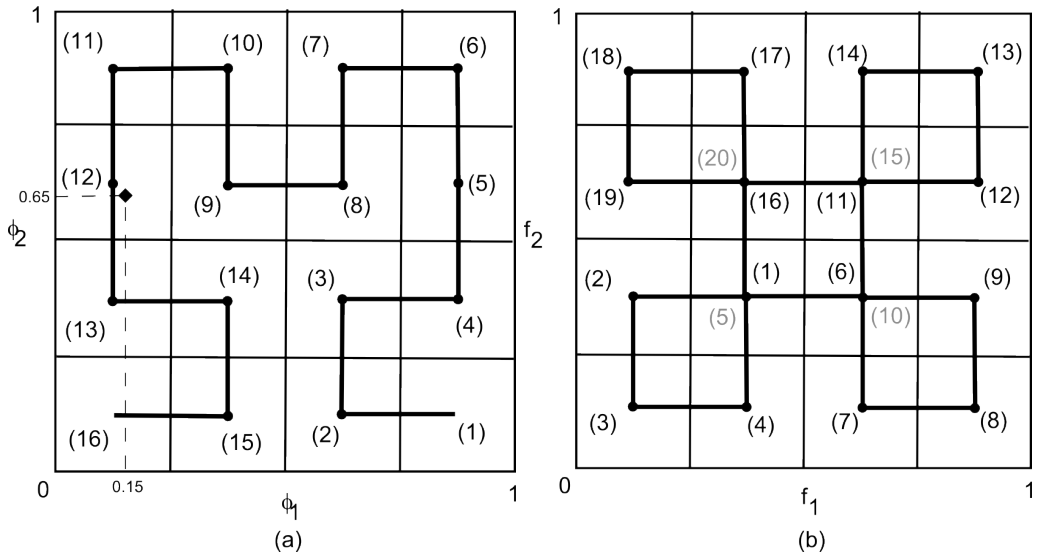


Figure 17: Example of data approximation using space-filling curves. (a) Hilbert curve. Each feature was divided into four equal subspaces. The Hilbert curve, denoted by bold lines, traverses each resulting hyper-cubical region in a predefined sequence. (b) Gray-code space-filling curve. The sequence of point on the curve is determined using a Hamiltonian path.

indexing. First of all, the difficulty in constructing such curves increases when the dimensionality increases. Most of the studies on space-filling curves are done in two or three-dimensional spaces [5]. Although the work in [4] generalizes the Hilbert curve to an arbitrary number of dimensions, we have chosen to use a customized variant of gray-code sequence that can be better fitted to customization. Another issue of using space-filling curves is that the number of indexing keys on the space-filling curve grows exponentially with increased size of the feature space. In such cases, the efficiency is reduced by searching many empty mapping points. For example, in a 25-dimensional space with a first order space-filling curve, 2^{25} indexing keys will be created. If one million or 2^{20} data points are indexed then statistically 99.7% of the indexing keys will contain no associating data point.

4.3.1 Proposed Approach

For the purpose of our research, we have chosen a variant of gray-code space-filling curves. This type of curve visits points of a multi-dimensional space in the gray-code order [53]. The gray-code algorithm encodes sequences of binary numbers so that any adjacent ones have only one bit difference. An example of such a gray code sequence for a 3-dimensional

space is the set (000, 001, 011, 010, 110, 111, 101, 100). Using gray-codes we can define a Hamiltonian path that cycles through all of the vertexes of the hypercube. Figure 17(b) shows such a space-filling curve of order two over a two-dimensional space. Our sequence starts from a point that is close to the center of the feature space and traverses it using a gray-code algorithm. For example, in Figure 17(b), the coordinates of indexing keys (1) and (5) are the same and one might say it is logical to directly connect the point noted (4) to (6). While this is true, such connection becomes difficult to link when the number of dimensions increases. In our implementation, although the two points exist on the sequence, the higher order one, shown with light color in Figure 17(b), is never used.

To avoid some of the drawbacks of using the space-filling curve approach, our method extends the traditional approach to be adaptive to the data set. The example in Figures 18(a) and (c) show some of these drawbacks. The data in Figure 18(a) has a chi-square distribution over both f_1 and f_2 features. In this case most of the data points will be approximated by the indexing key (3). Our approach, shown in Figure 18(b), creates a uniform distribution of data points to indexing keys by using flexible feature sub-spaces. Another case is shown in Figure 18(c) where the data has a uniform distribution over feature f_1 and a normal distribution over f_2. In such a case the indexing keys toward the limits of the feature f_2 will index few or no data points. Our solution, shown in Figure 18(d), is to assign a higher order space-filling curve for a feature that has high uniformness in data distribution. In this example we have a space-filling curve of third order for feature f_1 and first order for feature f_2. The advantages of such an approach is that although it uses the same number of indexing keys, it partitions the data points based on the characteristics of their distribution.

4.3.2 Index Creation

Let f_k be the k^{th} feature, n_m be the number of multimedia objects in the database, n_t be the expected number of multimedia objects for each indexing key in the space-filling curve, and $n_p = \left\lceil log_2^{\frac{n_m}{n_t}} \right\rceil$ be the total number of partitions across all feature dimensions. The following sub-sections discuss the three steps for creating an indexing structure using our

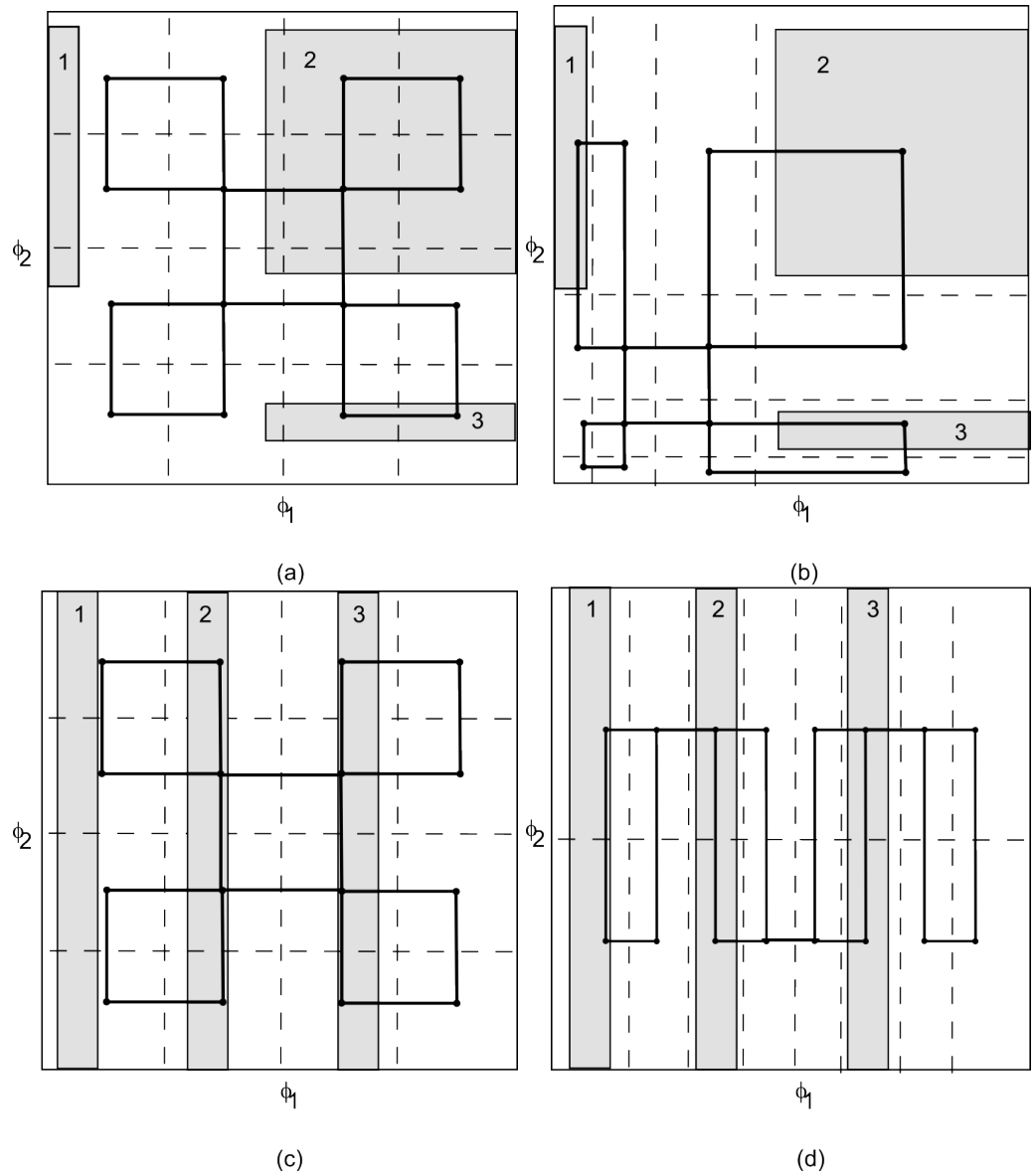


Figure 18: Example of space-filling curve customization to fit the existent data. The example in (a) shows a skewed distribution of data points per indexing key. The approach in (b) creates a more uniform distribution by varying the size of sub-spaces. Similar to (a) is the example in (c). The solution in (d) is to further split features that have a more uniform distribution of data.

customized space-filling sequences.

4.3.2.1 Creating a space-filling sequence

The space-filling curve is created iteratively. First we evaluate the uniformness of all of the features using an Anderson-Darling statistical test. This test quantifies the degree of uniformness of each feature. For example, a feature dimension which has a highly uniform distribution of the dataset will have more partitions. Each feature f_k is ranked by its uniformness using a ranking function. The order (c) of the space-filling sequence is then determined based on the ranks. Next, we compute the quantiles for each dimension to create a uniform distribution of data points across the indexing keys. The number of quantiles for each feature is determined by the order c. Using this information we recursively create the space-filling sequence.

4.3.2.2 Assigning an indexing key for each data point

To index a data point in a B^+-tree, we first compute its approximation on the space-filling sequence. This is done iteratively, using a similar approach to the one described in the previous sub-section. At each iteration, the data point is approximated to the closest indexing key of that order. The iteration stops when the maximum order of the sequence is reached.

4.3.2.3 Creating a B^+-tree index

Once we have all data points in a high-dimensional space mapped into a scalar from the space-filling sequence, we build a one-dimensional indexing structure using a B^+-tree. Each leaf node in this tree contains all data points that share the same indexing key.

4.3.3 Range Queries

As we mentioned previously, our approach is specialized to improve the efficiency of range queries using only a subset of features F_r in an n-dimensional space. The query region may overlap multiple sections of the space-filling sequence. For example, in Figure 19(a), the query rectangle overlaps with two sections of the space-filling sequence: one formed by the indexing keys (3) and (4) and the other by the indexing key (7). Our query will traverse all

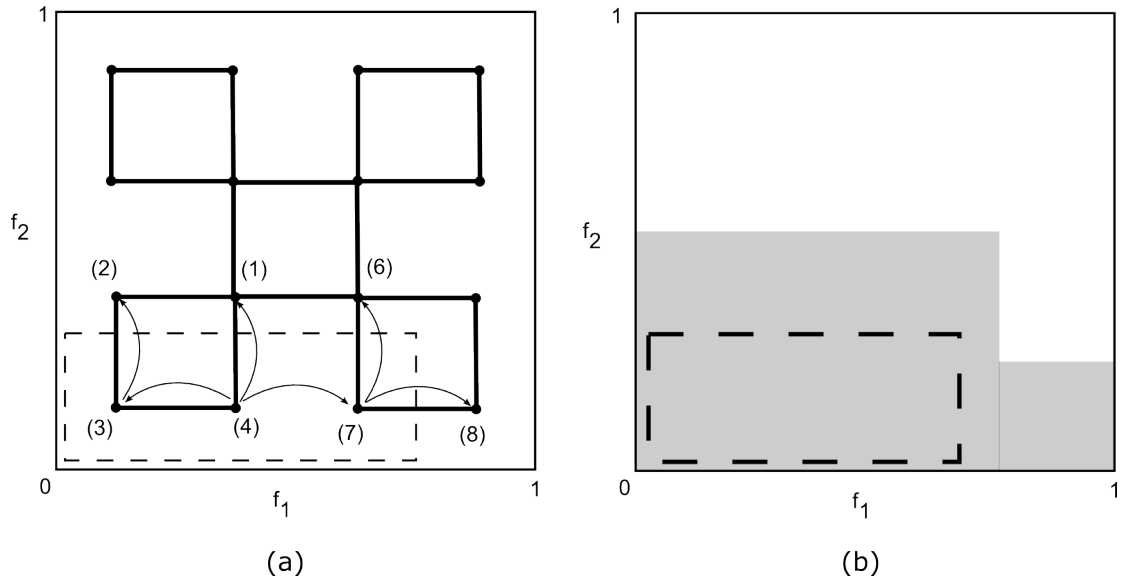


Figure 19: Range search query example. In this figure the query range is denoted by a dotted line. (a) The query starts by traversing the sub-sequence that includes the center of query range and then moving to the neighboring sub-sequences in an iterative manner. The iteration stops when all the indexing keys in the query range are exhausted or if the size of the result is reached. (b) For an exhaustive search of all of the data relevant to the query range, our method will search all of the data points in the grayed area.

space-filling regions that overlap with the query region. The query starts with the center of the query region and determines its indexing key. Then, the algorithm traverses the leaf nodes of the B^+-tree in both directions, until it reaches an indexing key that is outside the query range. For example, in Figure 19(a), the query starts at indexing key (4), which is the point closest to the center of the query region. Then, by traversing the tree in a descending direction, it also accesses the neighboring indexing keys (1), (2), and (3). Since keys (1) and (2) are outside of the query range, the query moves to another indexing region by searching other unvisited keys that are neighbors to those that have been visited in the F_r feature space. In this example, the algorithm traverses in an ascending order and discovers key (5). It then traverses the sub-sequence that contains key (5) to visit keys (6) and (8).

CHAPTER V

KNOWLEDGE EXCHANGE

In this section, we describe methods for knowledge exchange in the Essence Framework. The user has a wide variety of choices to exchange knowledge that is relevant to their mental models. These choices are: (1) Semantic customization, (2) System-level information exchange, and (3) Peer-to-peer information exchange. If the result of the semantic query is unsatisfactory for the user, they can customize the semantic assignment to an image feature. Upon completing this process, the results of the semantic query are expected to better reflect the user's subjective preferences. Also, the user can evaluate the query result using the semantic setting of other users of the system by using system-level information exchange. Finally, the peer-to-peer knowledge exchange will provide the user with similar cases from users' in different organizations.

5.1 User-specific Semantic Customization for Modeling Domain Knowledge

There are several reasons that a semantic retrieval could lead to an unacceptable result. In image retrieval, the process of articulating perceptual categories, as well as quantifying the associated semantic terms, proves to be highly subjective. Therefore, a robust semantic search engine should allow users to modify the quantification of existing semantic terms and to add new ones if needed [10]. Upon reviewing the retrieval results of query Q_u, if the user u decides that the results are not satisfactory, they can either modify the possibility distribution of each semantic term in Q_u, or add a new semantic term to the linguistic variable, with the help of the system's web-based interface.

5.1.1 Customization Procedure for High-level Feature Spaces

The flow of events for customizing the possibility distribution of a semantic term ς is as follows: 1) the system displays s training images having the measurement evenly distributed

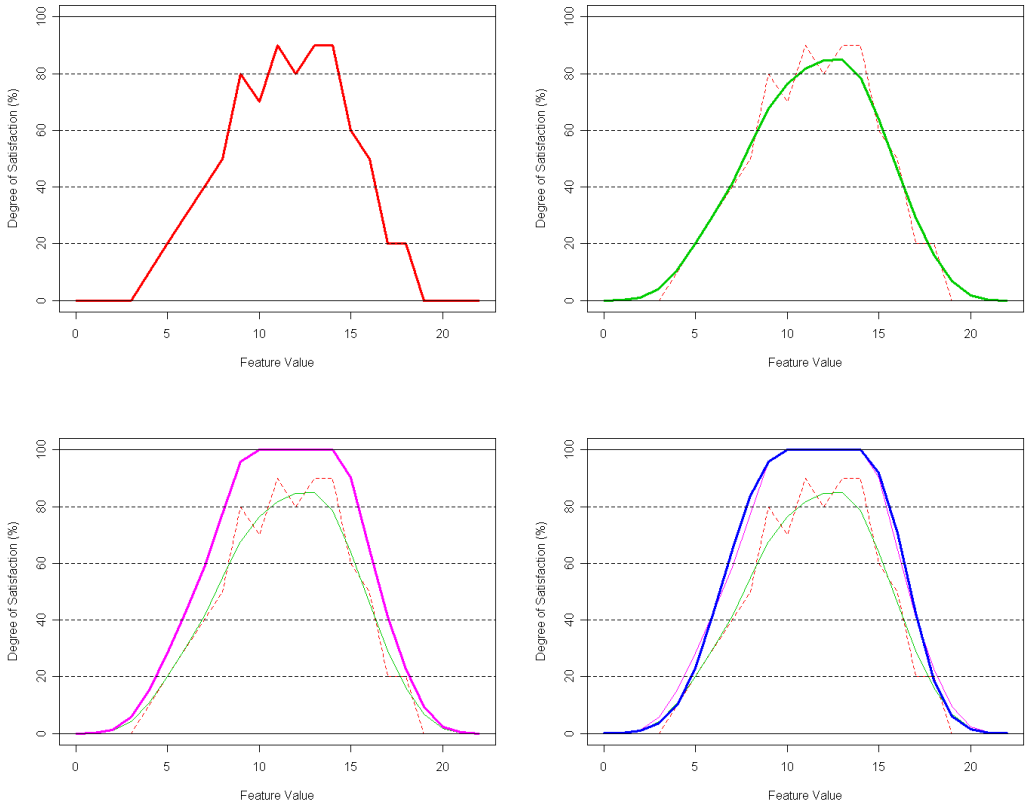


Figure 20: Determining the customized user-specific membership function. User rating is shown in (a). A kernel regression is applied to the user's input (b) and then it is compensated (c). The final sigmoid function is computed by nonlinear least square fitting algorithm and it is shown in (d).

over the universe of the linguistic variable l, and 2) the user rates the displayed images on a scale from 0 to 10. If the user's selection is not informative enough (few rated images with low ratings), the system will repeat the similar process based on the high rated images selected in the previous iteration.

Determining the possibility distribution that best matches the user's preferences (Figure 20) could be achieved by ensuring: (1) distribution completeness, (2) user preferences compensation, and (3) distribution regression. Let m be a measurement associated with an image feature and ς be the semantic term to be refined. As mentioned previously, the function will assign full degrees of significance for at least one measurement m in the universe of discourse. The system ensures the completeness by computing a function $g(m) = b = max(mean(g^{\varsigma}(m)))$, where m are measurements for any number of consecutive measurements in the training set. The rating of each image is then adjusted using $g^{\varsigma}(m) = min(10, g^{\varsigma}(m))$.

The sigmoid functions that best match the adjusted user's preferences are computed using a nonlinear least square fitting algorithm, and then the parameters λ_L^1, λ_L^2 , λ_L^3 , λ_R^1, λ_R^2 and λ_R^3 are decided. This setting is saved in both the user-specific and candidate semantic profiles, while the user selections are saved in the user preferences knowledge base.

5.1.2 Customization Procedure for Low-level Feature Spaces

To achieve better individualized settings, we developed an approach that uses individualized customizations of semantic into low-level features. This approach addresses the semantic gap that exists among analysts. This customization is available for analysts who provide feedback to the system by labeling retrieved images as either positive or negative examples. It provides three ways to quickly customize semantic assignments: (1) altering the existing rules to accommodate new user input, (2) removing existing rules that become irrelevant, and (3) adding negative rules to the semantic user profile. An image analyst can provide feedback on a continuous scale with values between -1 and 1, which correspond to a *negative example* and a *positive example*, respectively. These ratings are used for adjusting the semantic model that is specific to the image analyst without impacting other semantic

```
1   CUSTOMIZE(SM_u, I)
2    FOR EACH r ∈ SM_u  DO
3     Compute new support and confidence
4     IF σ_a < θ_σ OR χ_r < θ_χ THEN
5      Remove r from SM_u
6     END IF
7     FOR EACH g_φ ∈ a_r DO
8       Apply kernel regression to the feature υ_r ⊂ φ, φ ∈ F_I
9       Determine the new relevant subspaces V = {υ}
10      FOR EACH υ ∈ V DO
11        Compute new g_φ
12        Add g_φ to a_r
13      END FOR
14    END FOR
15   END FOR
16  RETURN SM_u
```

Figure 21: Algorithm for customizing the semantic profile SM_u of user u. Sigmoid functions that map the feature space into semantics are customized to the input of the user.

models.

The pseudocode for rule customization is shown in Figure 21. The customization function takes as input the semantic model, SM_u, of a user, u, and a set, I, of images that were labeled by the user. The customization process first modifies existing rules r by changing the level of support, σ_r, and confidence, χ_r, to account for the newly rated images (lines 4 to 6 in the pseudocode). For example, if the user adds a number of negative examples to the system, the confidence and support values of all the rules that have antecedents relevant to those images should be decreased. In doing so, the relevance of the whole rule is decreased. When the confidence and/or support values fall under predefined thresholds θ_σ and θ_χ, respectively, the rule is removed from the semantic model.

