ONTOLOGY_BASED DECISION SUPPORT FOR MULTIDISCIPLINARY MANAGEMENT OF BREAST CANCER

S Dasmahapatra*, D Dupplaw*, B Hu*, H Lewis+,

P Lewis*, M Poissonnier**, N Shadbolt*

+School of Engineering, *School of Electronics and Computer Science,

University of Southampton, Southampton SO17 1BJ, UK

**Oxford University, Medical Vision Lab, Ewert House, Oxford OX2 7DD UK

The decision-making process for the management of patients with breast cancer involves a consultation based on results of X-ray and other imaging technologies, clinical patient information, pathological evidence from cell and tissue extracts and so on. The evidence is presented in the specialised vocabulary of each of the experts and constitutes the input to the decisions taken for the management of the disease. We have developed a knowledge-based framework which builds on an index of key concepts from each of the fields of expertise organised in domain ontologies, which also organizes the relationships between these concepts with unambiguous machine-readable semantics. Appropriate evidential information and case notes can then be annotated with respect to this conceptual index, and we provide retrieval facilities for drawing upon the stored information via the concepts relevant to the domain at hand. Moreover, our system architecture is an example of a Semantic Web (Berners-Lee et al 2001) application which provides web-based access to and processing of relevant information, enabling different specialist departments in dispersed locations to collaborate. We demonstrate the relevance of such a system by showing the performance of classification services

integrated into this framework. As indicated, at the multi-disciplinary meeting the different specialists use their independent means and concepts to describe features of the case at hand. We used the Digital Database for Screening Mammography (DDSM) (Heath et al 1998) of the University of South Florida[1] to extract the image descriptors for X-ray mammograms to construct several classifiers for predicting likely malignancy of a case given its description. These classifiers can be run on remote servers with results returned to the user client in the meeting room.

After a brief presentation of the domain ontology in Section 1, Section 2 provides an overview of the MIAKT architecture and application framework within which the server side functionalities, such as classification as reported in Section 4 and image feature extraction are accommodated. Section 3 provides browsing facilities via a client-side case exploration method using a lattice-based visualisation technique called Formal Concept Analysis (FCA) (Ganter and Wille 1999).

1. Domain ontology

The concept terms of the ontology are compliant with the BI-RADS lexicon (ACR 2001) and are organized from the abstract to the concrete, *eg.*, concrete descriptors like "Spiculated Margin" are subsumed under high-level concepts such as "Medical Image" or "Image Descriptor." "Image Descriptor" has a subclass "Morphologic Descriptor" which in turn has a subclass "Mammogram Specific Margin Morphology" which has a concrete type "Spiculated Margin". Also, five top level roles (properties) are used to

[1] http://marathon.csee.usf.edu/Mammography/Database.html

represent five different generic categories of role referencing relationships among concepts. These abstractions allow ontology-based tools to be largely independent of details of data-typing at the specific end of the descriptive hierarchy. The ontologies are compliant with the Web Ontology Language (OWL) standard (McGuinnes and van Harmelen 2003), and stored as RDF triples (Lassila and Swick 1999) in the 3store database (Harris and Gibbins 2003).

2. The Architecture

The MIAKT architecture provides a generic remotely-accessible structure for rapidly prototyping new applications in new domains, and is thus deliberately abstracted from any particular application domain (and its description). The application ontology, which provides the link to the resources that will be available in this application, is divided into two distinct ontologies: the client ontology and the framework ontology. The client ontology describes the resources available to the client application. The framework ontology describes where and how many of those resources are accessed, mapped and initiated.

To provide access to web-services, a server-side service architecture is designed to be extended to provide access to different services. It is accessed via mapped task names, providing a flexible bridging mechanism between the client and the services. Bringing web-services into the application as component objects provides access to functionality that might otherwise be impossible to integrate into a desktop solution. Servers with large storage capacity or with large processing capability can be simply accessed through the generic client to provide domain specific and non-domain specific algorithms.

