4

Building Ontology from Knowledge Base Systems

Faten Kharbat[1] and Haya El-Ghalayini[2]
[1]*Zarqa Private University ,*
[2]*Petra University*
Jordan

1. Introduction

In the last decade, ontologies have been considered as the backbone technology in most knowledge-based applications. As ontologies have become more common, their applicability has ranged from artificial intelligence areas such as knowledge representation and natural language processing to different fields such as information integration and retrieval systems, requirements analysis, and lately in semantic web applications.
In the literature, several methodologies and methods have been introduced for building ontologies. Some of these methods allow the development of ontologies from existing ontologies or data sources. However, the proposed method for building ontologies integrates different data mining techniques to assist in developing a given domain ontology. Thus, the extracted and representative rules generated from the original dataset can be utilised in developing ontology elements.
The main research hypothesis in this chapter is that ontology can be developed from discovered hidden and interesting rules. In order to practically investigate this assumption, this chapter presents a complete developing discovery structure using one of the well known breast cancer test sets.
The chapter is organised into five sections. A general overview is found in section two with a brief description of the main components of this research. The development engine framework is introduced in the section three. Section four demonstrates proposed method using Wisconsin Breast Cancer dataset as a case study. Finally, this practical investigation ends by presenting the learned lessons and conclusions.

2. Background

2.1 What is ontology?

In the last decade, ontologies have been considered as the backbone technology in most knowledge-based applications. As ontologies have become more common, their applicability has ranged from artificial intelligence areas such as knowledge representation and natural language processing to different fields such as information integration and retrieval systems, requirements analysis, and lately in semantic web applications.
What is an ontology? This question may be answered from two viewpoints: philosophical and computing. From the philosophical viewpoint, *Ontology* (with an upper case 'O') is an

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

ancient branch of enquiry, initiated by the Ancient Greeks and continued through the Middle Ages till the Modern Age. Ontology is the study of what exists in the world: *beings, their nature, and essential properties*. In Ontology, philosophers try to answer questions such as what are things or beings, and how things can be classified.

The second viewpoint of *ontology* (with a lower case 'o') has been emerging into the discipline of computer science during the last 10-20 years. Ontology was initially proposed by the artificial intelligence community to model declarative knowledge for knowledge-based systems and shared with other systems. Recently, the utilisation of ontologies attracted attention in the development of information systems (Guarino, 1998). Also, the evolution of the semantic web has encouraged the development of ontologies. This is because an ontology represents the shared understanding and the well-defined meaning of a domain of interest, thereby enabling computers and people to collaborate better (Gómez-Pérez et al., 2004).

The most popular definition of ontology was proposed by Gruber (1993), who defined it as "*…a formal, explicit specification of a shared conceptualisation*". In this definition, Gruber placed emphasis on formalising the specification of concepts and relations, which in turn allows for knowledge representation and sharing among different agents. Studer et al. (1998) analysed this definition, and identified four main concepts: *formal, explicit, shared, and conceptualisation*. The term *formal* means that an ontology should be machine readable; *explicit* implies that all concepts and constraints used are explicitly defined; *shared* indicates that an ontology should capture consensual knowledge accepted by the communities involved; and *conceptualisation* refers to an abstract model of phenomena in the real world arrived at by identifying the relevant concepts of those phenomena. Another relevant definition of an ontology was introduced by Guarino (1998): "*a set of logical axioms designed to account for the intended meaning of a vocabulary*". In this definition, Guarino highlighted the role of logic theory as a means of representing an ontology.

As a conclusion, ontologies formalise the semantics of the domain explicitly by describing their elements; and thus, they consist of concepts that describe the internal features of the concepts, and the properties that describe the relationships between these concepts. Ontologies are based on a shared and consensual domain knowledge agreed by a community. Because of these properties, ontologies can support a wide variety of tasks in diverse research areas. Here are some examples:

1. The integration of heterogeneous data sources can benefit from the use of a domain ontology to overcome semantic heterogeneities (Lacroix and Critchlow, 2003).
2. An ontology enables explicit and consensual knowledge to be shared and reused between human and software agents (Uschold and Jasper, 1999).
3. An ontology can be used to build knowledge bases - a knowledge base being an ontology with a set of instances (Noy and McGuinness, 2001). Also, ontologies can be used in deriving aspects of information systems at development or run time (Guarino, 1998). For example, ontology-based retrieval systems can assist users to browse and understand domain concepts, and therefore, formulate better specialised queries (Baker et. al, 1999).