A second type of user-specific customization of the semantic model is performed by recomputing the sigmoid possibility function in the antecedents, a_r, of the semantic model. This process is shown from lines 7 to 14 of the pseudocode. First, we apply a kernel

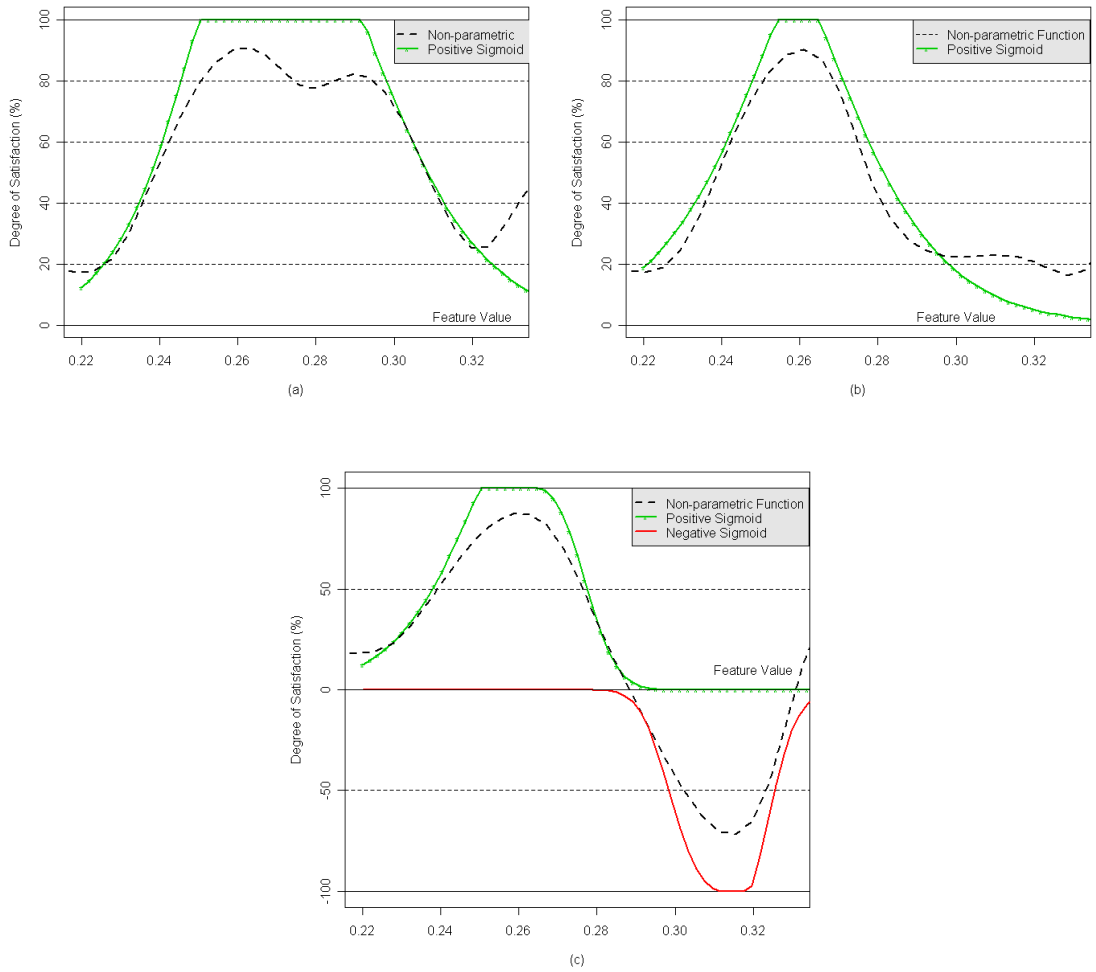


Figure 22: The process of semantic customization after a user provided a set of positive and negative examples for the semantic *construction*. (a) initial distribution of data and the sigmoid approximation. (b) the sigmoid approximation is adjusted to the new input from the user. (c) the negative examples provided by an image analyst will adjust the initial sigmoid function and then a new negative antecedent is created.

regression to the image ratings provided as feedback. Kernel regression is a method used for non-parametric smoothing of a variable. Figure 22(b) shows, with dotted lines, the result of the kernel regression using a normal kernel with a bandwidth of 0.04. The kernel regression was applied to the *construction* semantic and *spec_v_bin11* feature after the user provided a set of both positive and negative example images. As seen in the figure, the initial feature space is now divided into two subspaces at a value of approximately 0.29. In the left subspace, the number of positive *construction* examples exceed the negative ones and is consistent with the initial distribution shown in Figure 22(a). In the right subspace, the trend was reversed by the user's input. In this subspace there are more negative examples than positive ones. According to the new information, the two subspaces will be modeled separately and the negative antecedent will be added to the association rule. The new sigmoid possibility function is shown with a solid continuous line. For example if an image has the *spec_v_bin11* feature value of 0.3, the the negative antecedent will de-emphasize its relevance to the semantic *construction* by 0.17.

5.2 System-level Information Organization and Exchange

Periodically the system automatically triggers a learning component that updates possibility distributions for the default profile. To do that, the system searches user preferences for the most recently updated distributions that are highly correlated to the default profile (correlation greater than 0.7). It then computes two weights: $w_{un} = lg_2(n_{un})/n_{un}(lg_2(n_{un}) + lg_2(n_{ue}))$ for the qualified users' new ratings and $w_{default} = lg_2(n_{ue})/(lg_2(n_{un}) + lg_2(n_{ue}))$ for the default possibility distribution. In these ratios, n_{ue} is the number of users that have already contributed to the default profile and n_{un} is the number of qualified users that will contribute to it. This approach progressively increases $w_{default}$ to ensure the stability of the default profile. On the other hand, this system should be able to keep accepting new inputs from users even with a large number of users who previously contributed to the default profile. To deal with this, the logarithm function works by limiting the influence of $w_{default}$ when n_{ue} is large. After both updated weights are computed, the system builds a new non-parametric possibility distribution by taking a weighted average from the default possibility

distribution and the ratings from all qualified users. This is to adjust the default possibility distribution. An algorithm similar to the one described previously for user-specific profiles is then applied to form a new parametric default possibility distribution for the linguistic variable.

5.3 Peer-to-peer Information Exchange

If, during query process, a user considers that the result has a high degree of relevance, the user can save the result in their user preferences. The user can share the results of this successful query with peers by sending them a reference to this query. Peers are able visualize the resulting images directly, without an actual query action. A peer user could adopt the same possibility distribution in their user-specific profile for future retrievals.

5.3.1 Visual Semantic Synchronization

Experts may use different descriptions for the same pathology due to their training and geographical locations. For example, the "Tree-in-bud" (TIB) pattern is a direct CT scan finding of bronchiolar disease. The same pattern could also be called "Finger-in-glove" [51]. In order to effectively accommodate different users and ensure accurate and timely results, there is a need for our system to address this semantic-variation issue because it can negatively affect the system performance.

It is quite possible that users may have a clear visual picture of a candidate semantic term that describes a desired perceptual category but be unfamiliar with the linguistic variables and/or semantic terms used by others and deposited to our system in the shared ontology. In addition, the same semantic meaning may already exist in the shared ontology but be described differently. In such cases, querying the system by selecting semantic terms from the shared ontology will have limited relevance to the user. For these situations, we provide a system module to synchronize the meaning of the semantic terms between the user's semantics and the shared ontology if any inconsistency in wording exists. This module identifies linguistic variables or semantic terms that refer to the same perceptual category in the knowledge base and creates a synonym database for information exchange.

This process has an iterative approach that includes two steps: image set selection, and

semantic set refinement. To reduce the burden on the user, only representative images that cover all meaningful semantics in the shared ontology will be displayed. To accomplish this, we partition the semantic terms into groups. For each group, images displayed to the user maximize the relevance of the associating semantic terms. The user will be asked to rate the images on a scale from 0 to 10, with 0 corresponding to *Excellent Counter-example*, 5 corresponding to *Neutral*, and 10 to *Excellent example*. After several iterations, the system is expected to converge to the most relevant semantic term from the shared ontology. This module is implemented using the pseudocode shown in Figure 23.

5.3.2 Image Set Selection

As mentioned previously, image selection maximizes the relevance of displayed images to the semantic terms in the image set. Let $N_S = |S|$ be the number of relevant semantic terms from the shared ontology. The semantic terms are partitioned in N_G groups where:

$$N_G = \lceil min(\psi * ln(N_S), N_S) \rceil \tag{16}$$

This formula ensures that, when the semantic set is small, semantic terms are grouped individually. It also limits the number of groups when the semantic set is big by using the logarithmic function. The parameter ψ is used to scale up the result of the logarithmic function, so it determines a reasonable number of groups to be used. For each group in N_G, the system will display two images. For example, if the relevant semantic term set includes 53 terms, and considering $\psi = 3.5$, the system will display 28 images, partitioned into 14 groups. The semantic terms with the highest correlated degree of satisfaction will be clustered in the same group. The degree of correlation among semantic terms is computed off-line every time new images are added to the database.

The system selects relevant images to the g^th group S_g–shown in lines 09 to 10 of pseudocode 23–by computing a degree of relevance γ^i of each image i to the semantic terms in the group, using the formula:

$$\gamma^i_{n_G} = \underset{\varsigma_g}{\text{argmin}}(g^{s_G}) \quad | s_G \in S_G \tag{17}$$

```
1  SynconizeSemantics(user)
2   DIS ← database image set
3   STS ← initial semantic term set
4   B0 ← preset threshold
5   do
6      NSG ← number of semantic groups
7      NIG ← 2 ▷ number of images per group
8      for i ← 1 to length[DIS]
9         for j ← 1 to length[STS]
10           compute gamma[i][j]
11     RIS ← array[1, 2 * NSG] ▷ relevant image set
12     for j ← 1 to length[STS]
13        RIS[j] ← maximum gamma image for group j
14        RIS[j + NSG] ← second maximum gamma image for group j
15     display RIS to user
16     UIR[1, 2 * NSG] ← user ratings for images in RIS;
17     for i ← 1 to length[STS]
18        for j ← 1 to length[RIS]
19           if RIS[j] > 5 then
20              compute e[i][j];
21           else
22              compute c[i][j];
23        compute BE[i], BC[i]
24        B[i] = min(BE[i], BC[i])
25        if B > B0 then
26           add STS[i] to TMPSTS
27     STS ← TMPSTS;
28  while length[STS] > 1
29  return
```

Figure 23: Pseudo code for visual semantic synchronization.

This approach guarantees that all of the other semantic terms in the group will have a degree of significance greater than or equal to γ^i. Then, we maximize γ^i among all of the images in the database. As shown in lines 12 to 14 of the pseudocode Figure 23, an image i is selected to be displayed in the G_g group if:

$$\gamma_n^{i^*} = \underset{i}{\text{argmax}}(\gamma_{n_G}^i) \tag{18}$$

We repeat the same image selection process for other semantic groups, without including the already selected images. After images are selected for all of the groups, the system displays them to the user for rating. Table 4 shows an example of user rating on a scale from 0 to 10, with 0 corresponding to *Excellent Counter-example* and 10 to *Excellent example*. Once the system receives the user's ratings for this image set, it further evaluates the relevance of each term in the semantic set to decide the next relevant semantic term set. This iterative process stops when only one semantic term is determined to be relevant.

Table 4: Example of User Ratings

Image Name	Rating	Meaning
Essence-054	2	Good Counter-example
Essence-418	9	Very Good Example
Essence-809	10	Excellent Example
Essence-941	1	Very Good Counter-example
Essence-980	0	Excellent Counter-example
Essence-520	3	Fair Counter-example
Essence-963	6	Poor Example
Essence-565	7	Fair Example

5.3.3 Semantic Set Refinement

The initial semantic set selected by the system is often too general to finalize the synchronization of the semantic meaning. The system will take the ratings of positive examples and counterexamples from the user's feedback to select a more significant set of images for the next iteration. This process intends to create a much smaller set of semantic terms from the shared ontology. Once a new set of semantic terms is defined, a new set of images is presented to the user. The user can follow the same process described in the previous

sub-section to refine the synchronization results. Let i_e be a positive example image, ς be a term in the relevant semantic set, and r be the user rating. We define $\beta_e(\varsigma, i_e)$ as the relevance degree of the positive example i_e to the semantic term ς with:

$$\beta_e(\varsigma, i_e) = min(g^{\varsigma}, r) \tag{19}$$

Similarly, let i_c be a counterexample image selected by user, ς a semantic term, and r the rating attributed by the user to the image i_c. We define $\beta_c(\varsigma, i_c)$ as the degree of dissimilarity of the counterexample image to the semantic term ς:

$$\beta_c(\varsigma, i_e) = max(1 - g^{\varsigma}, r) \tag{20}$$

From the previous two equations we can compute $B_e(\varsigma)$ as the degree to which there exists at least one highly representative example for the semantic ς. We also compute $B_c(\varsigma)$ as the degree to which all highly representative counterexamples are irrelevant to the semantic ς,

$$B_e(\varsigma) = max(\beta_e(\varsigma, i_e)) = max(min(g^{\varsigma}, r)) \tag{21}$$

$$B_c(\varsigma) = min(\beta_c(\varsigma, i_e)) = min(max(g^{\varsigma}, r)) \tag{22}$$

Further, we can estimate the overall degree of relevance of a semantic term ς to a set of rated images by computing the following:

$$B(\varsigma) = min(B_e(\varsigma), B_c(\varsigma)) \tag{23}$$

A semantic term will be excluded from the set if the overall relevance falls below a threshold. This process helps us select the most relevant semantic terms that matches with the user's candidate semantic term. If the process does not converge to the most relevant one, the system applies this entire process for the next iteration, until no more semantic terms from the shared ontology are excluded.

5.3.4 Updating the Knowledge Base

When the query refinement is completed, the user is presented with an option to enter their description of the candidate semantic term. The new description is then populated into the knowledge base using entries with the attributes: <*indexing-code*, *type*, *value*>. For example, if the most meaningful semantic term in the shared ontology synchronized with the user's candidate term is *fin_gl_big*, and the user description is *Big finger-in-glove*, the following new description will be populated into the knowledge base: <*fin_gl_big*, *synonym*, *Big finger-in-glove*>.

CHAPTER VI

EXPERIMENTAL RESULTS

To evaluate our approach, we have used four image databases: (1) Maize, (2) Geospatial, (3) Medical, and (4) UCI. From these databases we have created a number of 48 datasets from these databases. Below, we describe these datasets, the methodology of creating the datasets, and experimental results.

6.1 Datasets

6.1.1 Maize Datasets

The maize database consists of 135 images of maize leaves collected at South Farm University of Missouri. For each image, feature extraction algorithms [121] are applied to form a 176-dimensional feature vector. The features used for semantic queries are roundness of lesions, lesion spectral features, leaf spectral features, lesion size, nearest-neighbor distances, lesion coverage ratio, average lesion number per pixel, and texture features [121]. Also, these images were labeled to include one or multiple mutant phenotypes from the following set: *les1*, *les6*, *les7*, *les8*, *les13*, *les17*, *les18*, and *les19*. [77] Using this data we created seven datasets with number of classes ranging from two to eight by randomly removing one class at a time. As seen in Table 5, we named these datasets *Maize-2*, *Maize-3*, *Maize-4*, *Maize-5*, *Maize-6*, *Maize-7*, and *Maize-8*.

Table 5: Detailed information of the 176-dimensional *Maize* datasets

Dataset	Sample Size	Classes
Maize-2	44	2
Maize-3	62	3
Maize-4	79	4
Maize-5	89	5
Maize-6	108	6
Maize-7	120	7
Maize-8	135	8

6.1.2 Geospatial Datasets

From the Geospatial database, we used a set containing 1032 image tiles from three cities in Missouri. For each image tile, feature extraction algorithms [122] are applied to form a 227-dimensional feature vector. The features used for semantic queries are spectral histogram, co-occurrence texture, linear structure, and aggregate object features. Also, these images were labeled by experts to include one or multiple labels from the following set: *Commercial*, *Industrial*, *Residential*, *Grassland*, *Cropland*, or *Forests*. A total of 155 images in this dataset were assigned with multiple labels. Using this data, we created seven datasets with number of classes ranging from two to eight by randomly removing one class at a time. As seen in Table 6, we named these datasets *Geospatial-2*, *Geospatial-3*, *Geospatial-4*, *Geospatial-5*, *Geospatial-6*, *Geospatial-7*, and *Geospatial-8*.

Table 6: Detailed information of the 227-dimensional *Geospatial* datasets

Dataset	Sample Size	Classes
Geospatial-2	278	2
Geospatial-3	407	3
Geospatial-4	529	4
Geospatial-5	652	5
Geospatial-6	786	6
Geospatial-7	915	7
Geospatial-8	1030	8

6.1.3 Medical Datasets

Table 7: Detailed information of the 40-dimensional *Medical* datasets

Dataset	Transactions	Classes
HRCT-2	606	2
HRCT-4	1456	4
HRCT-6	1615	6
HRCT-8	1696	8
HRCT-10	1769	10
HRCT-12	1808	12
HRCT-14	1823	14
HRCT-16	1841	16

The Medical database contains 1841 HRCT image of lungs. For each image, feature extraction algorithms [123] are applied to form a 40-dimensional feature vector. Also, these images were labeled by experts to include 23 perceptual categories of lung pathologies [120]. Using this data we created eight datasets with number of classes ranging from two to 16 by randomly removing three classes at a time. As seen in Table 7, we named these datasets *HRCT-2*, *HRCT-4*, *HRCT-6*, *HRCT-8*, *HRCT-10*, *HRCT-12*, *HRCT-14*, and *HRCT-16*.

6.1.4 University of California Irvine Datasets

Table 8: Detailed information of the *UCI* datasets

Dataset	Transactions	Features	Classes
anneal	808	38	5
austral	690	15	2
auto	205	25	7
breast	699	9	2
cleve	303	13	2
crx	690	15	2
diabetes	768	8	2
german	1000	20	2
glass	214	9	6
heart	270	13	2
hepatitis	155	19	2
horse	368	22	2
hypo	3772	29	4
iono	351	34	2
iris	150	4	3
labor	57	16	2
led7	3200	7	10
lymph	148	18	4
pima	768	8	2
sick	3772	29	2
sonar	208	60	2
tic-tac	958	9	2
vehicle	846	18	4
waveform	5000	40	3
wine	178	13	3
zoo	101	17	7

The UCI Machine Learning Repository [7] contains a collection of databases that are used by the machine learning community for the empirical analysis of machine learning

algorithms. This database maintains 160 datasets and has been widely cited used as a source of machine learning datasets. For our testing, we have selected 26 datasets that were also used by other researches in their evaluation. The characteristics of these datasets are shown in Table 8. In our experimenets we have used the raw data converted to the attribute-relation file format [51].

6.2 Data Preparation

Data preparation is very important in any data mining process. Features extracted from images are not always nicely distributed, and may contain noise and outliers. The purpose of the data preparation is to transform the raw data to uncover useful patterns hidden in the data, increase data mining efficiency and improve knowledge discovery [103]. For our experiments we have used the following data preparation methods:

- normalization to equalize ranges of the features and their effect in the computation of similarity.
- Cox and Box transformation [21] to reduce the heterogeneity variance. Theoretical studies show that this transformation can achieve a similar effect with a threefold increase in sample size [84].
- genetic feature selection [62] to reduce the dimensionality of feature space and increase the efficiency of the association rule mining process.

6.3 Classification Results

For classification purposes, we have compared our approach to the following classification algorithms: BayesNet [100], Artificial Immune Recognition System (AIRS) [139], Learning Vector Quantization (LVQ) [71], Support Vector Machine (SVM) [135], C4.5 decision tree [104], Classification Based on Associations (CBA) [82], Classification based on Predictive Association Rules (CPAR) [149], and Apriori Total From Partial Classification (AprioriTFPC) [35]. For the BayesNet, AIRS, LVQ, and C4.5 algorithms, we have used the WEKA implementation [51] using the default settings. The authors of the CPAR, AprioriTFPC,

and CBA algorithms provided original software through their website. For the LVQ algorithm, we varied the number of codebook vectors for empirically best performance. For the SVM algorithm, we have used the LIBSVM implementation [30] using a linear kernel. CBA, CPAR, and AprioriTFPC, were downloaded from authors' sites. For CBA and AprioriTFPC algorithms, we set a minimum level of support of 1%, confidence 50%, and 400,000 maximum number of frequent item sets.

Table 9: Classification accuracy (%) for the *Maize* datasets

Dataset	BayesNet	AIRS	LVQ1	SVM	C4.5	CBA	CPAR	ATFPC	Essence
Maize-2	**100**	90.9	**100**	**100**	**100**	95.5	**100**	*88.5*	**100**
Maize-3	98.38	*85.48*	98.38	98.38	93.54	92.4	90.47	86.66	**100**
Maize-4	97.46	*73.41*	94.93	97.46	94.93	83.75	91.07	83.75	**100**
Maize-5	92.13	*62.92*	92.13	91.01	89.88	81.12	76.38	72.91	**94.16**
Maize-6	91.66	69.44	92.59	89.81	84.25	84.54	83.45	*62.9*	**95.37**
Maize-7	85	*58.33*	85.83	86.66	78.33	63.25	78.33	70.83	**89.16**
Maize-8	83.7	*58.51*	85.92	86.66	80	58.59	78.46	73.35	**92.63**
Average	92.61	*71.28*	92.82	92.85	88.70	79.87	85.45	76.98	**95.90**

Table 9 shows the accuracy of classification for the *Maize* datasets. In comparison with the eight methods mentioned previously, the *Essence* algorithm ranks first for average accuracy over the eight maize datasets. SVM, LVQ1, and BayesNet also rank in the top four for classifying the maize data sets. Due to the fact that, for these datasets, the semantic classes are separable, CBA, AprioriTFPC, and CPAR methods stop the training process as soon as correct classification in training was achieved. This fact, correlated with the use of only one association rule for classification of new instances, results in overfitting and reduction in classification performance. This issue is avoided in *Essence* that is maximizes the difference in relevance among semantic classes and uses multiple association rules to classify new instances.

Table 10 shows the accuracy of classification for the *Geospatial* datasets. The *Essence* algorithm ranks second for average accuracy over the eight datasets behind SVM. In the top four also rank CPAR and C4.5. The high density of this dataset with multiple semantic labels present in geospatial images favors the SVM approach that projects the feature space

Table 10: Classification accuracy (%) for the *Geospatial* datasets

Dataset	BayesNet	AIRS	LVQ1	SVM	C4.5	CBA	CPAR	ATFPC	Essence
Geospatial-2	93.88	94.96	97.12	**97.84**	97.84	91.73	95.71	*90.99*	97.48
Geospatial-3	90.9	93.85	**95.08**	94.59	91.89	*84.9*	95.07	86.72	94.07
Geospatial-4	84.68	84.12	88.22	**92.62**	86.57	*71.1*	90.05	81.09	89.56
Geospatial-5	78.06	72.54	74.69	**85.42**	75.46	*57.36*	80.36	66.88	80.82
Geospatial-6	74.04	70.86	70.1	**83.33**	72.64	*49.84*	78.1	64.36	75.69
Geospatial-7	68.19	64.91	61.53	**78.68**	67.97	*36.32*	68.63	61.53	71.63
Geospatial-8	58.54	53.39	50.58	**66.79**	58.25	*22.87*	54.07	51.35	62.31
Average	78.32	76.37	76.90	**85.61**	78.66	59.16	80.28	*71.84*	80.63

into hyperspace for linear separation of the semantic labels. However we can also note that the associative classification methods such CPAR perform better on the geospatial datasets as compared with the maize datasets. This is because the training does not stop early and the model is less overfitted.

Table 11: Classification accuracy (%) for the *Medical* datasets

Medical	BayesNet	AIRS	LVQ1	SVM	C4.5	CBA	CPAR	A-TFP	Essence
HRCT-2	92.9	86.96	*76.86*	95.37	96.03	89.23	**97.04**	88.11	96.37
HRCT-4	81.52	*67.3*	79.03	85.41	**93.84**	75.78	92.81	81.97	92.96
HRCT-6	76.25	68.86	*59.36*	78.76	88.65	68.04	**90.1**	77.7	89.31
HRCT-8	66.31	60.83	*54.04*	77.53	85.95	61.45	**88.62**	72.38	85.59
HRCT-10	64.43	56.14	50.44	74.61	82.21	*49.5*	**84.32**	68.62	79.02
HRCT-12	62.39	50.77	*42.76*	72.73	79.32	44.36	**82.89**	61.76	76.85
HRCT-14	60.45	49.4	40.58	72.84	78.15	*37.01*	**81.16**	63.12	72.44
HRCT-16	59.73	52.78	52.62	73.03	78.32	*39.01*	**81.47**	62.98	73.57
Average	70.5	61.63	*56.96*	78.79	85.31	58.05	**87.30**	72.08	83.07

Table 11 shows the accuracy of classification for the *Medical* datasets. *Essence* algorithm ranks third for average accuracy over the eight datasets behind CPAR and C4.5 and ahead of SVM. The high density of these datasets combined with the reduced number of features makes SVM and Essence overfit by not finding best separation hyperplane and by using low confidence association rules respectively.

Table 12 shows the accuracy of classification for the *UCI* datasets. As seen on this

table, the *Essence* algorithm ranks first for average accuracy over the 26 datasets. Top four is completed by CPAR, SVM, and CBA. This table also shows the relevance of using association rules based methods for classification.

