Enriched Property Ontology for Knowledge Systems

A thesis presented in partial fulfilment of the requirements for the

Degree

of

Master of Information Systems

In

Information Systems

Massey University, Palmerston North, New Zealand

Conducted at LBD EPFL, Switzerland

Robert Minchin

October 2006

Contents

1 Introduction

"It is obvious that every individual thing or event has an indefinite number of properties or attributes observable in it and might therefore be considered as belonging to an indefinite number of different classes of things" [Venn 1876].

The world in which we try to mimic in Knowledge Based (KB) Systems is essentially extremely complex especially when we attempt to develop systems that cover a domain of discourse with an almost infinite number of possible properties. Thus if we are to develop such systems how do we know what properties we wish to extract to make a decision and how do we ensure the value of our findings are the most relevant in our decision making. Equally how do we have tractable computations, considering the potential computation complexity of systems required for decision making within a very large domain. In this thesis we consider this problem in terms of medical decision making.

Medical KB systems have the potential to be very useful aids for diagnosis, medical guidance and patient data monitoring. For example in a diagnostic process in certain scenarios patients may provide various potential symptoms of a disease and have defining characteristics. Although considerable information could be obtained, there may be difficulty in correlating a patient's data to known diseases in an economic and efficient manner. This would occur where a practitioner lacks a specific specialised knowledge. Considering the vastness of knowledge in the domain of medicine this could occur frequently. For example a Physician with considerable experience in a specialised domain such as breast cancer may easily be able to diagnose patients and decide on the value of appropriate symptoms given an abstraction process however an inexperienced Physician or Generalist may not have this facility.

Accordingly Physicians may be precluded from providing a correct or rapid diagnostic that ultimately has adverse affects on the patient or leads to the requirement of possibly unnecessary medical tests. Historically diagnostic KB Systems have not been tremendously successful within the medical practice, other than as simple support tools. This is thought to be caused by:

a) the limited scope (useful for a small specific domain only) of such tools

b) the inability to handle conflicting symptoms

c) the lack of consideration of the diagnostic process used by doctors.

In order to overcome these barriers, we propose the use of an extensible property rich ontology for mapping domains of decisions, with each sub-class/domain associated with a reference class and a set of KB systems. This approach guides the system user to the core set of properties that should be targeted in decision making or querying and should increase the relevance of each property used in decision making. This approach uses the existing knowledge of ontologies, decision systems such as Bayesian networks and statistics. It combines these fields so that we are able to:

a) Query an ontology that maps the domain of decision via properties, not only by sub domains. Enabling the potential of scope to map very large domains.

b) Increase the power of our decision systems within the applicable sub domain. Allowing inference of diagnosis when conflicting symptoms exist.

c) Build domain knowledge with domain experts (Physicians) thus the specific abstraction or decision making process may be mapped.

We proposed that this methodology presented in detail in section 4 could equally be used in conjunction with existing KB Systems to increase their scope and precision, for example; integrating a specialised diagnosis system Athena (for hypertension) with a general diagnostic system DXplain.

1.1 Motivation

In this thesis we propose that a property rich ontology may represent a domain and map expert abstraction to sub domains of decision. This enables the possibility to define key property variables for classifying an unknown thing in a large domain. This would make each finding obtained potentially more valuable for decision making. Equally we propose that an effective method of defining what a thing is in machine computational terms and in a large domain would enable a considerable advancement in knowledge based systems i.e. navigation systems, aids for the disabled, security systems, Medical KB and other KB systems.

Vast/complex classification systems cannot be effective or efficient if the system does not have a method of targeting what properties it wishes to consider. Current systems are too limited in scope, do not offer solutions of objectively defining specific property extraction and are essentially non extensible. Difficulties in overcoming problems of classification in a wide domain are illustrated by the limitations of use of belief revision methods and pattern recognition as shown in section 2.

As a case study we are considering medical expert systems. There has been considerable development of medical expert, decision support systems or KB systems since the 1970s. However, KB systems have still only had a very limited effect on the medical practice largely because these systems are either very specialised, are only accurate in specific domains and unreliable in others or the systems are just too simplified [9], (in terms of scope). In addition these systems may not have the possibility of deduction of error when a suspect incorrect classification has been made.