2.1.1 Types of ontology

Different kinds of ontologies exist that have been specified for different application domains thereby representing different types of knowledge. This section classifies ontologies along the following three dimensions: level of formality, level of generality, and primitive types.

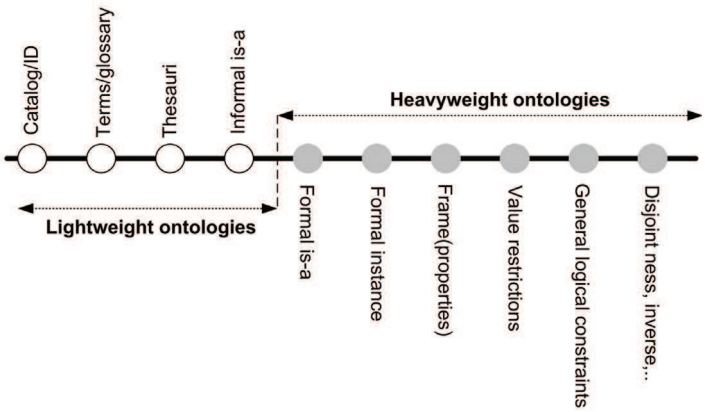


Fig. 1. An Ontology Expressiveness Spectrum adapted from McGuinness (2003)

In the first dimension, Uschold & Grüninger (1996) proposed distinguishing ontologies according to the degree of formality in the specification of the concepts. The basic types in this dimension are: *informal* ontologies comprising a set of concepts written in a natural language and organised in a hierarchy; *semi-formal* ontologies consisting of a hierarchy of concepts such as taxonomies defined by simple axiomatisation; and *formal ontologies* defining the semantics of the vocabulary in a formal language using complete axioms. McGuinness (2003) took his classification further by adding a number of levels of formality and expressiveness. The ontology spectrum in Fig. 1 depicts both the level of semantic expressiveness and the distinction between heavyweight and lightweight ontologies. Corcho et al. (2003) introduced *lightweight* ontologies as a set of concepts, properties and concept taxonomies**,** whereas *heavyweight* ontologies include, in addition, axioms and constraints.

In the second dimension (i.e., level of generality), Van Heijst et al. (1997) and Guarino (1998) classified ontologies according to their levels of conceptualisation. These are: *general* ontologies (or top-level ontologies) which define very general concepts that are independent of a particular domain such as space, time, thing, event, or property; *domain* ontologies which define concepts for a specific domain; *task* ontologies which describe concepts for specific task activities; *application* ontologies which describe concepts relating to both domain and the task activities; and *representation* ontologies which describe representation entities that are used in knowledge representation formalisms.

Finally, the third dimension for the classification of ontologies was proposed by Jurisica et al. (2004), which may be considered as the basis for capturing primitive concepts for large scale applications. This allows for a further classification of ontologies to cover the *static*, *dynamic*, *intentional*, and *social* aspects of the real world. A *static* ontology describes static aspects of the world which includes what things exist, their attributes and their relationships. Its main aim is to unify the domain concepts to enable information sharing and system cooperation. A *dynamic* ontology is concerned with the changing aspects of the world in terms of its states, state transitions, and processes. An *intentional* ontology covers the world of things that agents believe in, prove or disprove, and argue about, in terms of concepts such as issue, and goal. Finally, a *social* ontology involves social settings in terms of the key concepts of actor, position, and commitment.

After having introduced ontologies and their different classifications along with the potential support for them in different application areas, the next section introduces the different methodologies and methods for developing ontologies.