Table 12: Classification accuracy (%) for the *UCI* datasets

Dataset	BayesNet	AIRS	LVQ1	SVM	C4.5	CBA	CPAR	A-TFPC	Essence
anneal	95.43	92.09	*88.86*	**99.1**	98.88	98.66	98.44	89.86	98.33
austral	84.92	83.18	*68.69*	85.01	**85.36**	80.4	85.07	84.92	84.05
auto	80.97	80.48	72.68	80.48	89.26	*70.26*	**93.66**	84.47	92.66
breast	**97.13**	96.7	96.13	96.42	95.13	95.38	96.13	*94.56*	96.42
cleve	81.84	80.85	*66.6*	82.5	76.89	81.51	80.49	74.55	**83.13**
crx	84.92	83.18	*70.86*	84.63	**85.36**	80.99	85.07	84.89	84.05
diabetes	76.17	71.35	73.3	**77.47**	74.6	76.65	73.95	*70.7*	75.64
german	74.2	*68.8*	70.4	**76**	72	71.77	72	69.9	73.39
glass	71.02	*61.21*	71.96	63.08	67.57	74.99	**76.68**	65.08	70.6
heart	82.22	75.55	*64.44*	**84.44**	77.77	84.36	82.22	69.99	82.96
hepatitis	83.22	84.51	*75.48*	**85.16**	78.06	83.94	78.66	79.37	84.58
horse	80.7	*63.04*	68.75	84.23	**85.59**	80.95	82.89	76.63	82.1
hypo	98.56	*87.72*	92.84	96.18	**99.49**	98.81	99.25	92.31	96.89
iono	*85.18*	88.88	86.6	86.6	88.6	**92.53**	92.3	88.07	91.74
iris	94	96.66	96	**97.33**	95.33	*92.66*	94	94	94.66
labor	**91.22**	*77.19*	89.47	89.47	78.94	89.66	88.33	78.66	90
led7	73.43	*40.93*	**73.87**	73.59	73.84	69.25	73.25	66.93	73.46
lymph	**87.16**	79.05	77.7	83.78	76.35	79.64	79.85	*73.09*	82.57
pima	74.86	*70.83*	74.6	75	74.34	**75.5**	75	74.61	75.38
sick	96.95	*90.66*	93.16	96.6	**98.8**	97.15	96.89	93.87	97.53
sonar	77.88	70.67	75.48	77.4	74.03	*65.7*	**83.69**	70.8	79.4
tic-tac	69.62	89.45	80.68	*65.34*	85.49	99.6	**100**	65.76	98.95
vehicle	59.81	64.18	62.88	**80.26**	72.45	65.7	70.19	*58.49*	67.37
waveform	80.26	*71.98*	83.66	**86.58**	75.2	80.88	79.98	75.14	82.96
wine	**97.19**	93.82	*77.52*	95.5	93.82	93.77	95	89.34	96.11
zoo	94.05	93.06	*58.61*	95.04	92.07	94.18	91.09	92.18	**95.09**
Average	83.57	79.07	*77.35*	84.50	83.27	83.64	85.54	79.16	**85.77**

The overall conclusion is that, although the *Essence* system was designed for ranking images by semantics, it can be successfully applied to classification problems. Although Essence does not rank first over all individual datasets, in average, it ranks first over 48 datasets when compared with eight other methods. Besides this, Essence has the advantage of providing ranking for each individual semantic as discussed in Chapter 4. For detailed

classification results using the *Essence* method, the reader is referred to Appendix A.

6.4 Ranking Results

Ranking is very important to users because when searching an image database for visual patterns they expect images containing the desired visual pattern be at the top of the ranking result. High ranking of relevant images is a must for image retrieval systems because it determine the user's perception of the effectiveness of the application. Thus, in this experiment, we evaluate the ranking effectiveness of the *Essence* algorithm using two measures: (1) average ranking precision and (2) precision-recall [25]. We apply the ranking algorithm on the 48 datasets. The average ranking precision, $Prec = \frac{1}{|TP|}\sum_{i=1}^{|TP|}\frac{TP_i}{TP_i+FP_i}$, is reported in this section while complete detailed precision-recall charts are reported in Appendix B. Due to the fact that the eight data mining methods described in previous section are intended to classification and modifying them for ranking is not according to the original intent, in this section we display only ranking results from the *Essence* method.

Table 13: Ranking average precision (%) for the *Maize* datasets

Dataset	Average	Min	Max	Stdev
Maize-2	100	100	100	0
Maize-3	98.6	97.22	99.73	2.96
Maize-4	98.7	96.1	100	1.9
Maize-5	95.5	83.14	100	7.96
Maize-6	89.9	73.42	96.63	11.95
Maize-7	89.6	71.13	99.07	12.16
Maize-8	90	72.39	100	11.75

The results of the ranking experiment are shown in Tables 13 - 16. The ranking precision varies among different semantic classes. The worst performance is observed for labels that have very few training instances. In such cases the algorithm cannot identify enough association rules for that semantic. Another case is seen for the *Geospatial* datasets where *Isolated Road* (RD) represents a small fraction of an image tile and always shows together with *grassland*, *forest* or *farmland*. In this case, the algorithm retrieves images with similar visual patterns but different semantic such a *Stream*.

Table 14: Ranking average precision (%) for the *Geospatial* datasets

Dataset	Average	Min	Max	Stdev
Geospatial2	98.5	98.1	99	0.2
Geospatial3	95.6	92.7	97.1	0.8
Geospatial4	94	89.5	97.3	9.94
Geospatial5	82.1	67.7	93.2	2.7
Geospatial6	74.7	59.3	91.4	4.1
Geospatial7	75.2	61.8	90.8	5.1
Geospatial8	73.8	58	91.11	7.9

Table 15: Ranking average precision (%) for the *Medical* datasets

Dataet	Average	Min	Max	Stdev
HRCT-2	97.8	96.64	98.91	1.09
HRCT-4	88.1	69.78	95.97	5.63
HRCT-6	88.3	50.7	93.2	5.8
HRCT-8	69.9	49.7	88.9	14.4
HRCT-10	70.3	27.4	94.2	17.2
HRCT-12	69.6	37	89.5	18.4
HRCT-14	66.6	32	85.8	13.3
HRCT-16	63.5	31.7	100	26.9

6.5 System Customization of Semantic Settings

To demonstrate the performance of our model, we designed three experiments. The first experiment tests the improvement in retrieval precision after the system evolves its shared knowledge settings by adapting to domain expertise. The second one demonstrates the appropriateness of using the sigmoid functions described in Chapter 3 to quantize the semantic modeling. Finally, the third experiment evaluates the performance of the semantic integration mechanism described in Chapter 3 when searching for synonymous semantic terms. Upon completion of the experiments, the users were asked to complete a usability test. The results of the test are discussed at the end of this section.

Table 16: Ranking average precision (%) for the *UCI* datasets

Dataset	Average	Min	Max	Stdev
anneal	92.3	82.4	100	7.1
austral	86.4	85.7	87.1	0.8
auto	94.2	79	100	5.18
breast	97	94.3	99.7	0.4
cleve	84.6	84.1	85.2	1.2
crx	86.4	85.7	87.1	0.8
diabetes	78.3	73.2	83.5	1.5
german	84.5	51.7	84.5	1.4
glass	70.2	48.7	87.2	16.3
heart	87.9	87.8	881	1.4
hepatitis	82.4	72	92.8	2.4
horse	87.7	85.5	89.9	1
hypo	92.7	82.8	100	5
iono	93.1	92.8	93.3	1.1
iris	94.4	93.33	100	1.9
labor	93.6	88.8	93.6	1.7
led7	74	54.9	91.4	6.6
lymph	95.5	90	100	3.53
pima	77.7	72.6	82.9	1.2
sick	93.6	87.6	99.7	0.5
sonar	87.3	85.4	89.2	1.6
tic-tac	84.4	76.8	92	1
vehicle	68.5	33.4	96.3	3.3
waveform	80.9	76.7	83.6	0.8
wine	97.3	96.6	98.8	1.5
zoo	97.6	90	100	2.4

6.5.1 Simulated Scenario for Experiments

Users were explained, both visually and semantically, the typical appearance of cysts and nodules using the sketch shown in Figure 24 and comparing them with similar lung pathologies such as emphysema. Each user is then instructed by a domain expert, using a training image set, to identify the visual abnormalities of these pathologies on real HRCT lung images. This process emphasizes the semantic terms that will be used in the experiment such as *many small nodules* and *many big cysts*. For example, the term *Cyst* is used to refer to a lesion of a lung having the following characteristics [88]: well defined, circumscribed, air-containing, and thin-walled with size greater than 3 mm. It differentiates from *Emphysema*

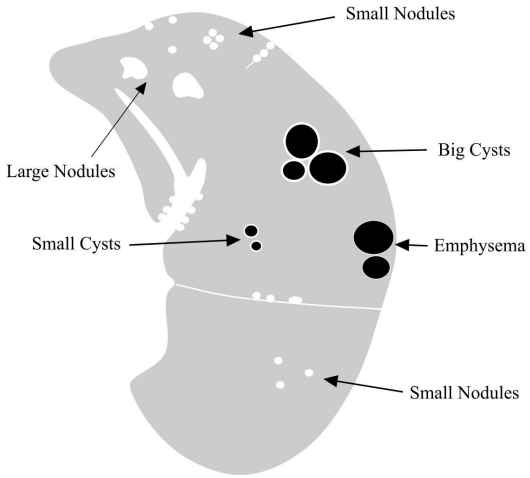


Figure 24: Typical appearance of different lung pathologies used in our experiment.

by the fact that the latter shows very thin and less defined walls. From HRCT images, a *Cyst* (perceptual category) with a diameter between 10 pixels to 20 pixels (range of values) might be classified as medium size. The term *Small nodule* [88] refers to a rounded density that does not correspond to vessels and is represented by a spherical structure having less than 1 cm in diameter.

All of the experiments reported in this dissertation require users to rate the relevance of HRCT images of lungs for one of these semantic terms. We used a rating scale from 0 to 10, where 0 corresponds to *Excellent Counter-example* and 10 to *Excellent Example*.

6.5.2 Improving the Retrieval Precision through Adapting the Shared Ontology Settings

In this experiment, users were assigned to two groups: *Group 1*: active users, including two domain experts and three computer scientists (*userA1* to *userA5*), and *Group 2*: inactive users, including two computer scientists (*userA6* and *userA7*). We assumed that the inactive users, although they have the expertise to customize their setting, prefer to use only the shared ontology. The purpose of this experiment is to evaluate the improvement in retrieval precision for the inactive users by benefiting from the active users' domain expertise. This process involves a system level knowledge exchange as described in section 3. During this experiment, 869 images were rated by both active and inactive users.

To capture the evolving nature of this process, active users were asked to customize their

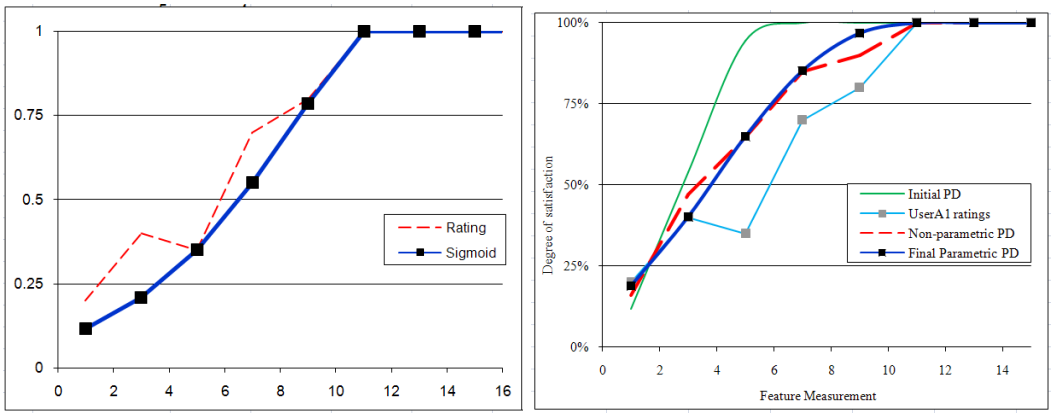


Figure 25: Possibility distribution (PD) for (a) *userA1*, and (b) shared ontology at stage 1 after *userA1* rating.

settings for *many big cysts* at different time intervals. At the end of each time interval, the system updated its default setting after each customization process. Each time interval represents a stage in the evolution of the shared ontology settings for this term. In this experiment, *userA1* customized his semantic profile at stage 1, *userA2* did at stage 2, and *userA3*, *userA4*, and *userA5*, did at stage 3. Figure 25(a) shows the shape of the user-specific possibility function after stage 1. It shows both the degree of satisfaction derived from the rating and its sigmoid approximation. For the shared possibility distribution shown in Figure 25(b), the system uses both the initial possibility distribution and *userA1*'s ratings to determine the updated non-parametric distribution, and later its sigmoid parametric approximation. Figure 26 follows the same idea but uses the initial possibility function, computed from stage 1. In Figure 27, the rating from three users contributes along with the previously determined possibility function to determine the shared possibility distribution.

At each stage, the inactive users were asked to query the database for *many big cysts*, using the shared ontology settings. Then, they evaluated the retrieval result by rating the displayed images. From their ratings, we computed the retrieval precision as the percentage of images rated above 0.7 in the retrieval result (*Good* to *Excellent Example*). It improved from 25% in the initial stage to 65% after all training experts updated their possibility distribution. We conclude that using domain expertise to evolve the shared semantic settings can improve the retrieval precision for new and inactive users. The average time to complete

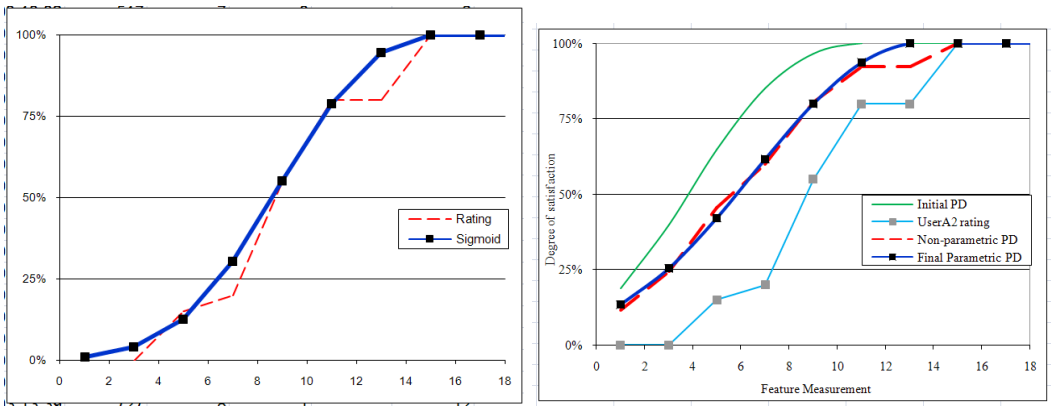


Figure 26: Possibility distribution (PD) for (a) *userA2*, and (b) shared ontology at stage 2 after *userA2* rating.

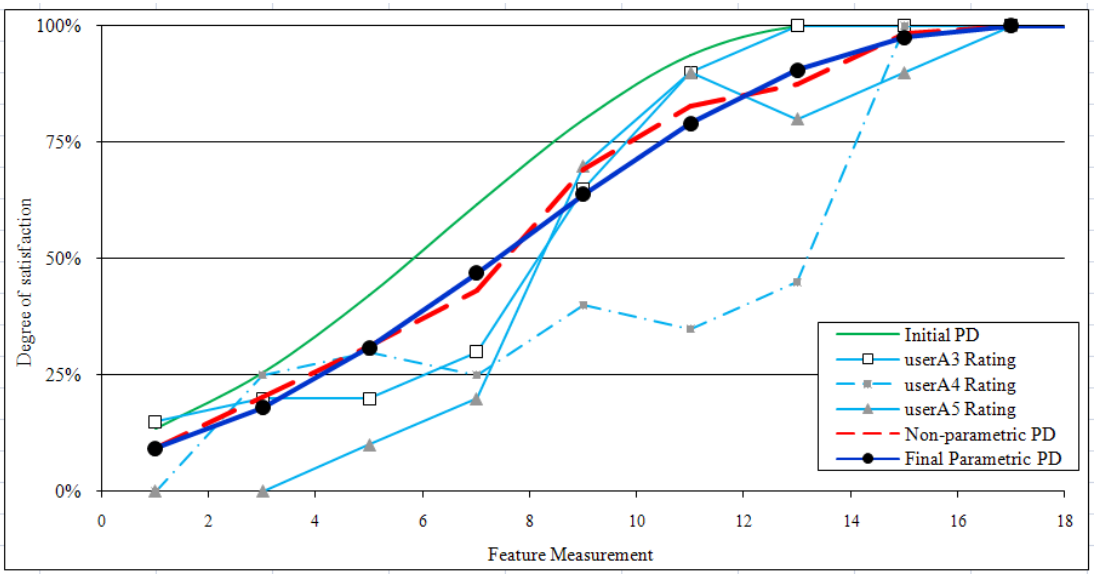


Figure 27: Default possibility distribution after their ratings upon *userA3*, *userA4*, and *userA5* ratings.

this task, which required ratings of 20 images, was 110 seconds for computer scientists (standard deviation of 53 seconds) and 184.5 seconds for physicians (standard deviation of 42 seconds). From the above observations, the efficiency of the semantic customization process is acceptable for users. This is also consistent with the usability test which will be discussed in details later in this section.

6.5.3 Evaluating the Usage of Sigmoid Functions to Approximate the Possibility Function

This experiment evaluates appropriateness of using the sigmoid function in approximation of the possibility function. For comparison, we use a linear function $g_l(m) = max(0, min(1, (m-b)/(a-b)))$, in which a and b are the values of low-level image measurements m with the degree of satisfaction 1 and 0, respectively. To measure the efficacy of both functions, we computed the approximation error for both linear and sigmoid functions:

$$\sum_{all\ ratings} |g_l(m) - rating(m)| \tag{24}$$

$$\sum_{all\ ratings} |g(m) - rating(m)| \tag{25}$$

where $rating(m)$ is the user's rating for the measurement m and g is the sigmoid possibility function. The approximation performance was evaluated in eleven cases - seven of them were related to user-specific possibility functions and four to shared ones. The sigmoid function outperformed the linear function in ten out of eleven cases by decreasing the error rate by 31% on average.

6.5.4 Evaluating the Semantic Integration Mechanism when Searching for Synonymous Semantics

We asked three physicians (*userB1* to *userB3*) and three computer scientists (*userB4* to *userB6*) to rate the existence of a candidate semantic term by using a set of HRCT images and the synchronization mechanism. The searched term has a visual pattern associated with terms archived in our shared ontology. For this experiment, we used candidate semantic terms that are synonymous to our targeting terms - *many big cysts* or *many small nodules*.

The images presented to the user were selected according to the algorithm in Section 5.3.1 to cover all of the significant terms. By asking users to search for these visual patterns, we evaluated both the accuracy and rate of convergence in matching the candidate semantic terms with the targeting ones. The rate of convergence is defined as the number of iterations needed by the algorithm to converge to a unique semantic term.

Our experiments show that this process accurately converges to the targeting semantic term in 92.8% of the cases (26 out of 28 synchronizations). For both targeting terms, the process converged in approximately 2 iterations on average, which demonstrates the viability of our approach in semantic set refinement. However, the convergence rates differ between these two targeting terms. Figure 28(a) shows that synchronizing a candidate term to *many small nodules* requires 26% more iteration on average than synchronizing to *many big cysts*. There are three reasons for this result: 1) *many small nodules* is more likely to be co-existed with other semantic terms, 2) *many big cysts* is more easily recognizable than the *many small nodules* even without in-depth training, and 3) the behavior of each user can differ depending on their subjectivity. Figure 28(b) shows the convergence rate of this process for each user with different sizes of initial sets. The average time required to complete this task was 203 seconds for computer scientists (standard deviation of 118 seconds) and 292 seconds for medical experts (standard deviation of 59 seconds). The time required to do semantic synchronization is higher due to its recursive nature. However, the times measured in this experiment are reasonable to learn a new perceptual category without knowing the exact associating semantic term.

6.5.5 Usability Evaluation

Due to the fact that Essence is used by both medical experts and computer scientists, it is very important to evaluate how easy is for them to collaborate in such an environment. To achieve this goal, we developed a usability questionnaire based on the SUS usability scale [8]. However, the ten questions in the original test are too general for the purpose of our study. We added six more questions from other usability questionnaires [14],[38] that addressed some more specific issues such as terminology, functionality and usefulness

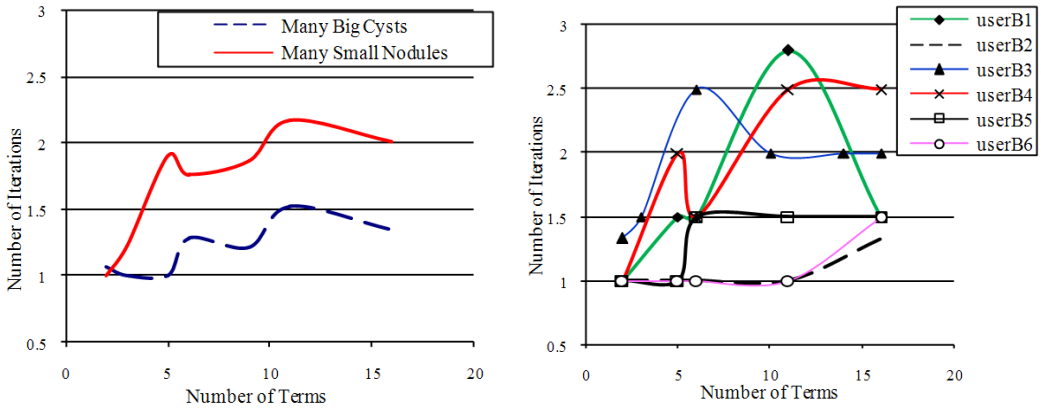


Figure 28: Average number of iterations performed for visual synchronization of semantic terms for each (a) semantic term and (b) user.

of retrieved images. Nine subjects, five computer scientists and four physicians, rated the usability according to the guidelines of the SUS test. The system was trained by experts before the experiment in order to stabilize the semantic assignments used in the experiment. Subjects completed the questionnaire at the end of the experiments discussed previously. The data was collected and further studied using analysis of means and variances of the usability ratings over the 16 questions. A perfect system would receive score 5 ratings. Listed in Table 17, we received 29.16% of score 5 ratings and 80.55% of score 4 or better ratings.

Table 17: Usability Test Result

Score	Computer scientists		Medical Experts		Overall	
	Percent	Cumulative	Percent	Cumulative	Percent	Cumulative
5	26.25	26.25	32.81	32.81	29.16	29.16
4	57.5	83.75	43.75	76.56	51.38	80.55
3	7.5	91.25	17.18	93.75	11.8	92.36
2	6.25	97.5	6.25	100	6.21	98.61
1	2.5	100	0	100	1.38	100

The lowest overall score was given to question "I understand the terminology used in the system" which received a score of 3. From the feedback provided by our subjects, medical terms are not intuitive to the computer scientists; while the interface terminology is not straightforward to the physicians. A noteworthy observation to report is that using

the search tree (as shown on the left panel of Figure 14) for semantic queries was new for most of the users at the beginning of the experiment. However, all users were successful in subsequent searches due to the intuitiveness of this type of search. On average, each semantic query takes 70.1 seconds to construct. The highest overall scores the system received were for its function integration and effectiveness. Computer scientists also appreciated the consistency of the system while the medical experts appreciated the manageability and results of the queries. The system was also evaluated on the SUS usability scale. The SUS-scale yields results between 0 and 100, with 0 for poor perceived usability and 100 for high perceived usability. The study of Nielsen and Levy [38] shows that a system with average usability gets a score around 64 on such a Likert scale, even though 50 represents neutral. The average SUS score for Essence was 77.22. Medical experts rated the system higher (average SUS score of 78.12) compared to the score of computer scientists (average SUS score 76.5). All of the physicians in this usability test considered the results of the semantic queries satisfactory without customizing their semantic assignment. Once the system is trained by domain experts, most of the physicians do not need to create new linguistic variables or customize existing semantic assignments. Under the condition when a new perceptual category is needed for certain newly discovered diseases, the community will ask for contributions from users. This knowledge exchange procedure consists of semantic synchronization and customization, and is believed to be acceptable by physicians who are enthusiastic about sharing their expertise to the databases.