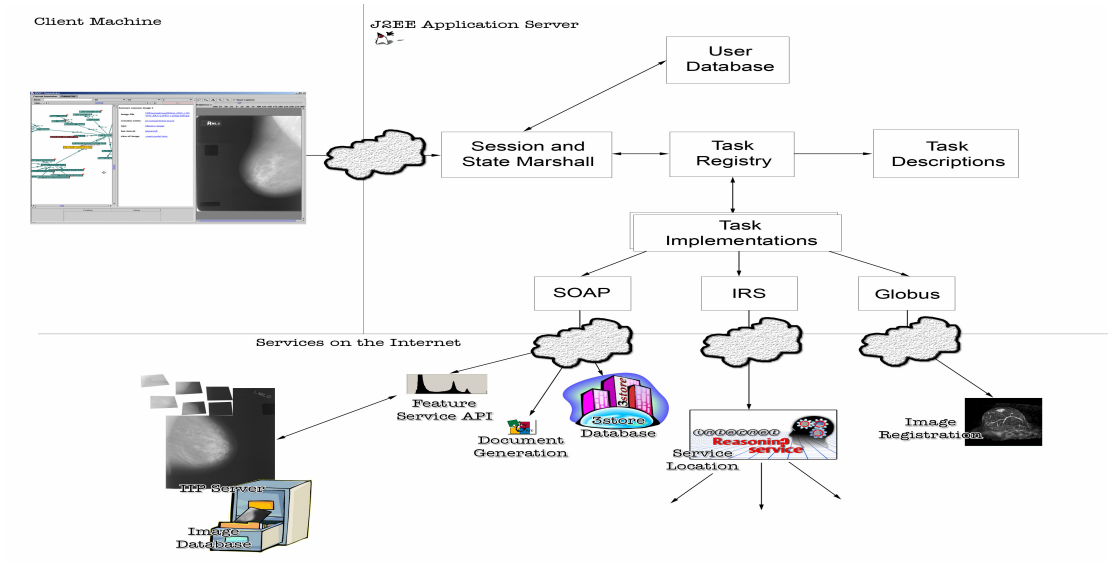

Figure 1 - The MIAKT Server-Side Server Invocation System

A user transparently interacts with the architecture through an application client that is also built around a generic architecture and invoked as instances of the application ontology.

2.1 Image retrieval over the Web

The distribution of the X-ray image data is achieved using a network protocol called the Internet Imaging Protocol (IIP) for transmitting large images across relatively small bandwidth pipes as image tiles (Martinez et al 1998). The image tiles that are returned from the servlet interface can be at various resolutions allowing low resolution, "overview" images to be downloaded quickly, providing better response times to the user interface. Image feature extraction modules can use tile retrieval to retrieve only the parts of the image necessary for their operation, thereby lowering bandwidth and latency costs. For example, should a user make an annotation on an image that some feature module will be providing feature extraction on, only that small area of the image needs to

be transferred (unless the feature module specifically requires otherwise). Similarly, using image tiles provides a method for only the necessary parts of images that are being displayed to be transferred, again increasing the response of the user interface. Although this provides a quicker response in initial image display, it potentially slows manipulation of the image view as tiles are retrieved during scrolling.

2.2 Image feature extraction

Association of features of an image or region of interest with the concepts in the ontology is the mechanism that provides a powerful model for storage of knowledge. Regions of interest withing an image may be manually or semi-automatically generated, or retrieved from legacy data. Association of a feature to concepts defined in the ontology is provided by a simple point and click mechanism. The user highlights the feature which they are going to associate with a concept, and finds the relevant concept in one of the concept browsers. They are then able to right click and associate this feature with a concept. During this process the feature vector is stored in a feature database, which provides indexed, feature-dependant retrieval of features, and the unique ID of the feature vector is inserted as an instance of the given concept.

Image feature extraction algorithms are provided to the client mainly through the web-service interface. A defined web-service interface called the Feature Service API has been developed which provides a simple, extendible web-service framework for publishing feature extraction modules. The API provides functionality for storing, retrieving, and comparing feature vectors from images, and automatically provides feature modules with the relevant regions from the source media. The image feature

extraction modules can use tile retrieval to retrieve only the parts of the image necessary for their operation, thereby lowering bandwidth and latency costs.