2.1.2 Methods and methodologies for developing ontologies

Ontologies have received much attention from researchers in different application areas in the computer science community. Therefore, different approaches have been reported for developing ontologies. This section presents methods and methodologies for developing ontologies from two perspectives: (1) Building ontology from scratch, and (2) Building ontologies from existing ontologies or from different data sources

In what follows a set of approaches related to the first perspective are introduced. In 1995 Grüninger & Fox (1995) proposed a methodology based on TOVE (TOronto Virtual Enterprise) project and Uschold & King (1995) proposed a method built upon the experience assembled from developing the Enterprise ontology. The two approaches were used to build ontologies about enterprise modelling processes (Pinto & Martins, 2004). The first activity in Grüninger and Fox methodology identifies the main scenarios that describe the purpose of the ontology with respect to the intended applications. Then, a set of competency questions are used to identify the scope of the ontology, thereby extracting the main concepts, properties, axioms of the underlying scope. After that, the elements of the ontology are expressed in first order logic. The Uschold and King's method proposes the following activities: (1) Identify the purpose of the ontology, (2) build the ontology by capturing knowledge and identifying key concepts and properties in the domain, coding knowledge, and reusing other ontologies inside the current one, (3) Evaluate the ontology, and (4) document the ontology. In 1996, the methodology METHONTOLOGY (Gomez-Perez et al., 1996) for building ontologies from scratch or from reusing other ontologies was proposed and influenced by software engineering methodologies. It identifies the ontology development process where the life cycle is based on evolving prototypes. In 2001, Noy and McGuinness proposed an iterative approach to ontology development. The approach starts with a rough first pass at the ontology. This is followed by revising and refining the evolving ontology and filling in the details.

Since building ontologies from scratch is not a simple task and is a time-consuming process, next we introduce the research work related to the second perspective, which studies the approaches for developing ontologies either from reusing existing ontologies or from reusing different data sources. For example, the developed ontology at Kactus (Bernaras et al, 1996) is built on the basis of an application knowledge base. In other words, the approach starts by building a knowledge base for an application. After that, when another knowledge base in a similar domain is needed, the first knowledge base can be generalized into an ontology. The output of repeating this process can lead to the development of an ontology that represents the consensual knowledge needed in all applications (Corcho et al, 2003). Also, Swartout and colleagues (1997) proposed an approach for deriving domain specific ontologies Sensus ontology (which contains more than 70,000 concepts) In this case, a set of related seed terms in a certain domain can be identified, then all the concepts in the path from the seed terms to the root of Sensus are included.

Furthermore, Maedche & Staab (2001) distinguished different approaches for developing ontologies from existing data sources based on the type of input. The input can be one of the following: (1) text where the ontology development is carried out by applying natural

language analysis techniques to texts; (2) dictionary where the relevant concepts and relations of an ontology is extracted from a machine readable dictionary; (3) knowledge base; is used an existing source for building an ontology; (4) semi-structured data is used for eliciting an ontology from sources which have any predefined structure; (5) relational schema aims to extract relevant concepts, properties, relations from databases schema or relations.

2.2 Knowledge based system & knowledge discovery

Knowledge based systems can be considered as a special type of database "that holds information representing the expertise of a particular domain [Milton, 2008]". Rule based systems are one of knowledge based systems where the each rules can be expressed by *If-Then* statement. The *if*-part is the Left Hand Side (LHS), which is also called the antecedent. It consists of one or more of condition elements. The representation of the conditions may be categorized for simple problems, integer/real intervals or combination of these for more complex problems. The *then*-part -which is called the Right Hand Side (RHS), consequent or action- consists of number of actions. However, in this chapter a rule has only one action. Usually, each rule is associated with some characteristics or features that strengthen or weaken the rule.

Developing and creating rule-based systems is carried out by knowledge discovery techniques which may vary from simple to complicated algorithms. Knowledge discovery is the broader process of turning low level data into high level knowledge which includes data mining with other essential steps; pre-processing and post-processing [Freitas, 2003]. All techniques have different capabilities and limitations; therefore, combining more than one technique is a beneficial way to enhance their capabilities and overcome their limitations. Many approaches tend to combine with evolutionary algorithms in order to make use of their search capability in complex spaces. This chapter uses the learning classifier systems (LCS) [Holland, 1986], which are considered as an evolution-based learning system [Peña-Reyes & Sipper, 2000]. The main advantage of using LCS is its extraction of comprehensible knowledge that provides higher level of readability which is not found in sub-symbolic approaches. LCS has been used in [Kharbat, 2006] to investigate generating readable, interpretable, and organised rules so as to extract high quality knowledge that can be utilised in understanding the real-domain problem.