6.6 User Customization of Semantic Settings

To demonstrate the performance of the customized semantic mappings, we designed a ten-fold validation experiment. We used an image database containing 1032 geospatial image tiles from three cities in Missouri. For each image tile, feature extraction algorithms [122] are applied to form a 227-dimensional feature vector. The features used for semantic queries are spectral histogram, co-occurrence texture, linear structure, and aggregate object features. Then, the dimensionality of feature space is reduced to 85 dimensions by applying a genetic feature selection algorithm [62]. Also, these images were labeled by experts to include one

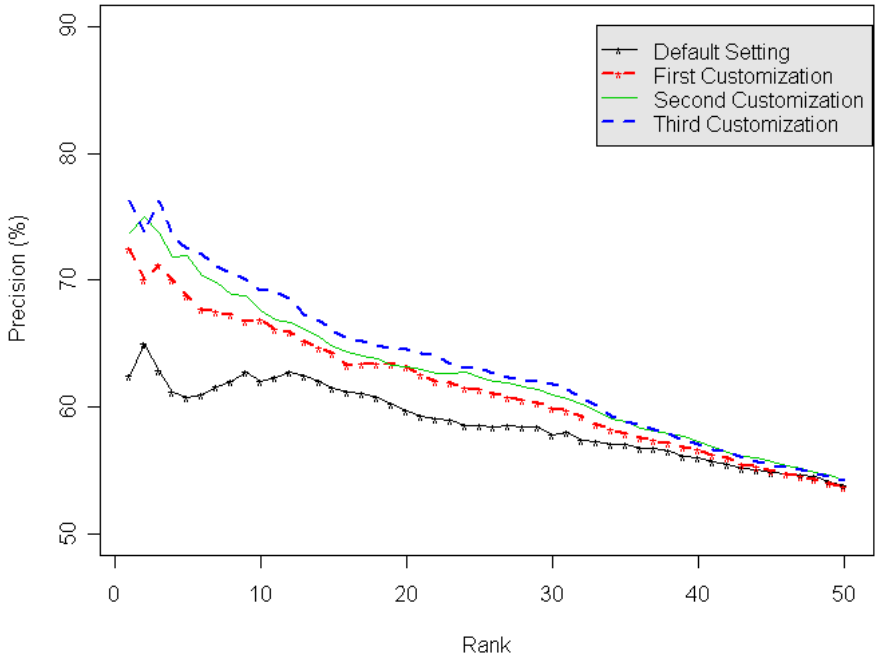


Figure 29: Improvement in precision of retrieval. The customization starts with the default setting and it is performed in three consecutive steps by adding both positive and negative examples to the training set.

or multiple labels from the following set: *Commercial*, *Industrial*, *Residential*, *Grassland*, *Cropland*, or *Forests*. A total of 155 images in this dataset were labeled with multiple labels. This dataset was then evenly divided into training and testing sets.

A semantic profile was built on the training set using the procedure described in [122]. This semantic profile was used as the initial profile for semantic queries and was evolved by successive customizations. For each semantic we performed three customization steps using the following procedure. First, a query is performed on the testing set for each semantic, and the top 50 images are retrieved. Then, we randomly add to the training set three examples and counter-examples from the query result. Finally, using the new training set, the semantic profile is customized according to the procedure described in Section 5 and another semantic query is evaluated.

The results of each semantic query were evaluated using precision, PR, and F-measure, $FM = \frac{2*PR*RC}{PR+RC}$, where RC is the recall. Figure 29 shows the average precision of retrieval

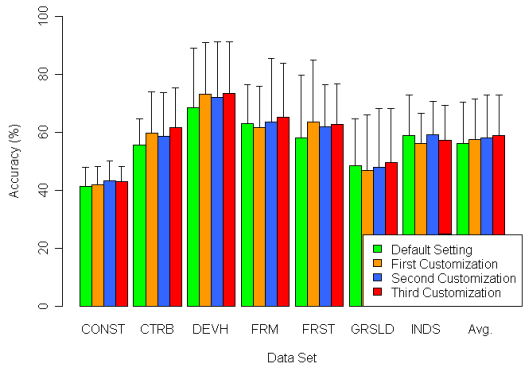


Figure 30: Improvement in F-measure for customization of each semantic of the *Geospatial-7* dataset.

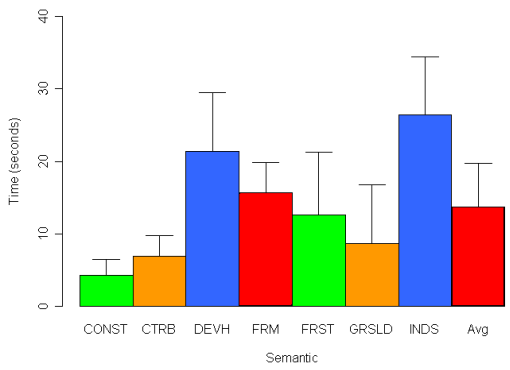


Figure 31: Time performance of semantic customization for customization of the *Geospatial-7* dataset.

results at each step of the customization process. For example, after customization, the precision of the top ten query results increased from 62% to 70%. Also, this figure shows that the gain tends to decrease after several iterations. This trend can be compensated by a complete rebuild of the semantic model. Figure 30 shows the average F-measure for each semantic. This shows us that for most semantics the customization process improves the quality of the semantic retrieval. However, the semantics *Industrial* (INDS) and *Grassland* (GRSLD) in the figure have mixed results. This is because the images in the dataset do not have a high clarity relevance to these semantics. The time necessary for semantic customization is shown in Figure 31. The average time was 14.3 seconds with a standard deviation of 7.16 seconds. This makes our approach more suitable to on-line semantic profile customization.

6.7 Performance Comparison for Using Sigmoid and Crisp Membership Functions

To demonstrate the relevance of using a customized sigmoid parametric approximation over the crisp approximation, we have designed a classification experiment on the geospatial and maize datasets. We maintained the same setting except for the type of function used to model the semantic assignment. We have preprocessed both datasets by applying: (1) feature normalization, (2) Box-Cox transformation, and (3) best-first feature selection. For

association rule we have used a maximum antecedent size of seven, 0.5% minimum support level, and 50% minimum confidence level. The results of this experiment are shown in Figures 32 and 33. As seen in this figure, using the sigmoid parametric approximation results in better accuracy than using crisp approximation, especially for increased number of semantic classes.

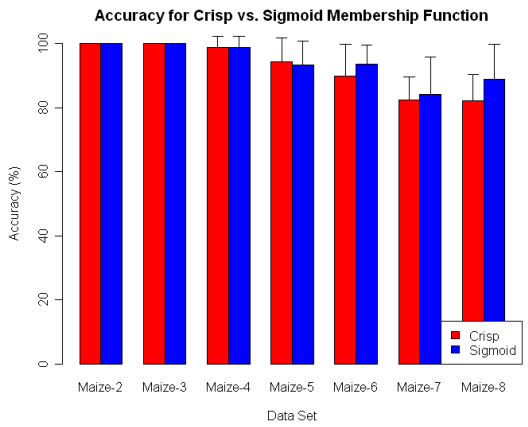


Figure 32: Improvement in average accuracy by using the sigmoid parametric approximation over crisp sigmoid parametric approximation for the *Maize* datasets.

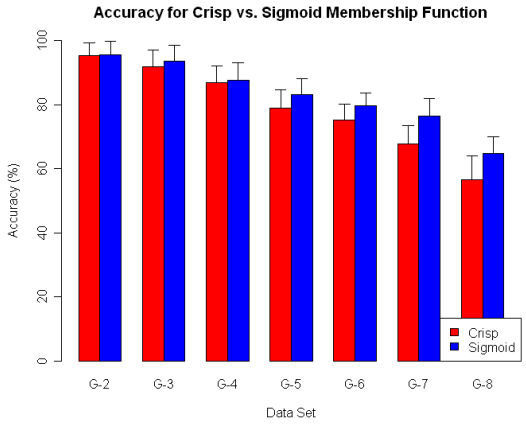


Figure 33: Improvement in average accuracy by using the sigmoid parametric approximation over crisp sigmoid parametric approximation for the *Geospatial* datasets.

6.8 Time Efficiency Experiments

6.8.1 Effects of Database Size on Mining Time

To evaluate the variation time necessary to mine association rules with the size of the database, we have designed an experiment on a 1–dimensional synthetic dataset. This dataset was generated using R [106]. In this dataset, data points assigned to each class have a normal mixture distribution as shown in Figure 34. Using this algorithm we have created six datasets that were named *Synt05* (5,000 training records), *Synt10* (10,000 training records), *Synt15* (15,000 training records), *Synt20* (20,000 training records), *Synt25* (25,000 training records), and *Synt30* (30,000 training records). Each of these datasets were further preprocessed as explained in Section 6.2. For association rule we have used a maximum antecedent size of four, 1% minimum support level, and 50% minimum confidence level.

The results shown in Figures 35 - 37 show that the size of the training dataset is not a

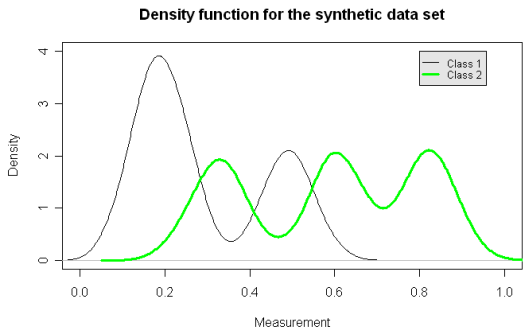


Figure 34: Distribution of the generated dataset. The dataset contains two classes with a normal mixture distribution.

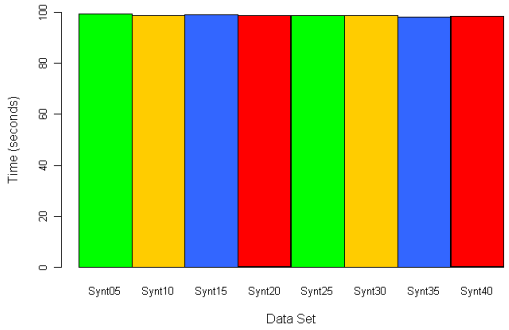


Figure 35: Average precision when varying the size of the synthetic training dataset between 5,000 and 30,000.

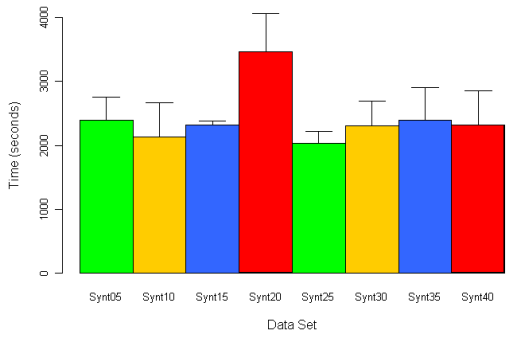


Figure 36: Average time for association rules mining when varying the size of the training dataset between 5,000 and 30,000.

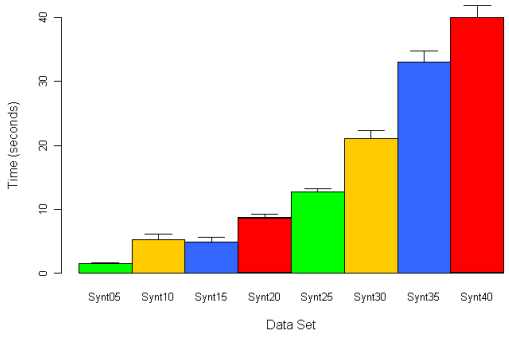


Figure 37: Average query time when varying the size of the training dataset between 5,000 and 30,000.

major factor in the time necessary to mine association rules.

6.8.2 Mining Time per Semantic

Figures 38 - 43 show the time necessary for mining association rules and ranking new images. These experiments were preformed on the *Maize-8*, *Geospatial-8*, and *HRCT-16* datasets. The times necessary for mining are proportional with the number of rules necessary for optimal ranking and separability of the semantic. For example, for the *Geospatial-8* dataset, images containing *Residential* (DEVH) semantic can also contain any other semantic such as *Forest*, *Grassland*, *Farmland*, *Industrial*, or *Commercial*. For this reason the algorithm needs more time to separate this semantic from the others. On the other hand *Isolated*

Road (RD) can be associated only with *Grassland*, *Forest*, or *Farmland* and therefore it reduces the size of the problem. For the *Maize-8*, ranking 26 images takes on average 0.02 seconds while for the *Geospatial-8* dataset it takes 0.6 seconds to rank 200 images on average. Similarly, for the medical dataset we need up to 3 seconds to rank 350 images of lungs. The search time for ranking is proportional with the number of association rules in the semantic model and with the number of ranked images. On average, each semantic model contains less than 118 association rules. However this number varies with the density of the dataset. For example, for the *Maize* datasets the average number of association ruled for each semantic model is 22, while for the *Geopatial* and *Medical* dataset we generated 86 association ruled and 118 respectively. The query time can be further reduced by using a space-filling curve indexing structure. Figure 44 shows the retrieval time as percentage of brute force. We conducted this experiment on the *Geospatial-8* dataset. For each semantic we retrieved between 25 and 200 relevant images. As shown in this figure the average search time for all semantics is less than 25% of the brute force. This is due to the fact that by using such an indexing structure we are able to avoid searching feature subspaces that are not relevant to the query semantic.

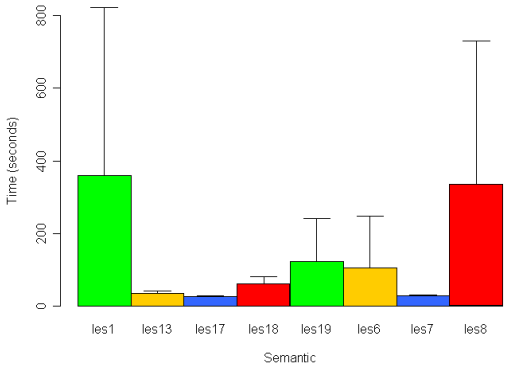


Figure 38: Average time for association rules mining for the *Maize-8* dataset.

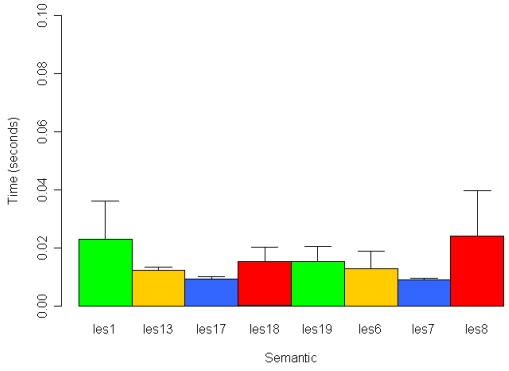


Figure 39: Average search time for the *Maize-8* datasets.

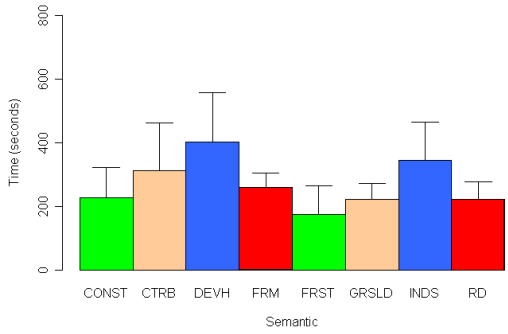


Figure 40: Average time for association rules mining for the *Geospatial-8* dataset.

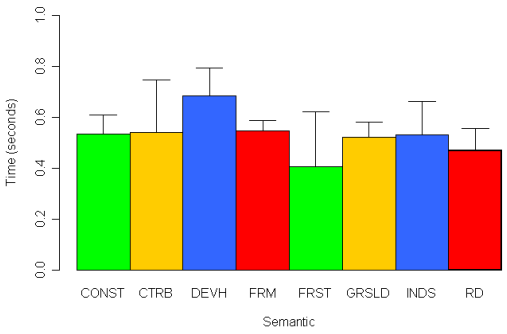


Figure 41: Average search time for the *Geospatial-8* datasets.

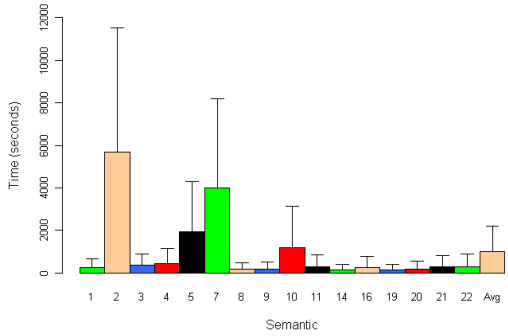


Figure 42: Average time for association rules mining for the *HRCT-16* dataset.

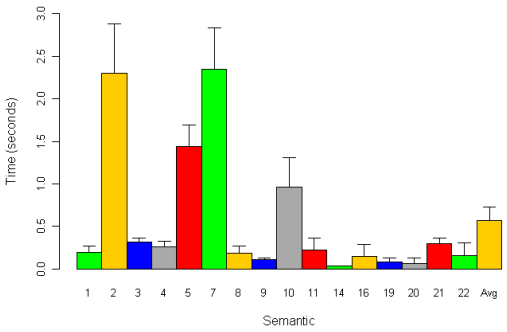


Figure 43: Average search time for the *HRCT-16* datasets.

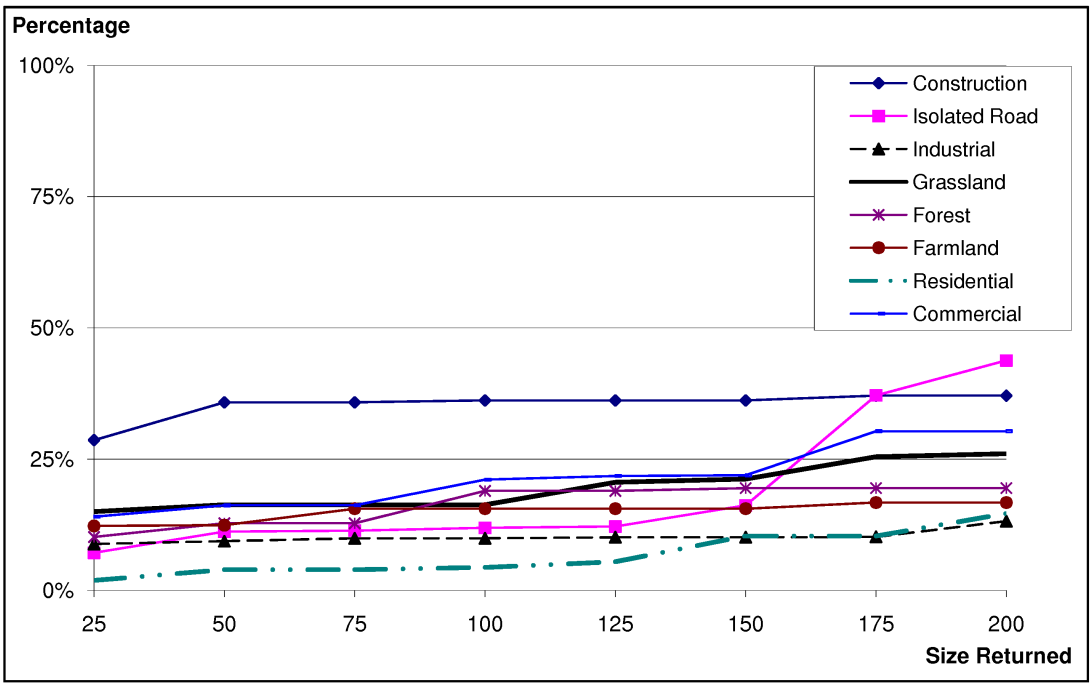


Figure 44: Average retrieval time as percentage of brute-force when using a space-filling curve indexing structure for the *Geospatial-8* dataset.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

The focus of this dissertation lies in investigating methods of knowledge management and exchange in content-based retrieval systems for large-scale biomedical media. In this area, we focused on providing researchers and experts with computer-based methods to acquire knowledge from different sources and on encoding this knowledge to make it available for peers' reference. Such methods use both explicit and tacit knowledge used by experts in their decision-making processes. Our approach was successfully applied to various domains, such as medical informatics for studying patterns of pulmonary diseases found in high-resolution computer tomography images, agronomy for modeling the phenotype-genotype correlations of maize mutants, and geospatial intelligence for discovering relevant knowledge from satellite images.

The research presented in this dissertation is concentrated in four areas: (1) semantic modeling in image retrieval systems, (2) knowledge exchange in collaborative environments, (3) conceptual change captured in computer-based applications, and (4) ontology integration in retrieval systems. To support this research, we have created a framework for knowledge representation and sharing called *Essence*. The backbone of this framework is a shared ontology based on common knowledge from experts and available literature. Using this framework, we extract and manage visual content found in domain-specific images. Visual patterns are mapped into semantics to provide methods for experts to search the database by these semantics. Experts can build their personalized semantic search criteria by customizing this mapping to fit their own preferences. Customization methods include adjusting mapping parameters using expert-in-the-loop methods or creating new personalized mappings that are presented to peers for evaluation and adoption. In the following sections, we provide an overview of the research in each of the above-mentioned areas.

7.1.1 Semantic Modeling in Image Retrieval Systems

Modern technology enables organizations to collect and store huge amounts of information at very low costs. Considering this large amount of data, human inspection of all images becomes infeasible. Computer algorithms can help analysts by filtering the image set that is presented for inspection. Most of the content-based techniques generate models that may not coincide with the information used by analysts for interpreting visual patterns. For a computer system to be successful, it needs to address the semantic gap between an analyst's models of visual patterns and the computer's representation of information. To reduce this semantic gap, we have researched methods of combining content-based and semantic techniques in information retrieval. The main contributions in this area are: (1) providing methods for mapping semantics into low-level features extracted by computer-vision algorithms; (2) evaluating methods for querying image databases by semantic using flexible association rules; and (3) modeling tacit knowledge captured from experts' feedback.

7.1.2 Knowledge Exchange in Collaborative Environments

Domain-specific knowledge exchange is difficult, especially due to the autonomy of research groups and the importance of the tacit component of the decision-making process. Domain experts have close concordance with their local environment setting, in which both previous experience and colleagues' opinions have a major influence. However, local knowledge is often limited and insufficient to deal with tough cases that have not been previously evaluated. To solve such cases, it is necessary for researchers to search for knowledge beyond the local setting. To facilitate knowledge exchange, we have researched methods to use web-based knowledge systems to connect users having similar research interests. Such systems deal with different settings for users and with complicated information exchange procedures. Knowledge exchange can be performed by using the expertise level and tacit knowledge encoded for each expert and can be beneficial in training procedures and in solving tough cases. The contributions of our research in this area are: (1) evaluating methods of knowledge sharing in the domain of image database retrieval systems; (2) evaluating methods for tacit and explicit knowledge exchange; and (3) evaluating automatic methods

of knowledge discovery.

7.1.3 Capturing Conceptual Change in Computer-Based Applications

Decision-making processes incorporate subjective evaluations of visual patterns and their relationships to neighboring patterns. A system that uses only globally accepted settings without addressing the subjectivity of individual experts is likely to be rejected by users. Individuals may have different representations of the same concept based on their level of expertise and individual specialties. These differences are typically resolved by direct and informal peer-to-peer communication and are hard to represent by computer models. Computer-based systems should take into account the evolving, unique views image analysts have about the same phenomenon. To address these issues, we have researched a methodology to create flexible, user-specific semantic mappings around the generally accepted knowledge representation model. This method offers a technique for bridging the semantic gap that may exist between users' models of the same concepts by customizing user-specific semantic assignments of semantics into low-level features. The contributions of our research in this area are: (1) capturing an expert's conceptual change when using computer-based information systems; (2) developing fast customization methods for adjusting system settings to an analysts preferences and conceptual dynamics; (3) evaluating the level of expertise of individual analysts, and (4) evaluating methods to adapt shared knowledge using a new individuals knowledge.