> Episodic skeletal plan refinement (ESPR) – A problem solving method that classifies and provides output on a defined protocol logic (skeletal plan) that is hierarchical and often time based to match medical treatment protocols. The skeleton is refined to the appropriate level of abstraction on an episodic basis (E.g. each patient visit).
>
> Computational Complexity – Evaluation of the required resources used during computation to solve a given problem, considering how many steps it takes to solve a problem (time) and how much memory is required. Time or space required to

solve the problem is considered as a function of the size of the input problem; for example the difficulty of finding a particular disease will become harder as we have a greater number of possible diseases and symptoms.

Current medical KB systems that use a system of episodic skeletal plan refinement (ESPR) [33] may well represent the temporal nature of patient treatment and medical guidance. However, these systems are only used for specific diagnosis/guidance process e.g. AIDS. In addition these systems may not consider reference class problems. A reference class is a like group or class having similar referenced attributes. The problem is that probability/inference is specific to a group or to a referenced class and should be interpreted according to the appropriate group. For instance, if we consider two different references, e.g. European Middle Aged Female vs. Polynesian Adolescent Male, the symptom inferences could be quite dissimilar. By using the reference class information we are potentially using known verifiable statistics to have more powerful variables in our classification and decision making systems for each reference group.

There is an obvious cause of limitations that affects all decision based systems, that is the complexity of making decisions in a large domain. Medical diagnoses processes are likely to have an extremely large set of properties, and in order to cope with this complexity experienced Physicians may work in different levels of abstraction by refining target symptoms. Thus if we are to improve the scope and accuracy of diagnostic KB systems we firstly need to look for a method of defining properties (symptoms and characteristics) that are relevant for given patient scenarios. Secondly we need to ensure that the relevant KB systems maximise the significance of variables in accordance with available findings, i.e. we need to consider the inference of reference class information.

In order to tackle these two issues we propose the use of a related four staged-approach:

a) The 1st stage being the design/formation of an extensible ontology considering the natural domains of decisions associated with a reference class or expert defined abstraction trees.

b) The 2nd stage being the collection of the information or statistics applicable to the class references.

c) The 3rd stage being the development of a set of KB systems associated with each domain from the reference class information.

d) The 4th stage being the enrichment of the ontology classes with attributes defined in the associated KB systems to create a property enriched ontology.

The three binary relationships between the ontology, Reference Classes and KB systems are illustrated in Figure 1-1. The ontology classifies the sub-domains of the universe of discourse and contains the properties that can be applied to each class or sub-domain (enriched). The reference classes are the statistics or information extracted considering the conditional implications of the considered sub-domain. The KB system(s) contains the decision formula(s) or model(s) used to define the next level of abstraction, constructed from the reference class information. We refer to KB systems in a plural sense with each domain as we put no restriction on the type of decision systems used in the methodology.

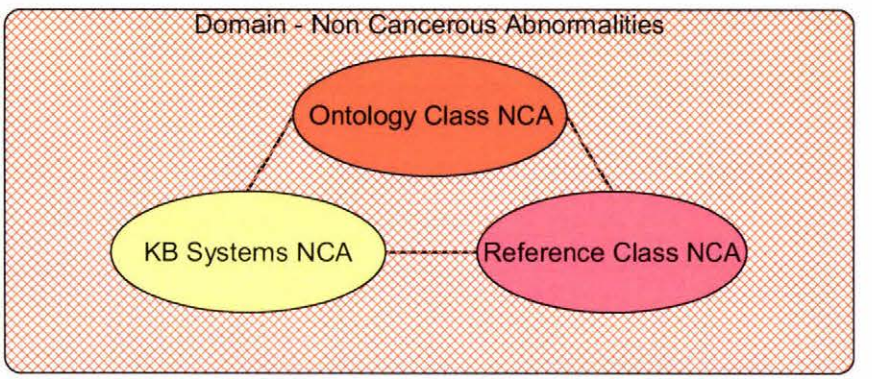

Figure 1-1 : Domain Triangle

The structure of querying implied is that we first verify whether a sufficient property exists to define a disease or a disease group (domain). Alternatively the disease domain can be established by the non sufficient properties or characteristics. From a specific domain or class reference the associated KB systems are used to target key properties for the decision. These target properties and findings then define the next level of decision or abstraction.

The symptoms ontology is static while being extensible i.e. a classification of disease symptoms is essentially fixed and will extend as new knowledge is learnt. The KB system or sets of worlds are likely to be reactionary to belief change and to the probabilities defined from reference class statistics, i.e. as we learn new information about a patient's characteristics or symptoms the implications of decision or diagnosis change. Thus an ontological mapping structure could enable the development of a vast database of diagnostic properties that could be queried effectively by symptoms/characteristics to classify patient disorders. This is possible because the data is stored in domain granules linked via the ontology and defined by class reference information. Such a structure manages the complexity of diagnostic methods. The structure increases the value of findings because unimportant symptoms for the domain are not requested and decisions or beliefs can be based on the associated reference class. In addition, such a system could know inherently when it has made an inappropriate domain allocation decision as new findings are added and could dynamically adjust i.e. when a conclusive decision cannot be defined.