3. Ontology development engine architecture

Fig. 2 illustrates the general framework to construct and develop an ontology based on the ruleset generated from previous discovery process. The architecture of the ontology development engine consists of the following phases.

1. Phase 1-Knowledge discovery and rules preparation

This phase is concerned with the extraction of patterns from the selected dataset over which a learning system, learning classifier system in particular, is applied. The generated rules are prepared in a suitable form to match the engine requirements.

2. Phase 2-Ontology development engine algorithm

This phase proposes a new algorithm to develop domain ontology from the generated ruleset. In this step, the ontology development engine considers a given ***domain ontology*** as a set of concepts used to describe a specific domain. The concepts are structured by the

means of two types of properties namely, *subsumption and domain properties*. The subsumption property represents the subtype relation in which one concept is more general than another whereas the domain property represents the relationships between domain concepts.

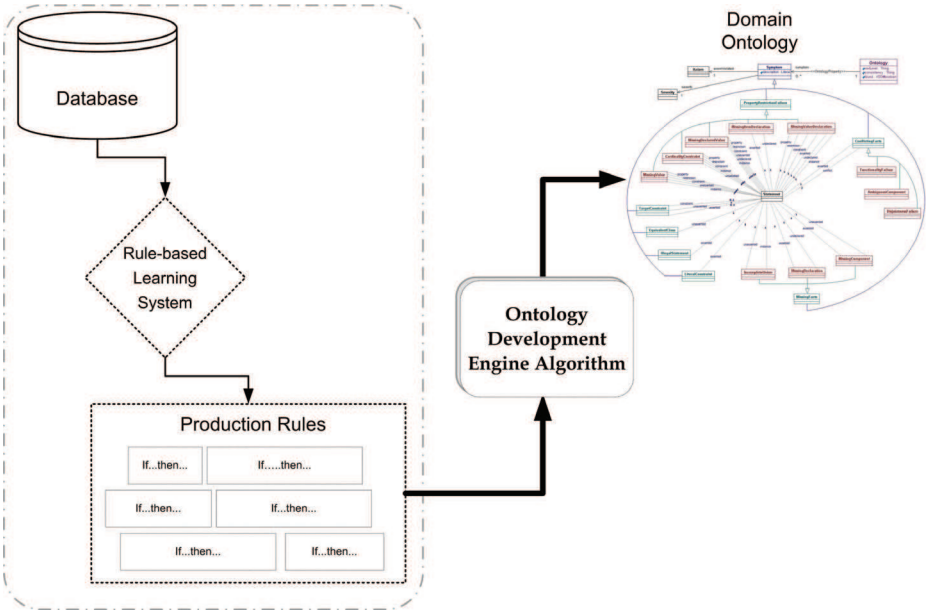


Fig. 2. A general framework.

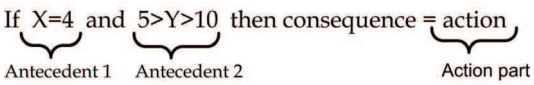
3.2 Knowledge discovery and rule preparation

In this phase, the initial data exploration is performed to verify the dataset completeness and missing attributes. Moreover, a technical preparation of the dataset in which pre-processing and reformatting mechanisms is performed to meet the requirements of the data mining techniques used within the investigation.

Preparing the original dataset is followed by applying a learning system, learning classifier system in this chapter, over the dataset. This allows the creation of a new knowledge base system which it is able to discover compacted, organised, and representative knowledge and to build a prediction model in order to predict future cases, and to deal with generating a readable human-interpretable output to describe the problem effectively**.**

The final step in this phase is to prepare the generated ruleset to suit the ontology development engine in the next phase. Firstly, weak rules from the ruleset are identified and removed based on their ***low experience*** or ***high prediction error***. The low experience of rules indicates that they either match a very small fraction of the dataset, which obviously could be matched by other rules, or by those that were generated late in the training process, which implies that the learning system did not have enough time to decide whether to delete them or approve their fitness. Moreover, the high prediction error of a rule indicates its inaccuracy and/or that it has very significant missing information.