7.1.4 Integrating Ontology in Retrieval Systems

Domain ontology can be used as a common base for knowledge representation and exchange because it can connect visual information to concepts stored in the knowledge base. Knowledge encapsulated in ontology can be successfully used to improve the accuracy of retrieval systems by giving extra information about domain concepts. Such synonymous information is included in an ontological hierarchy and could be used by computer algorithms. The contributions of our research in this area are: (1) researching methods of using ontology definitions to improve Retrieval Systems performance, and (2) developing models of knowledge mining using domain ontology.

7.2 Future Work

Our long-term research interest spans across all areas of computer science that would enable us to study methods of knowledge representation and exchange. These interests are grounded in the need to reduce the semantic gap between human and computer models of representing visual patterns and to develop a knowledge framework that is close to how experts reason when they make decisions. Our research plans include leveraging existing expertise in knowledge representation and exchange to tackle the following larger problems:

7.2.1 Develop computer-based models for describing domain knowledge

Experts use complex cognitive models to describe visual knowledge by connecting information from many areas. Computer-based models that map human cognitive models are desired but are difficult to create. For example, identifying the link between phenotypes and genotypes is hampered by complex mechanisms that cause the expression of complex phenotypes from underlying genetics and interaction with the environment. With recent improvement in technology, we can develop sophisticated models using quantitative trait loci (QTL) mappings into high-dimensional vectors. Using vectors instead of single values would help us better identify genes that contribute to the expression of complex phenotypes. Such an approach can be used in medical decision support. Using rich data captured at machine level, we can also create complex medical models that can be correlated with high-level semantics. Such models can be helpful in presenting the information in meaningful ways to the clinicians and help them distinguish abnormal or unique regions in HRCT images using content-based methods. The issues we want to address with this research are: (1) how to determine relevant and necessary knowledge to be represented by computer models; and (2) how to utilize computer models for efficient and accurate knowledge modeling.

7.2.2 Integrate Ontology in Knowledge Discovery and Exchange

While it is true that the use of ontology requires consensus on the ontological definitions, it can help knowledge exchange by reducing the ambiguities in communication. Ontological

representations of domain knowledge are essential for encoding domain in both bioinformatics and health informatics. In these domains, experts make inference about a new case by applying preexisting knowledge gained through learning and experience. Domain ontology is very useful for both encoding and exchanging such prior knowledge. It can be used to formalize the knowledge at both human and machine levels so it can be used by both scientists and domain applications. Great effort has been dedicated to create ontology such as Open Biomedical Ontology (OBO) or Unified Medical Language System (UMLS), and there is a need to use this knowledge in novel algorithms to further develop the related domains. The objectives of our research in this area are to: (1) how to evaluate methods of integrating knowledge from domain ontology into semantic discovery processes; (2) how to evaluate knowledge sharing using domain ontology; and (3) how to develop methods for using ontology for answering vague semantic queries.

7.2.3 Foster peer-to-peer knowledge discovery and exchange

Peer-to-peer networks have proven to be successful in tacit-knowledge elicitation and exchange by facilitating alternative opinions and revisions. These networks have the advantage of creating strong temporary connections from existing weak ones. This type of knowledge discovery is very relevant to domains such as bioinformatics and health informatics that require a great deal of cooperation between clinical and research scientists. We plan to use domain ontology and state-of-the-art computational techniques to facilitate domain knowledge discovery. For example a bioinformatics researcher may query the peer-to-peer network to discover genotype-phenotype linkage that is new to the local setting but which was researched by another scientific group. Similarly, in the medical domain, a clinical expert may request the opinion of a peer located in an outer hospital to evaluate tough cases such as "*mosaic pattern of lung attenuation*" in HRCT imagery. The objectives of our research in this area are to: (1) how to evaluate multi-layered peer-to-peer networks to facilitate tacit-knowledge exchange among local and global expert communities; and (2) how to develop local-to-local and local-to-global knowledge exchange protocols.

APPENDIX A

DETAILED CLASSIFICATION RESULTS

In this Appendix we present the detailed classification result for the 48 datasets presented previously. The quality measures used to determine the quality of classification are: (1) accuracy, (2) Kappa, (3) true positive rate (TPR), (4) false positive rate (FPR), (5) classification precision (Prec), (6) F-measure (F-meas), and (7) complete confusion matrix.

Accuracy A is defined as the percentage of correctly classified samples. Let TP be true positives, FN the false negatives, FP be the false positive, and FN be the false negative samples.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

The true positive rate (TPR) is the proportion of positive samples that were correctly identified, and it is calculated using the following formula:

$$TPR = \frac{TP}{TP + FN} \tag{27}$$

The false positive rate (FPR) is the proportion of negatives cases that were incorrectly classified as positive, and it is calculated using the following formula:

$$FPR = \frac{TN}{TN + FN} \tag{28}$$

Precision ($Prec$) is the proportion of the predicted positive cases that were correct, and it is calculated using the following formula:

$$Prec = \frac{TP}{TP + FP} \tag{29}$$

The F-measure is a weighted harmonic mean of precision and recall. We use the balanced F-measure givven by the following formula:

$$F - meas = \frac{2 * Prec}{Prec + TPR} \tag{30}$$

The Kappa coefficient K, used in these results and shown in Equation 31, indicates how the classifier assigns individual subjects into the same category on the measurement scale by comparing the agreement against that which might be expected by chance. [73] In this formula $P(A)$ is the observed agreement, while $P(E)$ is the expected agreement. Its values range from -1 corresponding to complete disagreement, and 1 corresponding to complete agreement.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{31}$$

A.1 UCI Public Datasets

Table 18: Classification result for the *UCI-Anneal* dataset

Accuracy = 98.33%
Accuracy Stdev = 1.91%
Kappa = 95.81%

CfM	1	2	3	5	U
1	7	0	1	0	0
2	1	96	2	0	0
3	4	2	678	0	0
5	0	0	0	67	0
U	0	0	5	0	35

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
87.5	0.6	58.3	70
97.0	0.3	98	97.5
99.1	3.7	98.8	98.9
100	0	100	100
87.5	0	100	93.3

Table 19: Classification result for the *UCI-Austral* dataset

Accuracy = 84.05%
Accuracy Stdev = 3.41%
Kappa = 67.76%

CfM	+	-
+	254	53
-	57	326

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
82.7	14.9	81.7	82.2
85.1	17.3	86	85.5

Table 20: Classification result for the *UCI-Auto* dataset

Accuracy=92.68%
Accuracy Stdev=5.79%
Kappa=84.39%

CfM	1	2	3	4	5	6	7
dohc	11	0	0	1	0	0	0
dohcv	0	0	0	0	0	1	0
l	0	0	11	1	0	0	0
ohc	2	0	2	142	1	0	1
ohcf	0	0	0	0	15	0	0
ohcv	3	1	0	2	0	7	0
rotor	0	0	0	0	0	0	4

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
91.7	2.6	68.8	78.6
0	0.5	0	0
91.7	1	84.6	88
95.9	7	97.3	96.6
100	0.5	93.8	96.8
53.8	0.5	87.5	66.6
100	0.5	80	88.9

Table 21: Classification result for the *UCI-Breast* dataset

Accuracy = 96.42%
Accuracy Stdev = 2.35%
Kappa = 92.09%

CfM	benign	malignant
benign	445	13
malignant	12	229

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
97.2	5	97.4	97.3
95	2.8	94.6	94.8

Table 22: Classification result for the *UCI-Cleve* dataset

Accuracy = 83.16%
Accuracy Stdev = 6.24%
Kappa = 66.08%

CfM	< 50	> 50.1
< 50	139	26
> 50.1	25	113

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
84.02	18.1	84.8	84.5
81.9	15.8	81.3	81.6

Table 23: Classification result for the *UCI-CRX* dataset

Accuracy = 84.05%
Accuracy Stdev = 3.41%
Kappa = 67.76%

CfM	+	-
+	254	53
-	57	326

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
82.7	14.9	81.7	82.2
85.1	17.3	86	85.5

Table 24: Classification result for the *UCI-Diabetes* dataset

Accuracy = 74.34%
Accuracy Stdev = 5.38%
Kappa = 41.88%

CfM	1	2
tested negative	418	82
tested positive	115	153

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
83.6	42.9	78.4	80.9
57.1	16.4	65.1	60.8

Table 25: Classification result for the *UCI-German* dataset

Accuracy = 73.4%
Accuracy Stdev = 4.45%
Kappa = 32.41%

CfM	good	bad
good	600	100
bad	166	134

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
85.7	55.3	78.3	81.8
44.7	14.3	57.3	50.2

Table 26: Classification result for the *UCI-Glass* dataset

Accuracy = 70.56%
Accuracy Stdev = 7.7%
Kappa = 59.53%

CfM	1	2	3	4	5	6
build wind float	59	10	1	0	0	0
build wind non-float	14	53	4	2	2	1
vehic wind float	8	3	6	0	0	0
containers	1	1	0	7	0	4
tableware	2	1	0	1	3	2
headlamps	1	1	1	3	0	23

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
84.3	18.1	69.4	76.1
69.7	11.6	76.8	73.1
35.3	3.0	50	41.4
53.8	3.0	53.8	53.8
33.3	1.0	60	42.8
79.3	3.8	76.7	78

Table 27: Classification result for the *UCI-Heart* dataset

Accuracy = 82.96%
Accuracy Stdev = 6.56%
Kappa = 65.61%

CfM	absent	present
absent	125	25
present	21	99

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
83.3	17.5	85.6	84.4
82.5	16.7	79.8	81.1

Table 28: Classification result for the *UCI-Hepatitis* dataset

Accuracy = 84.52%
Accuracy Stdev = 8.38%
Kappa = 46.56%

CfM	DIE	LIVE
DIE	15	17
LIVE	7	116

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
46.9	5.7	68.2	55.6
94.3	53.1	87.2	90.6

Table 29: Classification result for the *UCI-Horse* dataset

Accuracy = 82.06%
Accuracy Stdev = 5.81%
Kappa = 60.30%

CfM	yes	no
yes	209	23
no	43	93

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
90.1	31.6	82.9	86.4
68.4	9.9	80.2	73.8

Table 30: Classification result for the *UCI-Hypo* dataset

Accuracy = 96.89%
Accuracy StDev = 0.73%
Kappa = 75.43%

CfM	1	2	3	4
negative	3472	5	3	1
compensated hypothyroid	72	110	12	0
primary hypothyroid	16	6	73	0
secondary hypothyroid	2	0	0	0

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
99.7	30.9	97.5	98.6
56.7	0.3	90.9	69.8
76.8	0.4	83	79.8
0	0.03	0	0

Table 31: Classification result for the *UCI-Ionosphere* dataset

Accuracy = 91.73%
Accuracy Stdev = 4.91%
Kappa = 81.63%

CfM	b	g
b	105	21
g	8	217

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
83.3	3.6	92.9	87.8
96.4	16.7	91.2	93.7

Table 32: Classification result for the *UCI-Iris* dataset

Accuracy = 94.66%
Accuracy Stdev = 7.56%
Kappa = 92%

CfM	1	2	3
Iris-setosa	50	0	0
Iris-versicolor	0	45	5
Iris-virginica	0	3	47

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
100	0	100	100
90	3	93.8	91.9
94	5	90.4	92.2

Table 33: Classification result for the *UCI-Labor* dataset

Accuracy = 89.47%
Accuracy Stdev = 11.65%
Kappa = 76.89%

CfM	bad	good
bad	17	3
good	3	34

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
85	8.1	85	85
91.9	15	91.9	91.9

Table 34: Classification result for the *UCI-Led7* dataset

Accuracy = 73.47%
Accuracy StDev = 2.13%
Kappa = 70.49%

CfM	0	1	2	3	4	5	6	7	8	9
0	276	3	7	3	6	1	4	9	15	1
1	4	269	0	1	33	0	0	26	0	0
2	4	3	286	4	3	3	6	3	4	3
3	3	3	45	147	4	19	1	27	2	19
4	4	32	3	2	268	7	2	7	6	5
5	6	0	1	0	6	258	29	6	6	23
6	20	3	7	1	1	38	244	2	22	3
7	5	25	5	5	2	5	1	255	0	1
8	28	0	47	3	10	4	14	2	202	17
9	21	2	11	23	43	27	2	8	27	146

Class	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
0	84.9	3.3	74.4	79.3
1	80.8	2.5	79.1	79.9
2	89.7	4.4	69.4	78.3
3	54.4	1.4	77.8	64
4	79.8	3.8	71.3	75.3
5	77	3.6	71.3	74
6	71.6	2.1	80.5	75.8
7	83.9	3.1	73.9	78.6
8	61.8	2.9	71.1	66.1
9	47.1	2.5	67	55.3

Table 35: Classification result for the *UCI-Lymph* dataset

Accuracy = 82.43%
Accuracy Stdev = 11.38%
Kappa = 67%

CfM	1	2	3	4
normal	1	0	1	0
metastases	2	70	7	2
malign lymph	1	10	50	0
fibrosis	0	1	2	1

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
50	2.1	25	33.3
86.4	16.4	86.4	86.4
82	11.5	83.3	82.6
25	1.4	33.3	28.6

Table 36: Classification result for the *UCI-Pima* dataset

Accuracy = 75.39%
Accuracy Stdev = 4.82%
Kappa = 43.03%

CfM	41	42
41	434	66
42	123	145

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
86.8	45.9	77.9	82.1
54.1	13.2	68.7	60.5

Table 37: Classification result for the *UCI-Sick* dataset

Accuracy = 97.53%
Accuracy Stdev = 0.64%
Kappa = 76.61%

CfM	negative	sick
negative	3515	26
sick	67	164

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
99.3	29	98.1	98.7
71	0.7	86.3	77.9

Table 38: Classification result for the *UCI-Sonar* dataset

Accuracy = 79.32%
Accuracy Stdev = 7.98%
Kappa = 58.11%

CfM	1	2
Rock	69	28
Mine	15	96

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
71.1	13.5	82.1	76.2
86.5	28.9	77.4	81.4

Table 39: Classification result for the *UCI-TicTac* dataset

Accuracy = 98.95%
Accuracy Stdev = 1.55%
Kappa = 97.7%

CfM	negative	positive
negative	329	3
positive	7	619

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
99.1	1.8	97.9	98.5
98.9	0.9	99.5	99.2

Table 40: Classification result for the *UCI-Vehicle* dataset

Accuracy = 67.37%
Accuracy Stdev = 4.1%
Kappa = 56.50%

CfM	1	2	3	4
opel	129	50	19	14
saab	116	65	22	14
bus	2	0	212	4
van	30	0	5	164

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
60.8	23.3	46.6	0.528
30	7.9	56.5	39.2
97.2	7.3	82.2	89.1
82.4	4.9	83.7	83

Table 41: Classification result for the *UCI-Waveform* dataset

- Accuracy = 82.96%
- Kappa = 74.45%
- Accuracy Stdev=0.84%

CfM	0	1	2
0	1263	206	223
1	124	1425	104
2	91	104	1460

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
74.6	6.5	85.5	79.7
86.2	9.3	82.1	84.1
88.2	9.8	81.7	84.8

Table 42: Classification result for the *UCI-Wine* dataset

Accuracy = 96.06%
Accuracy Stdev = 5.88%
Kappa = 94.05%

CfM	1	2	3
1	57	2	0
2	1	66	4
3	0	0	48

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
96.6	0.8	98.3	97.4
93	1.9	97.1	95
1.0	3.1	92.3	96

Table 43: Classification result for the *UCI-Zoo* dataset

Accuracy = 95.04%
Accuracy Stdev = 5.18%
Kappa = 93.43%

CfM	1	2	3	4	5	6	7
mammal	41	0	0	0	0	0	0
bird	0	20	0	0	0	0	0
reptile	0	1	2	1	1	0	0
fish	0	0	0	13	0	0	0
amphibian	0	0	0	0	4	0	0
insect	0	0	0	0	0	8	0
invertebrate	1	0	1	0	0	0	8

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1.0	0.017	0.976	0.988
1.0	0.012	0.952	0.975
0.4	0.01	0.667	0.5
1.0	0.011	0.929	0.963
1.0	0.01	0.8	0.889
1.0	0.0	1.0	1.0
0.8	0.0	1.0	0.889

A.2 Classification Results for Maize Datasets

Table 44: Classification result for the *Maize-2* dataset

Accuracy = 100%
Accuracy Stdev = 0%
Kappa = 100%

CfM	les1	les13
les1	30	0
les13	0	14

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1.0	0.0	1.0	1.0
1.0	0.0	1.0	1.0

Table 45: Classification result for the *Maize-3* dataset

Accuracy = 98.3%
Accuracy Stdev = 5.27%
Kappa = 97.4%

CfM	les1	les7	les13
les1	30	0	0
les7	0	18	0
les13	1	0	13

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
100	3.1	96.8	98.4
100	0	100	100
92.9	0	100	96.3

Table 46: Classification result for the *Maize-4* dataset

Accuracy = 100%
Accuracy Stdev = 0%
Kappa=100%

CfM	les1	les8	les7	les13
les1	30	0	0	0
les8	0	17	0	0
les7	0	0	18	0
les13	0	0	0	14

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
100	0	100	100
100	0	100	100
100	0	100	100
100	0	100	100

Table 47: Classification result for the *Maize-5* dataset

Accuracy = 94.3%
Accuracy Stdev = 10%
Kappa = 92.7%

CfM	les1	les8	les7	les13	les19
les1	29	1	0	0	0
les8	0	15	0	0	2
les7	0	0	17	1	0
les13	0	0	0	14	0
les19	0	1	0	0	9

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
96.7	0.0	100	98.3
88.2	2.8	88.2	88.2
94.4	0	100	97.1
100	1.3	93.3	96.5
90	2.5	81.1	85.7

Table 48: Classification result for the *Maize-6* dataset

Accuracy = 95.3%
Accuracy Stdev = 7.7%
Kappa=94.3%

CfM	1	2	3	4	5	6
les1	30	0	0	0	0	0
les8	0	14	0	0	0	3
les6	0	0	19	0	0	0
les7	0	0	0	18	0	0
les13	0	0	0	0	14	0
les19	0	2	0	0	0	8

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
100	0	100	100
82.4	2.2	87.5	84.9
100	0	100	100
100	0	100	100
100	0	100	100
80	3.1	72.7	76.2

Table 49: Classification result for the *Maize-7* dataset

Accuracy = 89.16%
Accuracy Stdev=10.43%
Kappa = 87.09%

CfM	1	2	3	4	5	6	7
les1	28	0	0	0	0	2	0
les6	0	18	1	0	0	0	0
les7	0	0	18	0	0	0	0
les8	0	0	0	15	0	1	1
les13	0	0	1	0	13	0	0
les18	1	0	0	1	0	9	1
les19	0	0	0	2	0	2	6

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
93.3	1.1	96.6	94.9
94.7	0	100	97.3
100	2	90	94.7
88.2	2.9	83.3	85.7
92.9	0	100	96.3
75	4.6	64.3	69.2
60	1.8	75	66.7

Table 50: Classification result for the *Maize-8* dataset

Accuracy=92.59%
Accuracy Stdev = 4.9%
Kappa = 91.4%

CfM	1	2	3	4	5	6	7	8	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
les1	28	0	0	0	0	0	2	0	93.3	0	100	96.5
les6	0	19	0	0	0	0	0	0	100	0.9	95	97.4
les7	0	0	17	0	1	0	0	0	94.4	0	100	97.1
les8	0	0	0	15	0	0	0	2	88.2	1.7	88.2	88.2
les13	0	0	0	0	14	0	0	0	100	1.7	87.5	93.3
les17	0	1	0	0	0	14	0	0	93.3	0	100	96.5
les18	0	0	0	2	1	0	9	0	75	2.4	75	75
les19	0	0	0	0	0	0	1	9	90	1.6	81.1	85.7

A.3 Classification Results for Geospatial Datasets

Table 51: Classification result for the *Geospatial-2* dataset

Accuracy = 97.48%
Accuracy Stdev = 3.12%
Kappa = 94.91%

CfM	GRSLD	DEVH
GRSLD	122	2
DEVH	5	149

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
98.4	3.2	96.1	97.2
96.8	1.6	98.6	97.7

Table 52: Classification result for the *Geospatial-3* dataset

Accuracy = 94.07%
Accuracy Stdev = 3.7%
Kappa = 91.06%

CfM	INDS	GRSLD	DEVH
INDS	121	3	4
GRSLD	7	113	3
DEVH	6	1	147

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
94.5	4.7	90.3	92.4
91.9	1.4	96.6	94.2
95.5	2.8	95.5	95.5

Table 53: Classification result for the *Geospatial-4* dataset

Accuracy = 89.56%
Accuracy Stdev = 3.67%
Kappa = 86.04%

CfM	1	2	3	4
INDS	122	1	1	5
GRSLD	7	104	10	3
FRST	4	7	107	3
DEVH	10	2	2	139

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
94.6	5.3	85.3	89.7
83.9	2.5	91.2	87.4
88.4	3.2	89.2	88.8
90.8	2.9	92.7	91.7

Table 54: Classification result for the *Geospatial-5* dataset

Accuracy = 80.82%
Accuracy Stdev = 5.66%
Kappa = 75.97%

CfM	1	2	3	4	5
INDS	96	1	1	1	30
GRSLD	7	101	9	3	4
FRST	3	6	108	4	1
DEVH	3	0	3	142	6
CTRB	31	3	2	7	80

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
74.4	8.4	68.6	71.4
81.5	1.9	91	86
88.5	2.8	87.8	88.1
92.2	3	90.4	91.3
65	7.8	66.1	65.5

Table 55: Classification result for the *Geospatial-6* dataset

Accuracy = 75.7%
Accuracy Stdev = 2.97%
Kappa = 70.78%

CfM	1	2	3	4	5	6
INDS	89	3	0	1	6	30
GRSLD	4	86	10	17	3	4
FRST	3	5	105	3	4	2
FRM	1	24	3	104	0	2
DEVH	2	0	6	0	142	4
CTRB	33	5	2	3	11	69

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
69	6.5	67.4	68.2
69.4	5.6	69.9	69.6
86.1	3.2	83.3	84.7
77.6	3.7	81.3	79.4
92.2	3.8	85.5	88.7
56.1	6.3	62.2	59

Table 56: Classification result for the *Geospatial-7* dataset

Accuracy = 71.63%
Accuracy Stdev = 7.33%
Kappa = 66.86%

CfM	1	2	3	4	5	6	7
CONST	90	15	3	2	2	14	9
INDS	6	94	0	0	0	0	28
GRSLD	12	3	73	10	23	0	4
FRST	1	0	10	103	5	1	0
FRM	5	0	19	4	102	0	3
DEVH	21	5	1	2	0	124	3
CTRB	9	27	2	1	3	7	73

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
66.7	6.9	62.5	64.5
72.9	6.3	65.3	68.9
58.4	4.4	67.6	62.7
85.8	2.43	84.4	85.1
76.7	4.2	75.6	76.1
79.5	3	84.4	81.9
59.8	5.9	60.8	60.3