3. Lattice-based browsing of patient records

In this section we briefly report on a lattice-based visualization technique we have deployed to enable browsing an entire set of cases based on the various attributed by which these cases have been annotated. The technique used is called Formal Concept Analysis (FCA) which relies on a description of concepts in terms of a pairing of instances and their attributes (Ganter and Wille 1999). A lattice is constructed by identifying various intersections of predicative sets (each set containing elements possessing one common feature value) where the partial ordering relation is derived from set inclusion. This partial order is used to stitch a lattice together. The number of attributes increases as we go down the page and the corresponding number of elements in the intersection of these predicative sets reduces.

Instead of giving the appropriate mathematical description of FCA, we illustrate how it might be useful by means of an example. In Figure 2, we present a fragment of the DDSM cases. The descriptors for the anonymous cases of the DDSM such as age, metadata of the mammogram images, and abnormalities as found on the images, are mapped to instances of our ontology and re-expressed as description-logic (DL) instances in RDF format, which is the input to the FCA module. In the figure, the circular large nodes hold links to the cases which are described by the attributes that can be traced via the lines on the lattice above them. Thus the rightmost lowest node contains the cases which contain benign masses with the features Lobulated and Focal Asymmetric Density.

For the cases represented in the figure, it so happens that the Lobulated ones subsume the ones with Focal Asymmetric Density. However, there are other Lobulated cases which also have the value Irregular for the Shape feature which are all malignant, as can be seen from the node second from the left among the four on the same line. Thus the cases can be browsed in an exploratory fashion by following different pathways to the nodes reached by following the lines below the attributes in the figure.

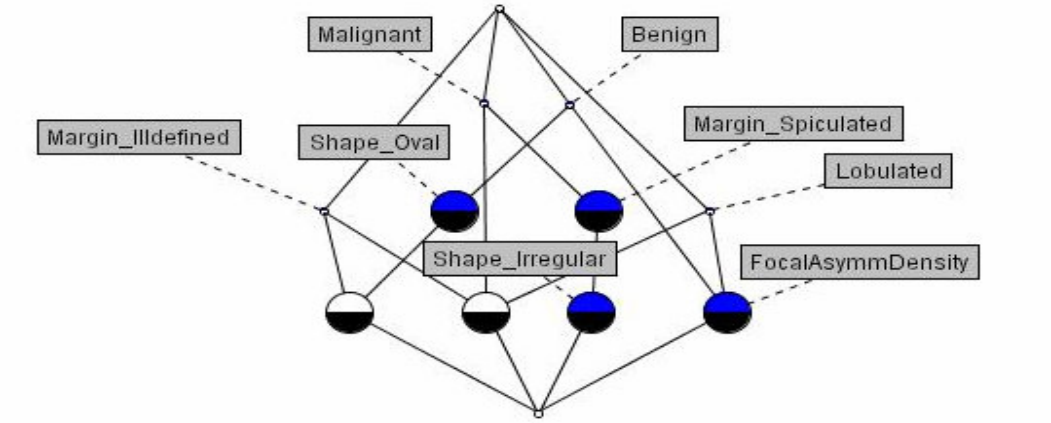

Figure 2 – FCA lattice for a sample of the DDSM cases.

4. Classification Services

The architecture described above has been designed to accommodate, among other methods, classification services based on features extracted from images via automatic means by image processing methods run on entire images or parts of the image segmented out by the hand-drawn region highlighting. In addition, in the multi-disciplinary meeting for patient management, the cases already come with expert labels attached – shape features for masses seen in X-rays, for example. The classifiers that we present all take BI-RADS descriptors for X-ray images as input and classify cases according to their likelihood of being benign or malignant. We illustrate the ideas by describing two of the classification methods we have tried out (Mitchell 1997).

We normalise the data extracted from the DDSM collections by assigning binary values for each of the metadata features like Shape, Margin, Architectural Distortion, and so on, representing whether or not such a feature has been recorded in the case notes accompanying each image. We used approximately 1500 of these cases to build our classifiers, with statistics accumulated from the (co)-occurrence or not of the features available.

4.1 Naive Bayes

The task of probabilistic classification of a case based on the statistics of occurrence of the features that represent it is one of finding which label maximizes the conditional probability of a label given the observed features P(label | features). Bayes' rule is used to interchange the order of conditioning, and in applying the Naive Bayes condition, we make the simplifying assumption (often violated) that the joint probability of the symptoms given the identification of the disease state (the target classification labels -- benign, malignant or unproven) factorizes into the product of the probabilities of the individual features given the classification label.