Secondly, any integer interval antecedent is converted to a categorical one to minimize the scope of the rules and to group the antecedent into fewer clusters. For example, figure ** shows a rule illustrating its parts as follows:



Each integer-interval antecedent (i.e., X=4 and 5>Y>10) is converted to a categorical antecedent based on a determined scale. The scale may vary from one application to another, but this research suggests the following procedure to assist in describing new concepts related to the underlying domain:

- Assign category="Low" for each interval between 1 and 3.
- Assign category="Mid" for each interval between 4 and 6.
- Assign category="High" for each interval between 7 and 10.
- If the interval joins more than one category, then it will be described by an OR operator.

For example, the above rule is transformed using the suggested procedure as follows:

- The first antecedent X=4 falls in the "Mid" category, therefore, it will be replaced by: X=Mid
- The second antecedent 5>Y>10 falls in two categories, that is, Y joins two categories, i.e., "Mid" and "High". Therefore, this antecedent can be replaced by: Y=Mid OR Y=High.

3.3 Ontology development engine algorithm

This section illustrates the second phase of the general framework in figure 2. In this phase, a new ontology development engine algorithm is proposed to accept the generated discovered ruleset as an input to develop a domain ontology that describes representative concepts of the underlying domain. The proposed algorithm is described as follows:

Ontology development Engine

Input : Set of rules (RS) in the form of if antecedent(s) Then Consequent

Output : Suggested domain ontology

Algorithm:

Define Ontology concepts set (OCS) =∅

For every $rule_i \in RS$

Begin-for

1. Map $consequent_i$ to a concept C_i.
2. if $Ci \notin OCS$ then
   Add Ci to OCS
   End-if
3. Define Description Set (DS)=∅
4. for every antecedent of the form ($antecedent_x$ =$category_y$)
   Begin-for
   a. Map $antecedent_x$ to a concept C_x .
   b. Map $antecedent_x$_$category_y$ to a concept C_{xy}.
   c. Attach a subsumption relation between C_x and C_{xy}.
   d. Add C_{xy} to DS.
   End-for
5. Describe concept C_i using the intersection logical operator between elements of DS.

6. Map $Rule_i$ To a concept $RuleC_i$.
7. Attach a subsumption relation between $RuleC_i$ and C_i

End-for.

4. Wisconsin Breast Cancer (WBC) ontology: a case study

4.1 Breast cancer & wisconsin breast cancer datasets

According to the statistics that Breast Cancer Care [2004] has recently presented, 41,700 women are diagnosed each year with breast cancer in the UK, which equals 25% of the total number of cancer diagnoses. Although breast cancer seems to primarily affect women, 1% of the cases are men. Cancer Research UK [online], defines cancer to be "a disease where cells grow out of control and invade, erode and destroy normal tissue". Breast cancer is defined as "a *malignant* growth that begins in the tissues of the breast" [Matsui & Hopkins, 2002].

Briefly, the breast is composed of lobules, ducts, and lymph vessels. The lobules produce milk, and are connected by the ducts that carry the milk to the nipple. Lymph vessels, which are part of the body's immune system, drain and filter fluids from breast tissues by carrying lymph to lymph nodes, which are located under the arm, above the collarbone, and beneath the breast, as well as in many other parts of the body [Highnam & Brady, 1999]. There are many types of breast cancer depending on the tumour's properties, location, and/or size. However, the main challenge in breast cancer treatment is to find the cancer before it starts to cause symptoms; the earlier the cancer is detected, the better chances cancer patients have for cure and treatment. One of the problems is the limitation of human observations: 10-30% of the cases are missed during routine screening [Cheng et al., 2003]. With the advances in data mining algorithms, radiologists and specialists have the opportunity to improve their diagnosis for current cases, and prognosis of the new ones. And thus, scientists have a chance to gain a better understanding of both cancer's behaviour and development.