Table 57: Classification result for the *Geospatial-8* dataset

Accuracy = 62.31%
Accuracy Stdev = 3.3%
Kappa = 56.9%

CfM	1	2	3	4	5	6	7	8
CONST	77	0	14	2	0	4	11	8
RD	6	48	0	24	36	10	2	2
INDS	6	0	91	0	0	1	4	28
GRSLD	19	35	1	53	8	11	5	3
FRST	4	18	0	3	101	5	2	1
FRM	12	14	0	14	5	77	1	4
DEVH	9	0	4	0	2	0	124	3
CTRB	14	2	24	0.21	5.4	60.2	60.2	

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
66.4	7.6	52.4	58.5
37.5	7.6	41	39.2
70	4.8	67.9	68.9
39.3	4.8	55.2	45.9
75.4	5.8	66	70.4
60.6	3.6	70	65
87.3	3.5	80.0	83.5

A.4 Classification Results for Medical Datasets

Table 58: Classification result for the *HRCT-2* dataset

Accuracy = 96.37%
Accuracy StDef = 2.76%
Kappa = 92.15%

CfM	5	7
5	209	13
7	9	375

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
94.1	2.3	95.9	95
97.7	5.9	96.6	97.1

Table 59: Classification result for the *HRCT-4* dataset

Accuracy = 91.35%
Accuracy Stdev = 2.24%
Kappa = 84.41%

CfM	3	5	7	8
3	44	11	5	0
5	3	208	11	0
7	4	9	369	2
8	0	0	14	2

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
73.3	1.1	86.3	79.3
93/7	4/3	91.2	92.4
96.1	10.1	92.5	94.3
12.5	0.3	50	20

Table 60: Classification result for the *HRCT-6* dataset

Accuracy = 89.31%
Accuracy Stdev = 2.57%
Kappa = 83.22%

CfM	1	3	5	7	8	14
1	38	2	2	2	0	0
3	2	50	5	2	0	1
5	3	3	201	13	0	2
7	6	2	6	365	0	5
8	0	0	8	8	8	0
14	1	0	1	12	3	15

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
86.4	1.7	76	80.9
83.3	1.0	87.7	85.4
90.5	2.6	93.5	92
95.1	9.9	90.8	92.9
50	0.4	72.7	59.3
46.9	1.1	65.2	54.6

Table 61: Classification result for the *HRCT-8* dataset

Accuracy = 84.76%
Accuracy Stdev = 3.26%
Kappa = 78.02%

CfM	1	3	5	7	8	14	19	21
1	37	2	4	0	0	0	0	1
3	3	44	10	3	0	0	0	0
5	1	5	206	8	0	1	1	0
7	3	4	12	348	2	5	1	9
8	0	0	0	5	9	1	0	1
14	0	0	2	21	1	7	0	1
19	2	0	1	1	0	0	8	2
21	0	0	1	13	0	1	0	53

TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
84.1	1.1	80.4	82.2
73.3	1.4	80	76.5
92.8	4.9	87.3	90
90.6	11.2	87.2	88.9
56.3	0.4	75	64.3
21.9	1	46.7	29.8
57.1	0.2	80	66.67
77.9	1.8	79.1	78.5

Table 62: Classification result for the *HRCT-10* dataset

Accuracy = 79.02%
Accuracy Stdev = 3.47%
Kappa = 72.26%

CfM	1	3	5	7	8	10	14	19	21	22
1	41	2	2	1	0	0	0	0	0	0
3	8	43	11	3	0	0	0	0	0	0
5	11	4	199	5	0	6	3	0	1	1
7	16	5	6	330	2	14	5	0	15	6
8	2	0	0	6	10	0	0	0	0	0
10	14	0	23	13	0	94	0	0	2	1
14	3	0	1	19	2	3	5	0	1	1
19	1	0	2	5	0	1	0	4	0	1
21	2	0	0	18	1	7	0	0	42	0
22	2	0	0	3	0	1	0	0	0	23

Semantic	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1	88.6	0.9	81.3	84.8
3	71.7	1.2	79.6	75.4
5	89.6	5.8	81.6	85.4
7	85.9	11.8	81.9	83.9
8	62.5	0.5	66.7	64.5
10	70.1	3.7	74.6	72.3
14	15.6	0.8	38.5	22.2
19	28.6	0	100	44.5
21	61.8	2	68.9	65.2
22	85.2	1	69.7	76.7

Table 63: Classification result for the *HRCT-12* dataset

Accuracy = 76.85%
Accuracy Stdev = 4.38%
Kappa = 70.78%

CfM	1	3	4	5	7	8	10	14	19	20	21	22
1	39	1	0	2	1	0	0	0	1	0	0	0
3	3	43	0	11	3	0	0	0	0	0	0	0
4	0	0	29	1	14	0	5	0	0	0	6	2
5	2	7	1	195	3	0	8	2	3	1	0	0
7	4	2	9	6	334	4	13	3	0	3	2	4
8	0	0	1	0	8	6	0	1	0	0	0	0
10	3	0	0	12	11	0	101	5	0	0	2	0
14	1	0	0	3	17	1	3	6	0	0	1	0
19	0	0	0	3	3	0	0	0	6	0	1	1
20	0	0	0	1	7	0	2	0	0	25	0	0
21	0	0	3	3	16	0	7	0	0	0	38	1
22	0	0	0	0	7	0	1	0	0	1	0	18

Semantic	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1	88.6	1.2	75	81.2
3	71.7	1	81.1	76.1
4	50.9	1.4	67.4	58
5	87.8	4.8	82.3	85
7	87	12.7	78.8	82.7
8	37.5	0.5	54.5	44.4
10	75.4	4.1	72.1	73.7
14	18.8	1	35.3	24.5
19	42.9	0.4	60	50
20	71.4	0.5	83.3	76.9
21	55.9	1.2	76	64.4
22	66.7	0.8	69.2	67.9

Table 64: Classification result for the *HRCT-14* dataset

Accuracy = 72.45%
Accuracy Stdev = 4.62%
Kappa = 65.86%

CfM	1	3	4	5	7	8	10	11	14	16	19	20	21	22
1	35	3	0	2	1	0	0	2	0	0	1	0	0	0
3	0	43	0	11	4	0	0	1	0	0	1	0	0	0
4	0	1	20	3	13	0	7	0	0	0	0	0	12	1
5	3	3	0	196	6	0	8	1	1	1	0	3	0	0
7	5	3	6	5	321	5	11	6	2	0	0	8	10	2
8	0	0	0	0	6	9	0	0	1	0	0	0	0	0
10	0	3	0	19	10	1	94	3	0	0	0	0	4	0
11	2	0	0	0	4	0	2	43	0	0	0	0	0	0
14	0	1	0	2	24	1	4	0	0	0	0	0	0	0
16	0	0	0	2	10	0	0	1	0	1	0	0	0	0
19	0	1	0	4	2	0	3	0	0	0	4	0	0	0
20	2	1	0	4	3	0	0	0	0	0	0	25	0	0
21	0	0	1	3	28	0	7	0	0	0	0	0	29	0
22	0	0	0	1	7	0	0	0	0	0	0	0	0	19

Semantic	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1	79.5	1.1	74.5	76.9
3	71.7	1.5	72.9	72.3
4	35.1	0.6	74.1	47.6
5	88.3	6	77.8	82.7
7	83.6	15.2	73.1	78
8	56.3	0.6	56.3	56.3
10	70.1	4.1	69.1	69.6
11	84.3	1.3	75.4	79.6
14	0	0.4	0	0
16	7.1	0.09	50	12.4
19	28.6	0.2	66.7	40
20	71.4	1	69.4	70.4
21	42.6	2.4	52.7	47.1
22	70.4	0.3	86.4	77.6

Table 65: Classification result for the *HRCT-16* dataset

Accuracy = 73.58%
Accuracy Stdev = 2.13%
Kappa = 66.14%

CfM	1	2	3	4	5	7	8	9	10	11	14	16	19	20	21	22
1	18	21	1	0	1	0	0	0	0	3	0	0	0	0	0	0
2	8	571	7	0	25	16	0	0	5	6	0	0	0	4	4	0
3	1	28	22	0	6	3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	18	1	26	0	0	4	1	0	0	0	0	7	0
5	1	27	2	0	179	4	0	0	4	2	0	1	0	2	0	0
7	1	10	1	7	9	328	2	0	9	1	0	0	0	7	4	5
8	0	0	0	2	0	3	10	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0
10	0	13	0	0	15	18	0	0	84	1	0	0	0	2	1	0
11	1	12	0	0	2	3	0	0	1	32	0	0	0	0	0	0
14	0	0	0	1	1	26	2	0	1	1	0	0	0	0	0	0
16	0	0	1	0	2	9	0	0	0	0	0	2	0	0	0	0
19	0	6	0	0	3	4	0	0	0	0	0	0	1	0	0	0
20	1	13	0	0	3	5	0	0	0	0	0	0	0	13	0	0
21	0	3	0	3	1	20	0	0	5	0	0	2	0	0	34	0
22	0	1	0	0	0	12	0	0	1	0	0	0	0	0	0	13

Semantic	TPR(%)	FPR(%)	Prec(%)	F-Meas(%)
1	40.9	0.7	58.1	48
2	88.4	11.5	81	84.5
3	36.7	0.7	64.7	46.8
4	31.6	0.7	58.1	40.9
5	80.6	4.3	72.2	76.2
7	85.4	10.4	68.8	76.2
8	62.5	0.2	71.4	66.7
9	100	0	100	100
10	62.7	1.8	73.7	67.8
11	62.7	0.9	68.1	65.3
14	0	0	0	0
16	14.3	0.2	40	21.1
19	7.1	0	100	13.3
20	37.1	0.8	46.4	41.2
21	50	1	66.7	57.2
22	48.1	0.3	72.2	57.7

APPENDIX B

DETAILED RANKING RESULTS

B.1 UCI Dataset

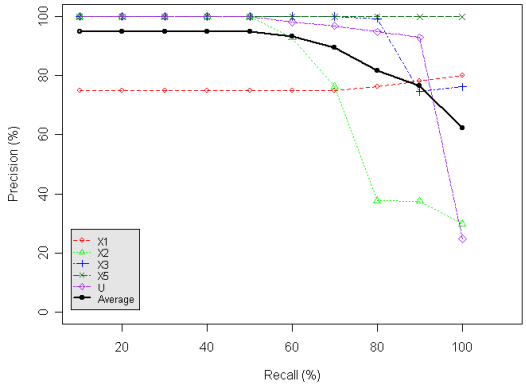


Figure 45: Ranking results and overall performance for *UCI-Anneal* dataset.

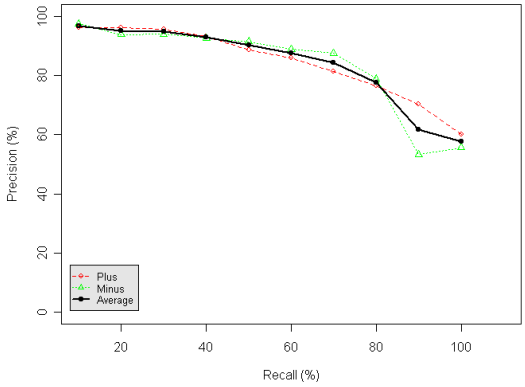


Figure 46: Ranking results and overall performance for *UCI-Austral* dataset.

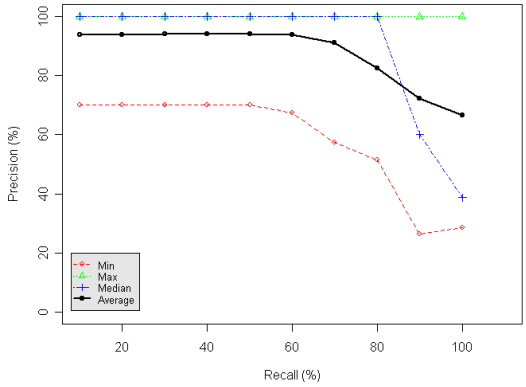


Figure 47: Ranking results and overall performance for *UCI-Autos* dataset.

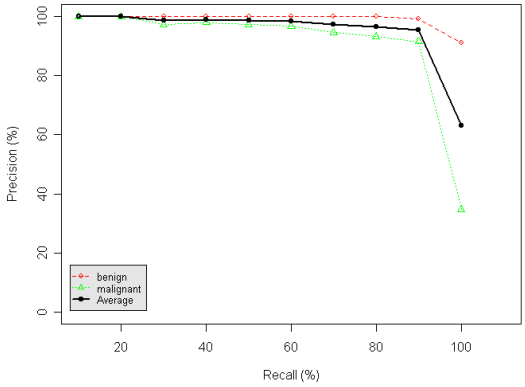


Figure 48: Ranking results and overall performance for *UCI-Breast* dataset.

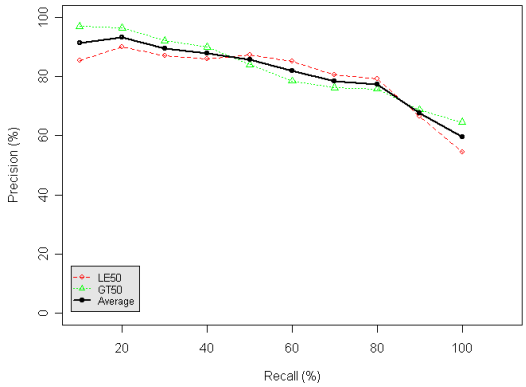


Figure 49: Ranking results and overall performance for *UCI-Cleve* dataset.

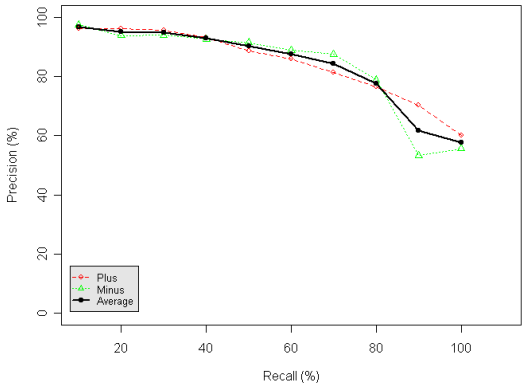


Figure 50: Ranking results and overall performance for *UCI-CRX* dataset.

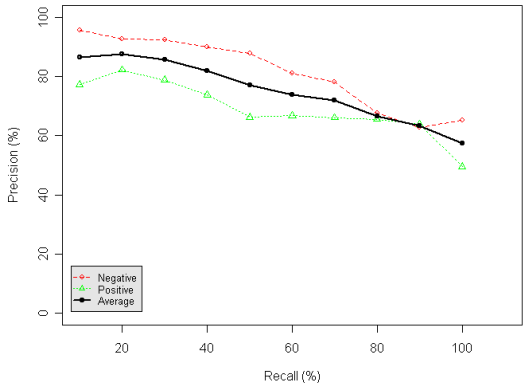


Figure 51: Ranking results and overall performance for *UCI-Diabetes* dataset.

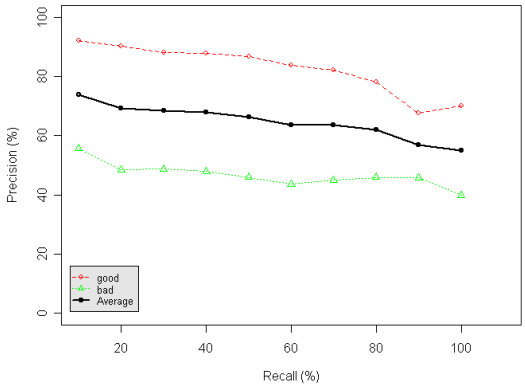


Figure 52: Ranking results and overall performance for *UCI-German* dataset.

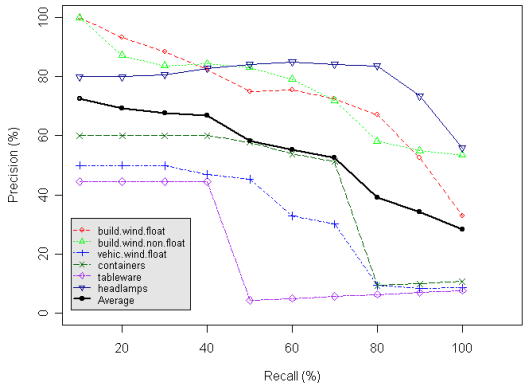


Figure 53: Ranking results and overall performance for *UCI-Glass* dataset.

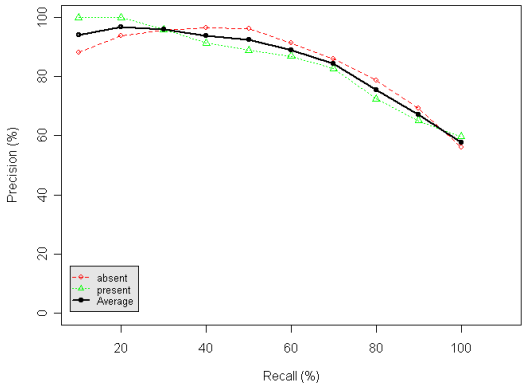


Figure 54: Ranking results and overall performance for *UCI-Heart* dataset.

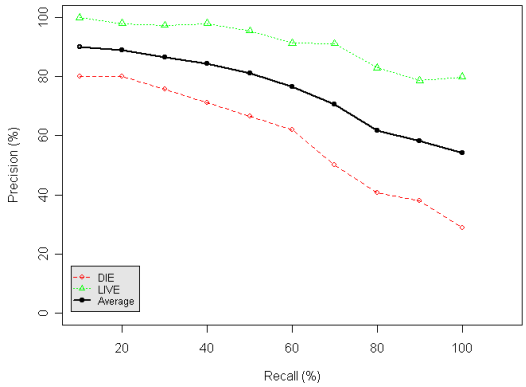


Figure 55: Ranking results and overall performance for *UCI-Hepatitis* dataset.

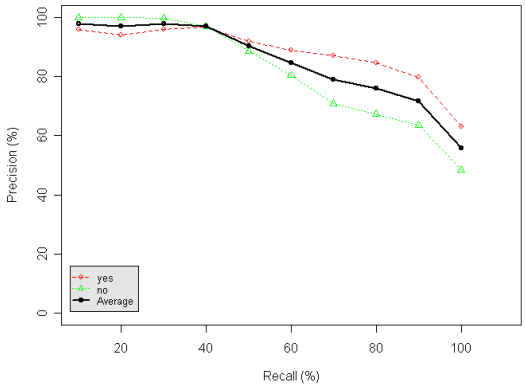


Figure 56: Ranking results and overall performance for *UCI-Horse* dataset.

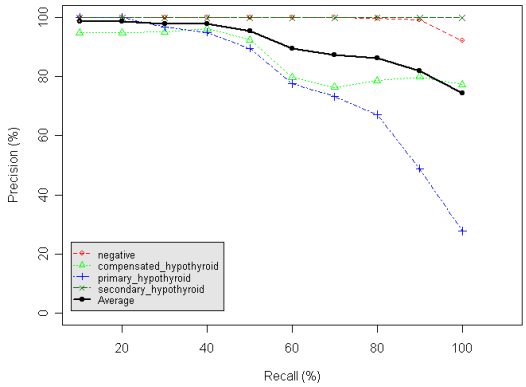


Figure 57: Ranking results and overall performance for *UCI-Hypo* dataset.

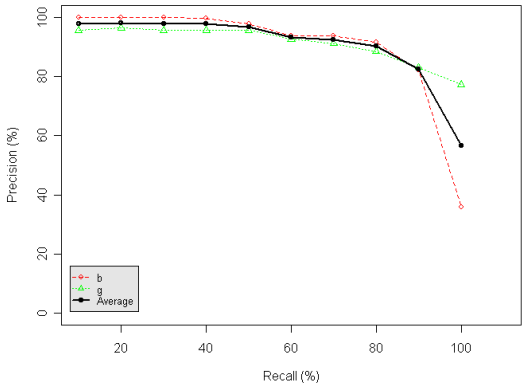


Figure 58: Ranking results and overall performance for *UCI-Ionosphere* dataset.

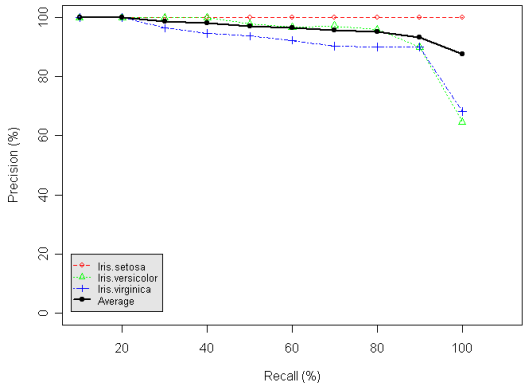


Figure 59: Ranking results and overall performance for *UCI-Iris* dataset.

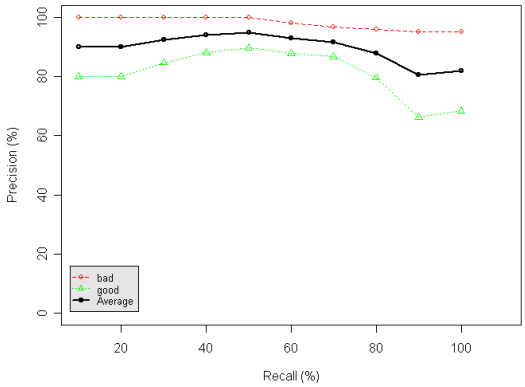


Figure 60: Ranking results and overall performance for *UCI-Labor* dataset.

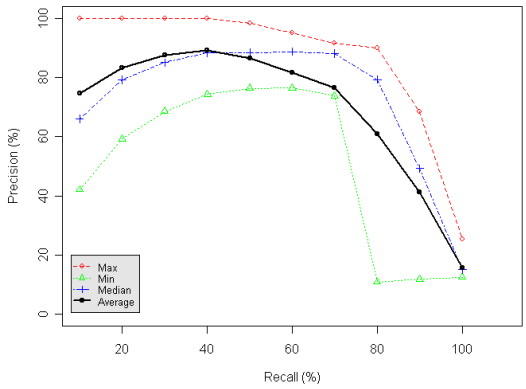


Figure 61: Ranking results and overall performance for *UCI-Led7* dataset.

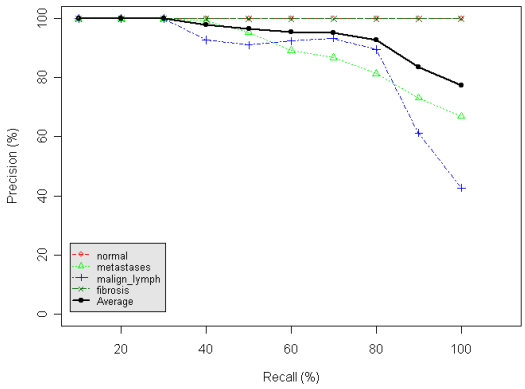


Figure 62: Ranking results and overall performance for *UCI-Lymph* dataset.

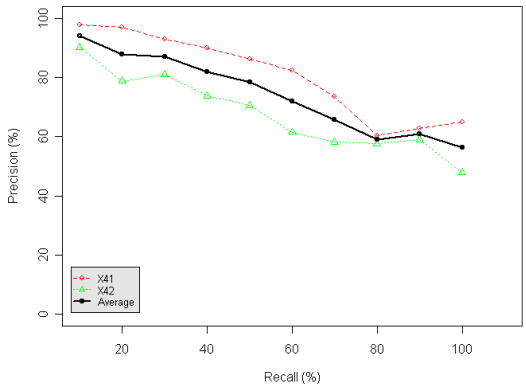


Figure 63: Ranking results and overall performance for *UCI-Pima* dataset.