For the dataset at hand, each of the features is assumed to be drawn from a binomial distribution except for the feature Age, which is assumed to be drawn from a Gaussian. To work around small sample errors, and in particular, the appearance of zero frequencies, we assume a uniform Laplace/beta prior (equivalent to introducing a pseudo-count of 1 for each feature, with the necessary normalization). After dividing up the data

set into those involving masses and those with microcalcifications present, we obtain the results summarized in Table 1.

Lesion	Correct Classification	False Pos	False Neg
Mass	77.9%	10.2%	11.1%
Calcification	73.6%	8.4%	17.7%

Table 1 -- Classification results for Naive Bayes.

4.2 Linear classifiers and Multi-layer Perceptrons

As an alternative method for classifying instances we tried training a multi-layer perceptron (MLP) on the data. We compared a simple linear classifier with a binary output variable obtained by taking a logistic function on a linear combination of input data with trainable coefficients, to a MLP with different numbers of hidden variables – 2, 4, 8, 16 and compared the results (see Figure 3 for illustration).

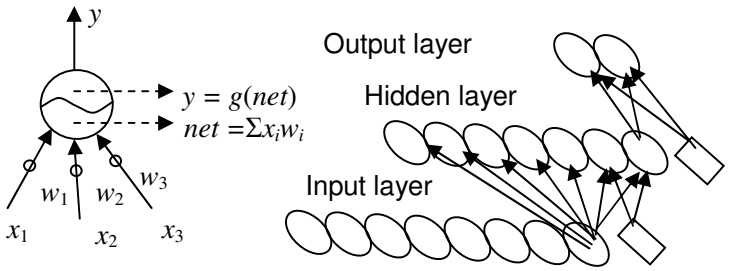

Figure 3 – Perceptron and MLP

The networks were trained by backpropagation with momentum. The classification accuracy is summarized in Figure 4.

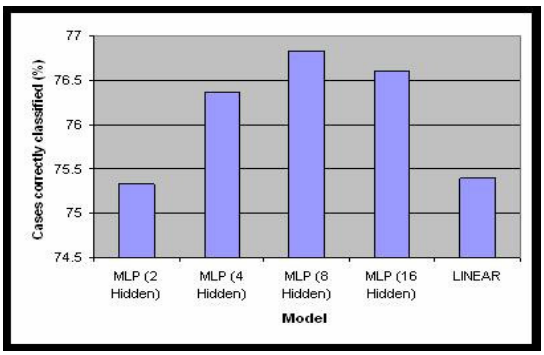

Figure 4 - Relative accuracy of MLP and linear classifiers.

5. Summary

This paper has illustrated how an ontology-based distributed system architecture can be invoked to run useful services such as diagnostic classification for decision support for breast cancer management.

References

Heath, M., K.W. Bowyer, D. Kopans. 1998. Current Status of the Digital Database for Screening Mammography, 457-460 in *Digital Mammography,* Kluwer.

Harris, S and N. Gibbins. 2003. 3store: Efficient Bulk RDF Storage. In *Proceedings 1st Int Wkshp on Practical and Scalable Semantic Web Systems*, Sanibel Is., Florida, USA.

Martinez, K., J. Cupitt, and S. Perry. 1998. High resolution Colorimetric Image Browsing on the Web, *Int. Conf. World Wide Web, WWW-7*, Elsevier. 30(1-7) 4.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American*, May issue.

American College of Radiology (ACR). 2001. Breast Imaging Reporting and Data System: BIRADS. Available at http://www.acr.org/departments/stand_accred/birads/.

McGuinnes, D. and F. van Harmelen. 2003. Ontology Web Language (OWL) Overview. Available at http://www.w3.org/TR/owl-features/.

Ganter, B. and R. Wille. 1999. *Formal Concept Analysis,* Springer-Verlag, Berlin.

O. Lassila and R.R. Swick. 1999. Resource Description Framework (RDF) Model and Syntax Specification. W3C. Available at http://www.w3c.org/.

Mitchell, T. 1997. *Machine Learning.* McGraw Hill.