Wisconsin Datasets are three well-known breast cancer datasets from the UCI Machine Learning Repository [Blake & Merz, 1998]: (1) Wisconsin Breast Cancer Dataset (WBC) which has the description of histological images taken from fine needle biopsies of breast masses, (2) Wisconsin Diagnostic Breast Cancer Dataset (WDBC) where 30 characteristics of the cell nuclei present in each image are described, and (3) Wisconsin Prognostic Breast Cancer Dataset (WPBC) which contains follow-up data on breast cancer cases.

Development of the WBC dataset started in 1989 in Wisconsin University Hospitals by Dr. William Wolberg, and since then it has been heavily used as a test bed for machine learning techniques as a medical dataset [Mangasarian & Wolberg, 1990]. It consists of 699 test cases, in which every case has nine integer attributes associated with the diagnosis. Also, each attribute ranges between 1 and 10 where 1 indicates the normal state of the attribute and 10 indicates the most abnormal state. The diagnostic parameter (action) has binary possibility as either *malignant* or *benign*. The nine attributes are: *Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli* and *Mitoses*.

4.2 Knowledge discovery over the wisconsin datasets

Table 1 shows the classification accuracy of LCS (XCS in particular) over the WBC (average and standard deviation) compared to other popular learning algorithms. The experiments were performed using the well-known and traditional classification techniques namely, C4.5

[Quinlan, 1993] and XCS [Wilson, 1995]. C4.5 is a well known decision tree induction learning technique which has been used heavily in the machine learning and data mining communities. The output of the algorithm is a decision tree, which can be represented as a set of symbolic rules in the form of if-then. For the current dataset, results showed that the C4.5 technique achieved using the Weka software [Witten & Frank, 2005] are $95.4_{(1.6)}$ classification accuracy.

The Learning Classifier System (LCS) [Holland, 1976] is a rule-based system which uses evolutionary algorithms to facilitate rule discovery. It has been applied to different data mining problems and shown effectiveness in both predicting and describing evolving phenomenon (e.g., [Holmes *et al.*, 2002]). The vast majority of LCS research has made a shift away from Holland's original formalism after Wilson introduced XCS [Wilson, 1995]. XCS uses the accuracy of rules' predictions of expected payoff as their fitness. In addition, XCS uses a genetic algorithm (GA) [Holland, 1975] to evolve generalizations over the space of possible state-action pairs of a reinforcement learning task. XCS has been shown to perform well in a number of domains (e.g., [Bull, 2004]). For the current dataset, results showed that the XCS technique achieved $96.4_{(2.5)}$ classification accuracy.

	C4.5	XCS
WBC	$95.4_{(1.6)}$	$96.4_{(2.5)}$
WDBC	$92.61_{(1.98)}$	$96.13_{(2.48)}$

Table 1. Accuracy of XCS and C4.5 on the WBC and WDBC averaged over 10 trials, one standard deviation shown in parentheses

It can be seen from table 1 that XCS achieved the highest classification accuracy showing the efficiency and ability of XCS to tackle real complex problems; therefore, the generated rules (knowledge) from XCS are to be applied in the next step for ontology development. Appendix A illustrates the generated ruleset from Wilson (2001) which contains 25 rules all of which considered as strong and efficient patterns that assist in describing breast cancer domain.

Before implementing the Ontology Development Engine over the ruleset, a preparation phase will be performed as explained in the section 3.2.

Example 1:

Rule#1 in Appendix A states:

If ***1>Uniformity of Cell Shape>4 and***
1>Bare Nuclei>4 and
1>Bland Chromatin>3 and
Normal Nucleoli= 1 and
Then ***the diagnosis=benign***

The preparation of this rule converts all the antecedents in the rule to categorized antecedents. Thus, Rule#1 is prepared as follows:

If ***Uniformity of Cell Shape=Low or Mid***
Bare Nuclei=Low or Mid and
Bland Chromatin=Low and
Normal Nucleoli= Low and
Then ***the diagnosis=benign***

4.2 Applying ontology development engine to WBC-ruleset

Applying the ontology development engine algorithm begins after preparing the ruleset which is considered as a source input knowledge for developing the WBC ontology. In what follows, the process of applying the proposed algorithm to WBC ruleset is described by a walked-through example to a specific rule.