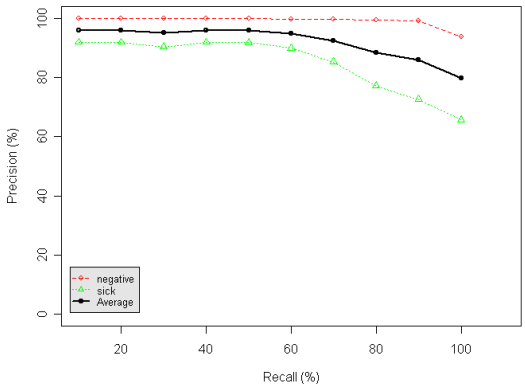


Figure 64: Ranking results and overall performance for *UCI-Sick* dataset.

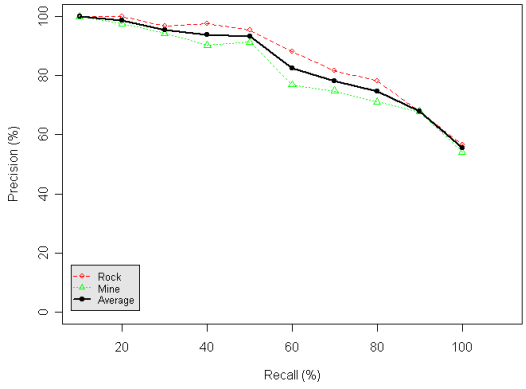


Figure 65: Ranking results and overall performance for *UCI-Sonar* dataset.

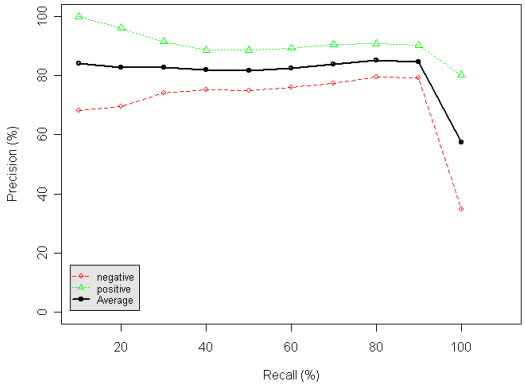


Figure 66: Ranking results and overall performance for *UCI-TicTac* dataset.

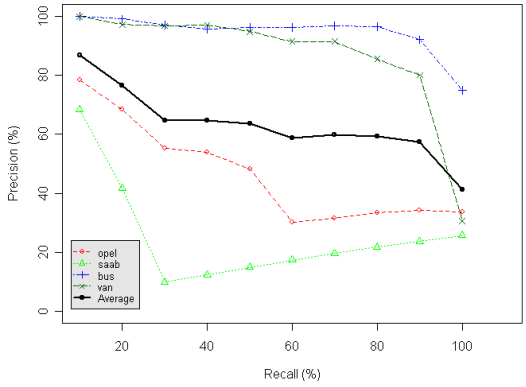


Figure 67: Ranking results and overall performance for *UCI-Vehicle* dataset.

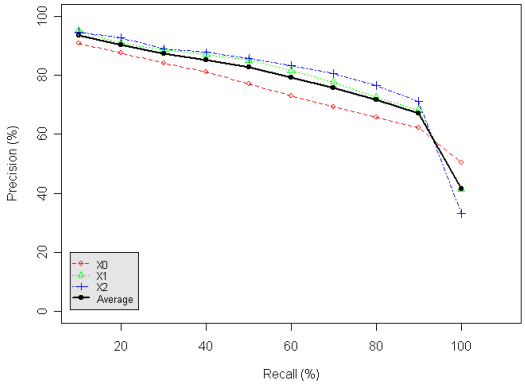


Figure 68: Ranking results and overall performance for *UCI-Waveform* dataset.

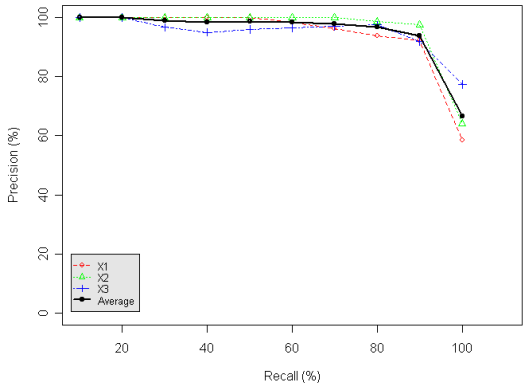


Figure 69: Ranking results and overall performance for *UCI-Wine* dataset.

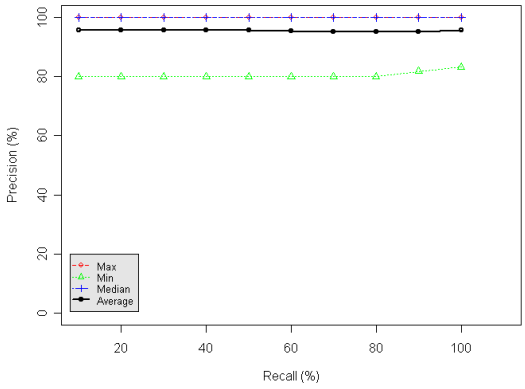


Figure 70: Ranking results and overall performance for *UCI-Zoo* dataset.

B.2 Maize Dataset

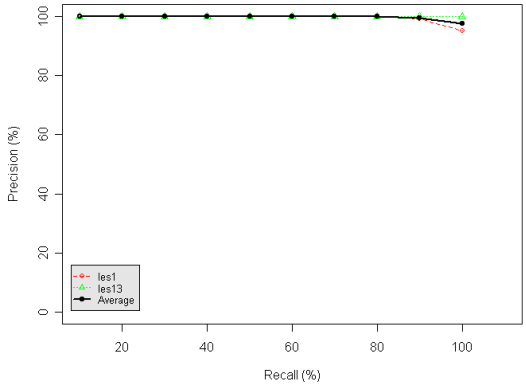


Figure 71: Ranking results and overall performance for *Maize-2* dataset.

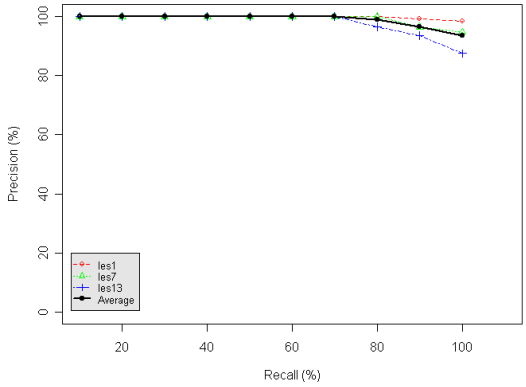


Figure 72: Ranking results and overall performance for *Maize-3* dataset.

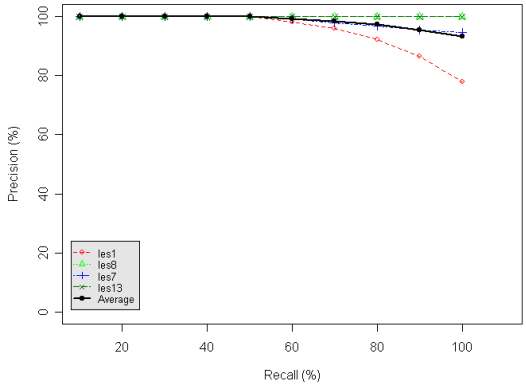


Figure 73: Ranking results and overall performance for *Maize-4* dataset.

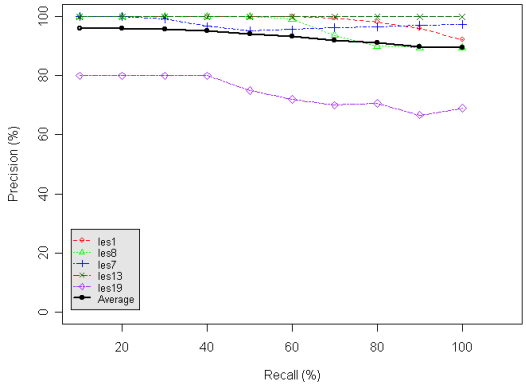


Figure 74: Ranking results and overall performance for *Maize-5* dataset.

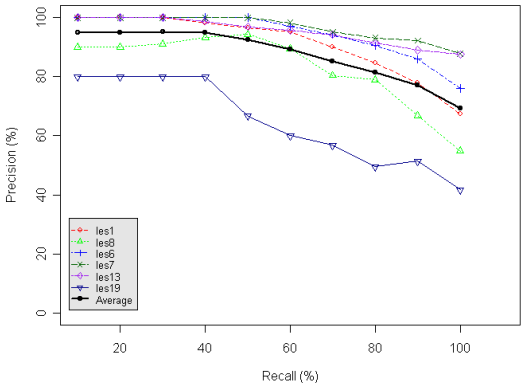


Figure 75: Ranking results and overall performance for *Maize-6* dataset.

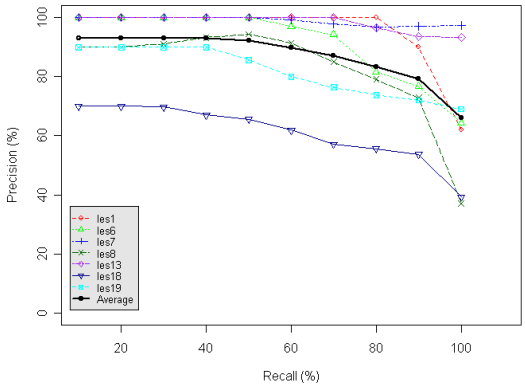


Figure 76: Ranking results and overall performance for *Maize-7* dataset.

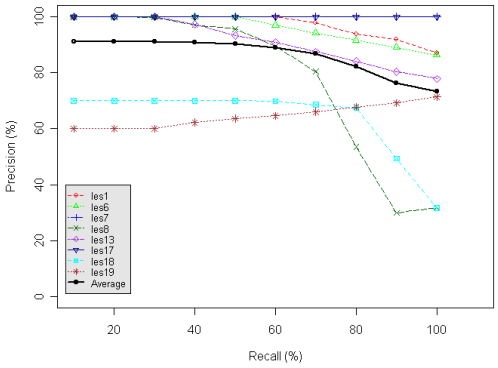


Figure 77: Ranking results and overall performance for *Maize-8* dataset.

B.3 Geospatial Datasets

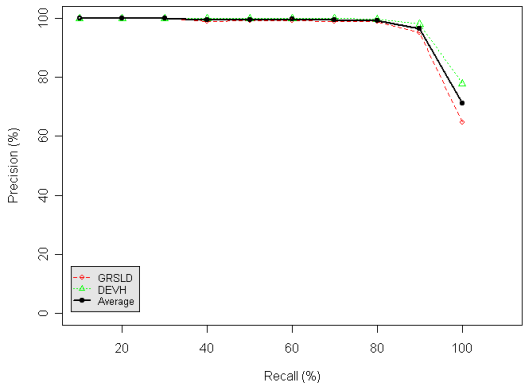


Figure 78: Ranking results and overall performance for *Geospatial-2* dataset.

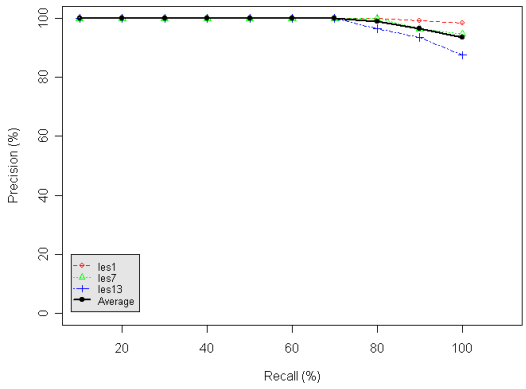


Figure 79: Ranking results and overall performance for *Geospatial-3* dataset.

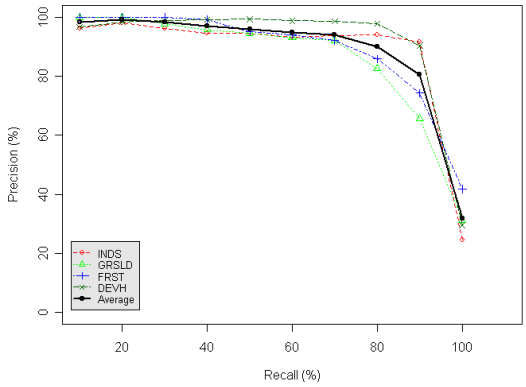


Figure 80: Ranking results and overall performance for *Geospatial-4* dataset.

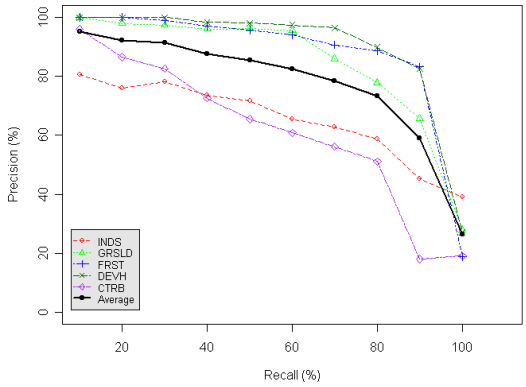


Figure 81: Ranking results and overall performance for *Geospatial-5* dataset.

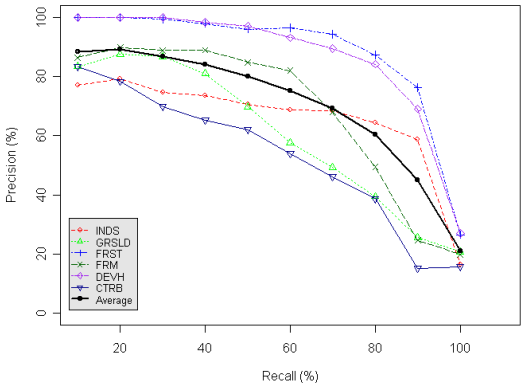


Figure 82: Ranking results and overall performance for *Geospatial-6* dataset.

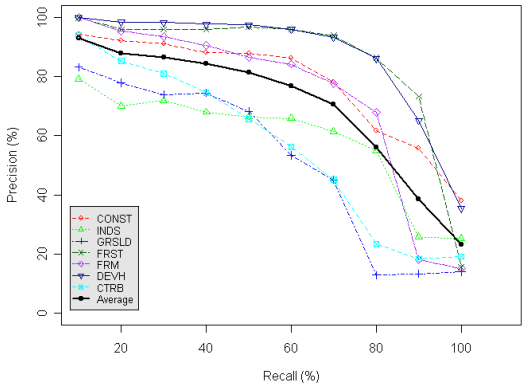


Figure 83: Ranking results and overall performance for *Geospatial-7* dataset.

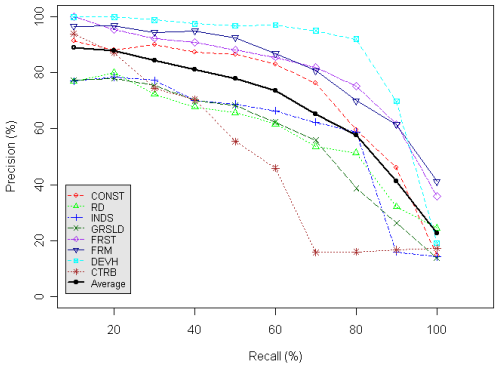


Figure 84: Ranking results and overall performance for *Geospatial-8* dataset.

B.4 Medical Datasets

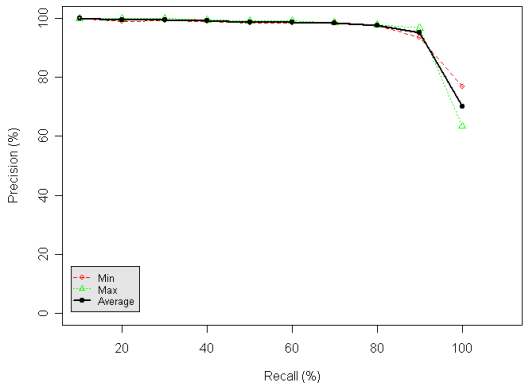


Figure 85: Ranking results and overall performance for *HRCT-2* dataset.

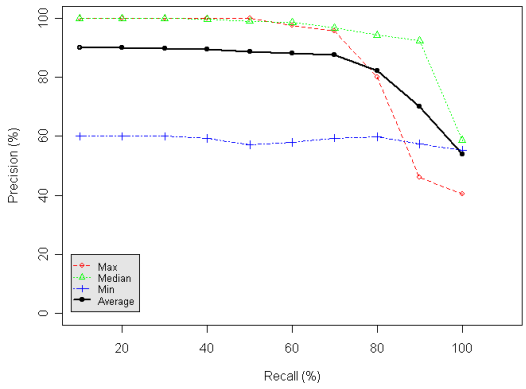


Figure 86: Ranking results and overall performance for *HRCT-4* dataset.

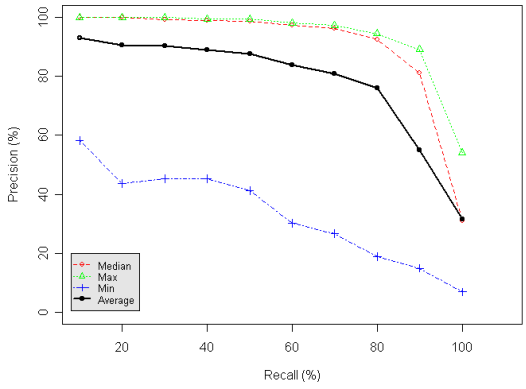


Figure 87: Ranking results and overall performance for *HRCT-6* dataset.

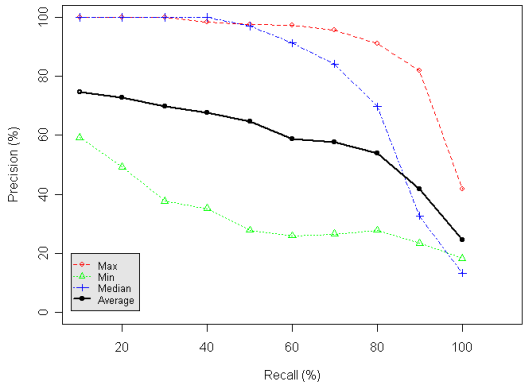


Figure 88: Ranking results and overall performance for *HRCT-8* dataset.

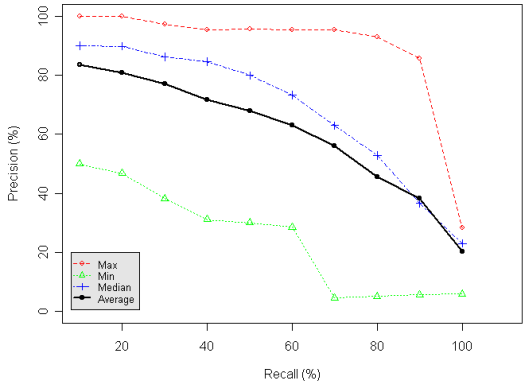


Figure 89: Ranking results and overall performance for *HRCT-10* dataset.

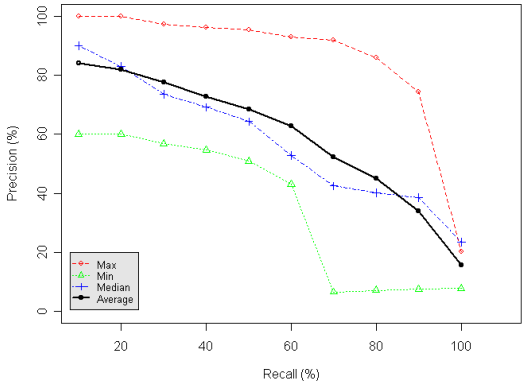


Figure 90: Ranking results and overall performance for *HRCT-12* dataset.

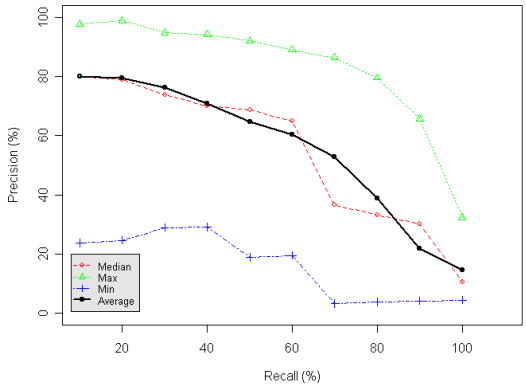


Figure 91: Ranking results and overall performance for *HRCT-14* dataset.

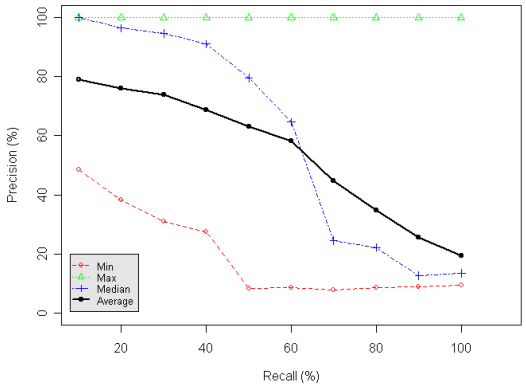


Figure 92: Ranking results and overall performance for *HRCT-16* dataset.

REFERENCES

[1] Ackoff, R. L., "From data to wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3–9, 1989.

[2] Agrawal, R., Imieliński, T., and Swami, A., "Mining association rules between sets of items in large databases," in *SIGMOD '93: Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216, 1993.

[3] Aguilera, A., Suber, A., Martinez, L., Subero, A., and Tineo, L., "Fuzzy image retrieval system," in *Fuzzy Systems, The 10th IEEE International Conference on*, vol. 3, pp. 1247–1250, 2001.

[4] Alber, J. and Niedermeier, R., "On multi-dimensional hilbert indexings," in *Computing and Combinatorics*, (London, UK), pp. 329–338, Springer-Verlag, 1998.

[5] Aref, W. G. and Kamel, I., "On multi-dimensional sorting orders," in *DEXA '00: Proceedings of the 11th International Conference on Database and Expert Systems Applications*, (London, UK), pp. 774–783, Springer-Verlag, 2000.

[6] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., "Gene ontology: tool for the unification of biology. the gene ontology consortium.," *Nat Genet*, vol. 25, pp. 25–29, May 2000.

[7] Asuncion, A. and Newman, D., "UCI machine learning repository," 2007.

[8] Awad, E. and Ghaziri, H., *Knowledge Management*. Pearson Education International, 2004.

[9] Baker, P. G., Goble, C. A., Bechhofer, S., Paton, N. W., Stevens, R., and Brass, A., "An ontology for bioinformatics applications.," *Bioinformatics*, vol. 15, pp. 510–520, June 1999.

[10] Barb, A. and Shyu, C.-R., "User-specific semantics for modeling content-based information in geospatial knowledge," in *IEEE International Geoscience and Remote Sensing Symposium*, 2007.

[11] Barb, A. S. and Shyu, C.-R., "Semantics Modeling in Diagnostic Medical Image Databases Using Customized Fuzzy Membership Functions," in *Fuzzy Systems, 2003. FUZZ '03. The 12th IEEE International Conference on*, vol. 2, pp. 1159– 1164, May 2003.

[12] Barb, A. S., Shyu, C.-R., and Sethi, Y., "Semantic Integration and Knowledge Exchange for Diagnostic MedicalImage Databases," in *Bioinformatics and Bioengineering, BIBE 2004. Proceedings. Fourth IEEE Symposium on*, vol. 00, (Los Alamitos, CA, USA), p. 175, IEEE Computer Society, 2004.

[13] Barb, A. S., Shyu, C.-R., and Sethi, Y., "Knowledge Representation and Sharing using Visual Semantic Modeling for Diagnostic Medical Image Databases," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 4, pp. 538–553, 2005.

[14] Bentley, J., "Multidimensional binary search trees in database applications," *Software Engineering, IEEE Transactions on*, vol. SE-5, no. 4, pp. 333–340, July 1979.