In terms of medical diagnosis this could mean faster, more accurate diagnosis and reduced cost by reducing the number of medical tests to form an acceptable certainty in diagnosis. To demonstrate this methodology we have developed a prototype SOMKS (Symptoms Ontology for Mapping Knowledge Systems) that maps Knowledge domains of Breast Health using Stanford's Protégé tool and Netica developed Bayesian Networks to represent domain specific Knowledge bases.

1.2 Related Work

Classification systems generally avoid global decision domains in order that assumptions and prior knowledge about the domain can be applied; i.e. the problem is overcome by avoidance in creating systems that have very defined and limited domain of operation such as a sensory based quality control in chip manufacturing. Methods of belief revision. are introduced in later sections, have not had a large impact because they are restricted in terms of computational complexity or they are unable to provide rational revisions, as developed in [28]. Concepts of granular computing introduce the human decision making process of hierarchy and abstraction that we attempt to better map in our system.

For our case study we consider specifically Medical KB systems. The medical informatics community has built a considerable number of KB systems to aid medical practitioners in many ways. Recent systems address extensively temporal nature of medicine and use ontologies in defining medical protocol through processes including episodic skeletal plan refinement (ESPR) see [33]. The common complaints about these systems are that they are highly domain specific or are excessively general.

> Rational Revision – Revision of a belief that meets the basic AGM postulates (section 2), for example when a revision is added to a belief formula and then removed the original belief formula should be obtained.

1.3 Our Contribution

The methodology that we have developed uses the pillars of existing knowledge concerning ontologies, KB systems and statistics. We combined these approaches to develop a manageable method on increasing the scope of classification or decision systems.

We recognise that there are many systems and proposals for managing classification operations. However it has been generally concluded that these systems, either, do not manage classifications/decisions well in a large domain, are excessively complex to be practically used or are just too simplified. In order to overcome these limitations in KB Systems, we propose the use of an extensible property rich ontology that maps target reference classes that have associated KB systems. The KB systems defined from reference class information directs the system to a set of target properties that can lead the system to a more precise abstraction or lower level reference class.

We are applying a granular approach to specify target properties, to increase the value in decision making of each property defined (finding) and to allow a system to potentially know when an inappropriate reference class has been defined and dynamically correct this. In addition the core of the ontology is essentially static and extensible. For example, LCIS is likely to continue to be defined as a type of non-invasive cancer with specific symptoms and if a new type of invasive cancer is defined it can be added to the super-class of Invasive cancer without affecting LCIS (see the complete ontology in appendix 1). The KB systems in turn could be dynamically adjusted via traditional methods while limiting impact of computation complexity.

1.4 Thesis Outline

In the preliminaries we introduce issues in pattern recognition, class reference and belief change affected by the limitations of scope in classification, computational complexity and KB systems. We then introduce Bayesian networks and OWL, that are used in our prototype SOMKS. We further discuss Medical Expert systems and potentially why they have had a less than expected impact on medical practice.

In the related work section we review the historic developments in belief change and discuss their limited use due to either not being rational or having complexity limitations. We introduce some of the principle medical KB systems. We also introduce the granular computing whose objectives relate strongly to our methodology.

In the theory section we review the advantages that our design and querying approach of property enriched ontology for mapping domains, would bring to KB systems. We then consider such an approach in conjunction with medical KB systems.

In section 5 we review the prototype SOMKS functionalities, and the tools/software used for its construction. In section 6 we introduce our case study using 'SOMKS' for the diagnoses of breast abnormalities. SOMKS uses a breast health ontology containing classifying properties of symptoms and patient characteristics that are then used to lead the system to a class reference domain. The user is then focused on a finite set of features that can be defined to diagnose patient abnormalities.

The prototype is outlined in figure 1-2. SOMKS ontology reasoner finds the most appropriate sub-class or domain granule from initial specified symptoms/characteristics. If SOMKS can not distinguish between a possible disorder and a healthy patient from initial information, it then requests additional information based on the defined key variables for the domain of decision using the knowledge reasoner. The knowledge reasoner then defines the next level of abstraction. Theoretically SOMKS should also enable the re-querying of the ontology with the new findings.