Example 2:

Having prepared Rule#1 that describes *benign* diagnosis in example 1, the algorithm of ontology development starts as follows:

1. Figure 3 shows the mapping of the consequent of Rule#1 to a ***benign*** concept using Step-1 since it is not included in the ontology concepts set (ODC).

Fig. 3. mapping a consequent to a concept

2. The new concept of a ***benign*** is added to **OCS** using Step-2.
3. A new description set (**DS**) is defined as an empty set to accumulate the rule antecedents' definitions using Step-3.
4. The four antecedents in Rule#1, will be transformed using Step-4 as follows:
   i. The first antecedent of Rule#1 is mapped to a concept of ***Uniformity-of –Cell-Shape***.
   ii. The antecedent of Uniformity of Cell Shape=Low or Mid is mapped to a concept of Low-Uniformity-of-Cell-Shape and to a concept of Mid- Uniformity-of-Cell-Shape.
   iii. A subsumption relation is attached between the concept of ***Uniformity-of Cell-Shape*** and the sub-concepts ***of*** (***Low-Uniformity-of-Cell-Shape*** and Mid-***Uniformity-of-Cell-Shape)*** as shown in Fig. 4.

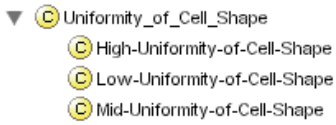


Fig. 4. mapping the subsumption relation between *Uniformity of Cell Shape* and all its antecedent_categories

   iv. Add the concepts of Low-Uniformity-of-Cell-Shape and Mid-Uniformity-of-Cell-Shape to the Description Set (DS) using the union logical operator. i.e., DS contains: Low-Uniformity-of-Cell-Shape ⊔ Mid-Uniformity-of-Cell-Shape.

The sub-steps of Step-4 will be repeated for the following antecedents:

Bare Nuclei=Low or Mid
Bland Chromatin=Low
Normal Nucleoli= Low

Thus, the following concepts and properties will be generated:

   i. The concepts of Bare-Nuclei, Bland-Chromatin, and Normal-Nucleoli..
   ii. The concepts of Low-Bare-Nuclei, Mid-Bare-Nuclei, Low-Bland-Chromatin, and Low-Normal-Nucleoli.
   iii. A subsumption relation are attached between (1) the concept of ***Bare-Nuclei*** and the sub-concepts of (***Low-Bare-Nuclei*** and ***Mid-Bare-Nuclei);*** (2) the concept of ***Bland-Chromatin*** and the sub-concept of ***Low-Bland-Chromatin***; (3) the concept of ***Normal-Nucleoli*** and the sub-concept of ***Low-Normal–Nucleoli.***
   iv. DS contains: (Low-Uniformity-of-Cell-Shape ⊔ Mid-Uniformity-of-Cell-Shape), (Low-Bare-Nuclei ⊔ Mid-Bare-Nuclei), Low-Bland-Chromatin, and Low-Normal-Nucleoli.

5. The concept of ***benign*** is described using the intersection logical operator between elements of DS as follows:

$$(\textit{Low-Uniformity-of-Cell-Shape} \sqcup \textit{Mid-Uniformity-of-Cell-Shape}) \sqcap (\textit{Low-Bare-Nuclei} \sqcup \textit{Mid-Bare-Nuclei}) \sqcap \textit{Low-Bland-Chromatin} \sqcap \textit{Low-Normal-Nucleoli}$$

6. Rule#1 is mapped to a concept as shown in Fig. 5.

Fig. 5. mapping Rule#1 to a concept

7. A subsumption relation is attached between Rule#1 and the concept of a ***benign*** as illustrated in Fig. 6.

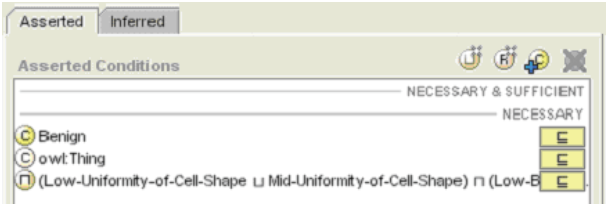


Fig. 6. Rule#1 is a sub-concept from the benign concept and has an intersection concept

This process proceeds for the 25 rules generated from the first-phase of the approach where the WBC ontology is illustrated from different snapshots in Figures 7, 8 and 9.