[15] Berchtold, S., Keim, D. A., and Kriegel, H. P., "The X-tree: An index structure for high-dimensional data," in *Proceedings of the 22nd International Conference on Very Large Databases* (Vijayaraman, T. M., Buchmann, A. P., Mohan, C., and Sarda, N. L., eds.), (San Francisco, U.S.A.), pp. 28–39, Morgan Kaufmann Publishers, 1996.

[16] Berchtold, S., Bohm, C., and Kriegal, H.-P., "The pyramid-technique: towards breaking the curse of dimensionality," in *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of data*, (New York, NY, USA), pp. 142–153, ACM Press, 1998.

[17] Bianconi, F. and Fernandez, A., "Evaluation of the effects of gabor filter parameters on texture classification," *Pattern Recognition*, vol. 40, pp. 3325–3335, December 2007.

[18] Blum, A. L. and Langley, P., "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1-2, pp. 245–271, 1997.

[19] Bodenreider, O., "The unified medical language system (umls): integrating biomedical terminology.," *Nucleic Acids Res*, vol. 32, January 2004.

[20] Bohm, C., Berchtold, S., and Keim, D. A., "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," *ACM Comput. Surv.*, vol. 33, no. 3, pp. 322–373, 2001.

[21] Box, G. and Cox, D., "An Analysis of Transformations (with discussion)," *Journal of the Royal Statistical Society*, vol. B26, pp. 211–252, 1964.

[22] Brehmer, B., "Models of diagnostic judgments," in *Human Detection and Diagnostic of System Failures* (Rasmussen and Rouse, W. B., eds.), (New York: Plenum), pp. 231–241, 1981.

[23] Brewster, C., O'Hara, K., Fuller, S., Wilks, Y., Franconi, E., Musen, M. A., Ellman, J., and Shum, S. B., "Knowledge representation with ontologies: The present and future," *IEEE Intelligent Systems*, vol. 19, no. 1, pp. 72–81, 2004.

[24] Brown, M. S., McNitt-Gray, M. F., Goldin, J. G., and Aberle, D. R., "An extensible knowledge-based architecture for segmenting computed tomography images," in *ICIP '97: Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 3*, (Washington, DC, USA), p. 516, IEEE Computer Society, 1997.

[25] Buckland, M. and Gey, F., "The relationship between recall and precision," *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, 1994.

[26] Burgun, A., Botti, G., Fieschi, M., and P., L. B., "Sharing knowledge in medicine: semantic and ontologic facets of medical concepts," in *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, vol. 1, pp. 300–305, 300-305, .

[27] Cai, W., Feng, D. D., and Fulton, R., "Content-based retrieval of dynamic pet functional images.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 2, pp. 152–158, 2000.

[28] Carey, S., *Conceptual change in the childhood*. Cambridge, MA, 1985.

[29] Cha, G.-H. and Chung, C.-W., "Multi-mode indices for effective image retrieval in multimedia systems," in *ICMCS '98: Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, (Washington, DC, USA), p. 152, IEEE Computer Society, 1998.

[30] Chang, C.-C. and Lin, C.-J., *LIBSVM: a library for support vector machines*, 2001.

[31] Chawla, S., Shekhar, S., Wu, W., and Ozesmi, U., *Geographic Data Mining and Knowledge Discovery*, ch. Modeling spatial dependencies for mining geospatial data: An introduction. Taylor and Francis, 2001.

[32] Chu, W. W., Ieong, I. T., and Taira, R. K., "A semantic modeling approach for image retrieval by content," *The VLDB Journal*, vol. 3, no. 4, pp. 445–477, 1994.

[33] Ciaccia, P., Patella, M., and Zezula, P., "M-tree: An efficient access method for similarity search in metric spaces," in *The VLDB Journal*, pp. 426–435, 1997.

[34] Coenen, F., "LUCS KDD implementation of CMAR (Classification based on Multiple Association Rules)." Department of Computer Science, The University of Liverpool, UK., 2004.

[35] Coenen, F., Leng, P., and Ahmed, S., "Data structure for association rule mining: T-trees and p-trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 6, pp. 774–778, 2004.

[36] Cruz, I. F. and Rajendran, A., "Semantic data integration in hierarchical domains," *IEEE Intelligent Systems*, vol. 18, no. 2, pp. 66–73, 2003.

[37] Dafner, R., Cohen-Or, D., and Matias, Y., "Context-based space filling curves," *Computer Graphics Forum*, vol. 19, no. 3, pp. ??–??, 2000.

[38] de Oliveira, J. V., "Semantic constraints for membership function optimization.," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 29, no. 1, pp. 128–138, 1999.

[39] de Saussure, F., *Course In General Linguistics*. McGraw-Hill Humanities/Social Sciences/Languages, June 1965.

[40] Demastes, S., R., G., and Peebles, P., "Patterns of conceptual change in evolution," *Journal of Research in Science Teaching*, vol. 33, no. 4, pp. 407–431, 1996.

[41] DiSessa, A., "Towards an epistemology of physics," *Cognition and Instruction*, vol. 10, pp. 105–225, 1993.

[42] Dubois, D., Prade, H., and Sedes, F., "Fuzzy logic techniques in multimedia database querying: A preliminary investigation of the potentials," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 3, pp. 383–392, 2001.

[43] Eco, U., *The limits of interpretation*. Bloomington: Indiana University Press, 1990.

[44] Economou, G.-P. K., Lymberopoulos, D. K., Karavatselou, E. I., and Chassomeris, C. A., "A new concept toward computer-aided medical diagnosis - a prototype implementation addressing pulmonary diseases.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 5, no. 1, pp. 55–65, 2001.

[45] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M., "The sequence ontology: a tool for the unification of genome annotations.," *Genome Biol*, vol. 6, no. 5, 2005.

[46] Eklund, P. W., Kirkby, S. D., and Salim, A., "Data mining and soil salinity analysis.," *International Journal of Geographical Information Science*, vol. 12, no. 3, pp. 247–268, 1998.

[47] Faloutsos, C. and Roseman, S., "Fractals for secondary key retrieval," in *PODS '89: Proceedings of the eighth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, (New York, NY, USA), pp. 247–252, ACM Press, 1989.

[48] Fonseca, F., Egenhofer, M., Agouris, P., and Camara, G., "Using ontologies for integrated geographic information systems," 2002.

[49] Fox, J. and Thomson, R., "Decision support and disease management: A logic engineering approach.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, no. 4, pp. 217–228, 1998.

[50] Gardner, D., Knuth, K. H., Abato, M., Erde, S. M., White, T., DeBellis, R., and Gardner, E. P., "Common data model for neuroscience data and data model exchange.," *J Am Med Inform Assoc*, vol. 8, no. 1, pp. 17–33, 2001.

[51] Garner, S., "Weka: The waikato environment for knowledge analysis," in *In Proc. of the New Zealand Computer Science Research Students Conference*, pp. 57–64, 1995.

[52] Giacobazzi, R. and Mastroeni, I., "Compositionality in the puzzle of semantics," in *PEPM '02: Proceedings of the 2002 ACM SIGPLAN workshop on Partial evaluation and semantics-based program manipulation*, (New York, NY, USA), pp. 87–97, ACM Press, 2002.

[53] Gilbert, E., "Gray codes and paths on the n-cube," *Bell System Tech. J.*, vol. 37, pp. 815–826, 1958.

[54] Gkoutos, G. V., Green, E. C. J., Mallon, A.-M., Hancock, J. M., and Davidson, D., "Building mouse phenotype ontologies," in *Pacific Symposium on Biocomputing*, pp. 178–189, 2004.

[55] Gruber, T. R., "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[56] Gruber, T. R., "Toward principles for the design of ontologies used for knowledge sharing," *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5-6, pp. 907–928, 1995.

[57] Guttman, A., "R-trees: a dynamic index structure for spatial searching," in *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, (San Francisco, CA, USA), pp. 599–609, Morgan Kaufmann Publishers Inc., 1988.

[58] Guyon, I. and Elisseeff, A., "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[59] Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M., and Niblack, W., "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 7, pp. 729–736, 1995.

[60] Han, J., Pei, J., and Yin, Y., "Mining Frequent Patterns Without Candidate Generation," in *2000 ACM SIGMOD Intl. Conference on Management of Data* (Chen, W., Naughton, J., and Bernstein, P. A., eds.), pp. 1–12, ACM Press, May 2000.

[61] Haralick, R. M., Shanmugam, K., and I., D., "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybertinetics*, vol. 3, pp. 610–621, 1973.

[62] Holland, J. H., *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, April 1992.

[63] Hsu, C.-C., W., C. W., and Taira, R. K., "A knowledge-based approach for retrieving images by content," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 4, pp. 522–532, 1996.

[64] Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R., "Image indexing using color correlograms," in *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, (Washington, DC, USA), p. 762, IEEE Computer Society, 1997.

[65] Jonassen, D. H., Beissner, K., and Yacci, M. A., *Structural knowledge: Techiques for representing, conveying and acquiring structural knowledge*. Lawrence Erlbaum, 1993.

[66] Jonassen, D., Strobel, J., and Gottdenker, J., "Model building for conceptual change," *Interactive Learning Environments*, vol. 13, pp. 15–37, 2005.

[67] Kavouras, M. and Kokla, M., *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*. CRC Press, 2007.

[68] Kazic, T., "Semiotes: a semantics for sharing," *Bioinformatics*, vol. 16, no. 12, pp. 1129–1144, 2000.

[69] Kelly, P., Cannon, T., and D.R., H., "Query by image example: the candid approach," *Storage and Retrieval for Image and Video Databases III*, vol. 2420, pp. 238–248, 1995.

[70] Kindberg, T., Bryan-Kinns, N., and Makwana, R., "Supporting the shared care of diabetic patients," in *GROUP '99: Proceedings of the international ACM SIGGROUP conference on Supporting group work*, (New York, NY, USA), pp. 91–100, ACM Press, 1999.

[71] Kohonen, T., "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[72] Korn, P. F., Sidiropoulos, N., Faloutsos, C., Siegel, E., and Protopapas, Z., "Fast and effective retrieval of medical tumor shapes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 10, no. 6, pp. 889–904, 1998.

[73] Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, December 2003.

[74] Kruempel, K., "Making the right (interactive) moves for knowledge-producing tasks in computer-mediated groups," *IEEE Transactions on Professional Communication*, vol. 43, pp. 185–195, 2000.

[75] Kruskal, W. and Wallis, W., "Use of ranks on one criterion variance analysis," *J. Am. Statistical Assoc.*, vol. 47, pp. 119–127, 1952.

[76] Lawder, J. K. and King, P. J. H., "Using space-filling curves for multi-dimensional indexing," *Lecture Notes in Computer Science*, vol. 1832, pp. 20–??, 2000.

[77] Lawrence, C., Dong, Q., Polacco, M., Seigfried, T., and Brendel, V., "Maizegdb, the community database for maize genetics and genomics," *Nucleic Acids Res.*, vol. 1, no. 32, pp. 393–397, 2004.

[78] Lees, B. and Ritman, K., "Decision tree and rule induction approach to integration of remotely sensed and gis data in mapping vegetation in disturbed or hilly environments," *Environmental Management*, vol. 15, pp. 50–71, 1991.

[79] Leroy, G. and Chen, H., "Meeting medical terminology needs-the ontology-enhanced medical concept mapper," *IEEE Transactions on Information Technology in Biomedicine*, vol. 5, pp. 261–270, 2001.

[80] Liang, A., Salokhe, G., Sini, M., and Keizer, J., "Towards an infrastructure for semantic applications: Methodologies for semantic integration of heterogeneous resources," *Cataloging & Classification Quarterly*, vol. 43, pp. 161–189, April 2006.

[81] Liu, B., Hsu, W., and Ma, Y., "Integrating classification and association rule mining," in *KDD*, pp. 80–86, 1998.

[82] Liu, B., Hsu, W., and Ma, Y., "Integrating classification and association rule mining," in *Knowledge Discovery and Data Mining*, pp. 80–86, 1998.

[83] Liu, H. and Yu, L., "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[84] Loh, W.-Y., "Annals of statistics," *Bounds on AREs for restricted classes of distributions defined via tail-orderings*, vol. 12, pp. 685–701, 1984.

[85] Lorette, A., Descombes, X., and Zerubia, J., "Texture analysis through a markovian modelling and fuzzy classification: Application to urban area extraction from satellite images," *International Journal of Computer Vision*, vol. 36, pp. 221–236, February 2000.

[86] May, D. and Taylor, P., "Knowledge management with patterns," *Commun. ACM*, vol. 46, no. 7, pp. 94–99, 2003.

[87] Medasani, S. and Krishnapuram, R., "A fuzzy approach to content-based image retrieval," in *ICMCS '99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems Volume II-Volume 2*, (Washington, DC, USA), p. 964, IEEE Computer Society, 1999.

[88] Miller, A., *High-Resolution CT of the Chest: Comprehensive Atlas, 2nd Edition*, vol. 120. Lippincott, Williams & Wilkins, 2001.

[89] Miller, G. A., "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, November 1995.

[90] Mitaim, S. and Kosko, B., "The shape of fuzzy sets in adaptive function approximation," *IEEE-FS*, vol. 9, pp. 637–656, Aug. 2001.

[91] Mitchell, T. M., *Machine Learning*. New York: McGraw-Hill, 1997.

[92] Mouaddib, N. and Bonanno, N., "New semantics for the membership degree in fuzzy databases," in *ISUMA '95: Proceedings of the 3rd International Symposium on Uncertainty Modelling and Analysis*, (Washington, DC, USA), p. 655, IEEE Computer Society, 1995.

[93] Muller, H., Michoux, N., Bandon, D., and Geissbuhler, A., "A review of content-based image retrieval systems in medical applications–clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, pp. 1–23, February 2004.

[94] Mungall, C. J., "Obol: integrating language and meaning in bio-ontologies," *Comparative and Functional Genomics*, vol. 5, pp. 509+, August 2004.

[95] Nah, Y. and Sheu, P. C.-Y., "Image content modeling for neuroscience databases," in *SEKE '02: Proceedings of the 14th international conference on Software engineering and knowledge engineering*, (New York, NY, USA), pp. 91–98, ACM Press, 2002.

[96] O'Leary, D. E., "Enterprise knowledge management," *Computer*, vol. 31, no. 3, pp. 54–61, 1998.

[97] Otsu, N., "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 62–66, January 1979.

[98] Palmer, S. E., *Vision science: photons to phenomenology*. Cambridge, MA: MIT Press, 1999.

[99] Pass, G. and Zabih, R., "Histogram refinement for content-based image retrieval," in *WACV '96: Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)*, (Washington, DC, USA), p. 96, IEEE Computer Society, 1996.

[100] Pearl, J., "Bayesian networks: A model of self-activated memory for evidential reasoning," in *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, pp. 329–334, August 1985.

[101] Peirce, C. S., *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press, 1935-66.

[102] Pudil, P., Novovičová, J., and Kittler, J., "Floating search methods in feature selection," *Pattern Recogn. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[103] Pyle, D., *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[104] Quinlan, J., "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[105] Quinlan, J. R., "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[106] R Development Core Team, *The R reference manual: base package*. Network Theory, 2004.

[107] Rasmussen, J., "Diagnostic reasoning in action," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 4, pp. 981–992, 1993.

[108] Reeves, B. N. and Shipman, F., "Tacit knowledge: icebergs in collaborative design," *SIGOIS Bull.*, vol. 17, no. 3, pp. 24–33, 1996.

[109] Robinson, G. P., Tagare, H. D., S., D. J., and C., J. C., "Medical image collection indexing: shape-based retrieval using kd-tree," *Computerized Medical Imaging and Graphics*, vol. 20, pp. 209–217, 1996.

[110] Rowley, J. E., "The wisdom hierarchy: representations of the dikw hierarchy," *Journal of Information Science*, pp. 1–17, 2007.

[111] Russ, J. C., *Image Processing Handbook, Fourth Edition*. Boca Raton, FL, USA: CRC Press, Inc., 2002.

[112] Sagan, H., *Space-Filling Curves*. Springer-Verlag, 1994.

[113] Saint-Paul, R., Raschia, G., and Mouaddib, N., "Prototyping and browsing image databases using linguistic summaries," in *Proc. of the 11th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE'2002)*, vol. 1, pp. 476–481, May 2002.

[114] Samet, H., *Indexing Issues in Supporting Similarity Searching.*, pp. 463–470. Advances in Multimedia Information Processing - PCM 2004,5th Pacific Rim Conference on Multimedia, Tokyo, Japan, 2004.

[115] Santini, S., "Query paradigm to discover the relation between text and images," in *Proc. SPIE Vol. 4315, p. 161-171, Storage and Retrieval for Media Databases 2001, Minerva M. Yeung; Chung-Sheng Li; Rainer W. Lienhart; Eds.* (Yeung, M. M., Li, C.-S., and Lienhart, R. W., eds.), vol. 4315, pp. 161–171, Dec. 2000.

[116] Santini, S. and Jain, R., "Image databases are not databases with images," in *ICIAP (2)*, pp. 38–45, 1997.

[117] Shannon, C. E., "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, july, october 1948.

[118] Shyu, C. R., *A Physician-in-the-Loop Content-Based Image Retrieval System for Medical Image Databases*. PhD thesis, Purdue University, 1999.

[119] Shyu, C.-R., Barb, A., and Davis, C., "Mining Image Content Associations for Visual Semantic Modeling in Geospatial Information Retrieval And Indexing," in *Proc. of International Geoscience and Remote Sensing Symposium*, pp. 5622–5625, July 2005.

[120] Shyu, C.-R., Brodley, C. E., Kak, A. C., Kosaka, A., Aisen, A. M., and Broderick, L. S., "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," *Computer Vision and Image Understanding: CVIU*, vol. 75, pp. 111–132, / 1999.

[121] Shyu, C.-R., Harnsomburana, J., Green, J., Barb, A. S., Kazic, T., Schaeffer, M., and Coe, E., "Searching and mining visually observed phenotypes of maize mutants.," *J Bioinform Comput Biol*, vol. 5, no. 6, pp. 1193–213, 2007.

[122] Shyu, C.-R., Klaric, M., Scott, G., Barb, A. S., Davis, C., and Palaniappan, K., "GeoIRIS: Geospatial Information Retrieval and Indexing System - Content Mining, Semantics Modeling, and Complex Queries," *IEEE Transactions on Geoscience and Remote Sensing, Special Issue on Image Mining*, vol. 45, pp. 839–852, 2007.

[123] Shyu, C.-R., Pavlopoulou, C., Kak, A. C., Brodley, C. E., and Broderick, L. S., "Using human perceptual categories for content-based retrieval from a medical image database," *Comput. Vis. Image Underst.*, vol. 88, no. 3, pp. 119–151, 2002.

[124] Sidhu, A. S., Dillon, T. S., Chang, E., and Sidhu, B. S., "Ontology-based knowledge representation for protein data," in *Proceedings of IEEE's 3 rd International Conference on Industrial Informatics*, pp. 535–539, 2005.

[125] Smith, J. R. and Chang, S.-F., "Tools and techniques for color image retrieval," in *Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 426–437, 1996.

[126] Stern, E. J. and Collins, J., "Hrct: Know your buzz words," *Society of Thoracic Radiology Annual Meetings*, pp. 205–207, 2000.

[127] Stevens, R., Goble, C. A., and Bechhofer, S., "Ontology-based knowledge representation for bioinformatics," *Brief Bioinform*, vol. 1, no. 4, pp. 398–414, 2000.

[128] Stricker, M. A. and Orengo, M., "Similarity of color images," in *Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 381–392, 1995.

[129] Swain, M. and Ballard, D., "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, November 1991.

[130] Tayar, N., "A model for developing large shared knowledge bases," in *CIKM '93: Proceedings of the second international conference on Information and knowledge management*, (New York, NY, USA), pp. 717–719, ACM Press, 1993.

[131] Tiwana, A., "Affinity to infinity in peer-to-peer knowledge platforms," *Commun. ACM*, vol. 46, no. 5, pp. 76–80, 2003.

[132] Trombert-Paviot, B., Rodrigues, J. M., Rogers, J. E., Baud, R., van der Haring, E., Rassinoux, A. M., Abrial, V., Clavel, L., and Idir, H., "Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures," *International Journal of Medical Informatics*, vol. 58-59, pp. 71–85, September 2000.

[133] Tuceryan, M. and Jain, A. K., "Texture segmentation using voronoi polygons," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-12, no. 2, pp. 211–216, 1990.

[134] Tuceryan, M. and Jain, A. K., "Texture analysis," *Handbook of Pattern Recognition and Computer Vision*, pp. 235–276, 1993.

[135] Vapnik, V. N., *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, November 1999.

[136] Von Helmholtz, H. L. F., *Treatise on Physiological Optics (translated and edited by J. P. C. Southhall)*. New York, USA: Dover, 1962.

[137] Vosniadou, S., "Capturing and modeling the process of conceptual change," *Learning and Instruction*, vol. 4, pp. 45–69, 1994.

[138] Wang, H., Liu, S., and Chia, L.-T., "Image retrieval with a multi-modality ontology," *Multimedia Systems*, vol. 13, no. 5-6, pp. 379–390, 2008.

[139] Watkins, A., Timmis, J., and Boggess, L., "Artificial immune recognition system (AIRS): An immune-inspired supervised machine learning algorithm," *Genetic Programming and Evolvable Machines*, vol. 5, pp. 291–317, September 2004.

[140] Weber, R., Schek, H.-J., and Blott, S., "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, (San Francisco, CA, USA), pp. 194–205, Morgan Kaufmann Publishers Inc., 1998.

[141] Wei, C. P., Hu, P. J.-H., and R., S. O., "A knowledge-based system for patient image pre-fetching in heterogeneous database environments - modeling, design, and evaluation.," *IEEE Transactions on Information Technology in Biomedicine*, vol. 5, no. 1, pp. 33–45, 2001.

[142] West, L. and Pines, A., *Cognitive structure and conceptual change*. New York: Academic Pr, 1985.

[143] White, D. A. and Jain, R., "Similarity indexing with the ss-tree," in *ICDE '96: Proceedings of the Twelfth International Conference on Data Engineering*, (Washington, DC, USA), pp. 516–523, IEEE Computer Society, 1996.

[144] Wilcoxon, F., "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.

[145] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Sys, Morgan Kaufmann, second ed., June 2005.

[146] Wong, S. T. C., *Medical Image Databases*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[147] Xu, L., Yan, P., and Chang, T., "Best first strategy for feature selection," in *ICPR88*, pp. II: 706–708, 1988.

[148] Yamazaki, Y. and Jaiswal, P., "Biological Ontologies in Rice Databases. An Introduction to the Activities in Gramene and Oryzabase," *Plant Cell Physiol.*, vol. 46, no. 1, pp. 63–68, 2005.

[149] Yin, X. and Han, J., "CPAR: Classification based on Predictive Association Rules," in *Proceedings of 2003 SIAM International Conference on Data Mining*, pp. 369–376, SIAM Press, 2003.

[150] Yoon, H. and Yang, K., "Feature subset selection and feature ranking for multivariate time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1186–1198, 2005. Member-Cyrus Shahabi.

[151] Zeleny, M., "Management support systems: Towards integrated knowledge management," *Human Systems Management*, vol. 7, no. 1, pp. 59–70, 1987.

VITA

Adrian S. Barb received the bachelor degree in industrial engineering from the University of Bucharest, Romania, School of Engineering in 1990, a master's degree in business administration from the University of Missouri Columbia (MU) in 2002. He is currently working on a PhD degree in Computer Science at MU. He also worked as a database programmer analyst for the Information Access and Technology Services at MU. His research interests include knowledge representation and exchange in content-based retrieval systems, semantic modeling and retrieval, capturing and predicting conceptual change in knowledge-base systems, ontology integration, and expert-in-the-loop knowledge exchange. He is a student member of IEEE.