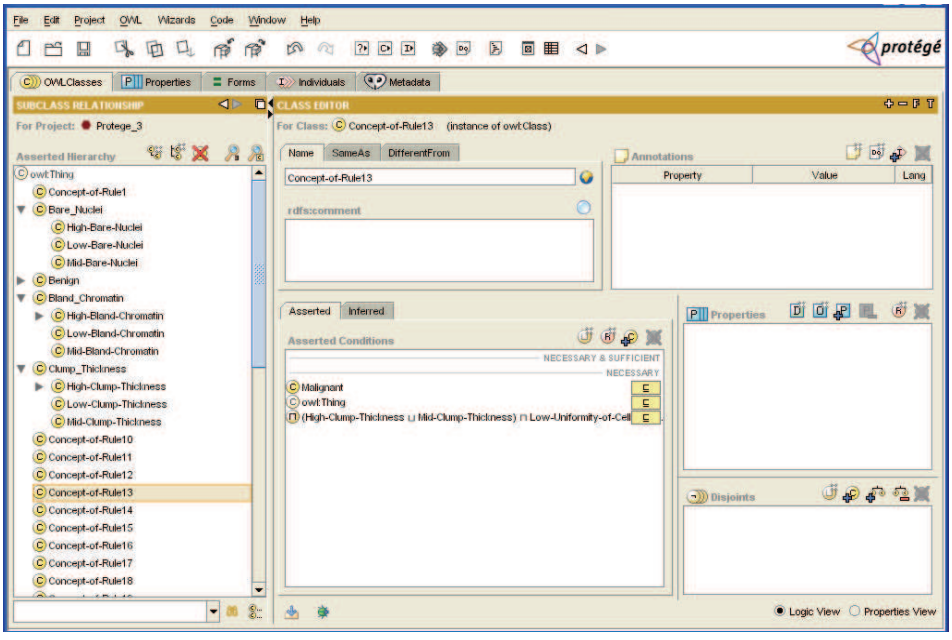


Fig. 7. A snapshot of the generated WBC ontology for rule#13 concept

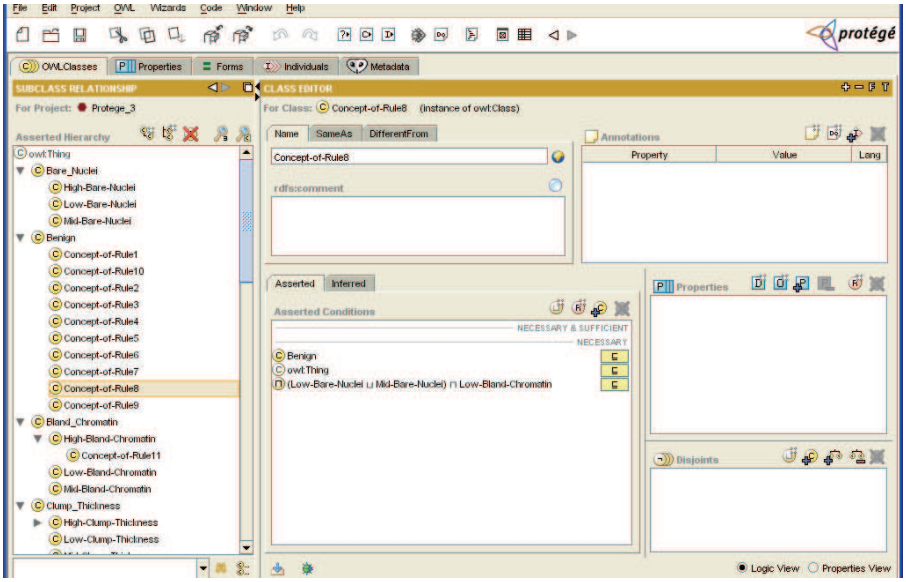


Fig. 8. A snapshot of the generated WBC ontology for Benign concept

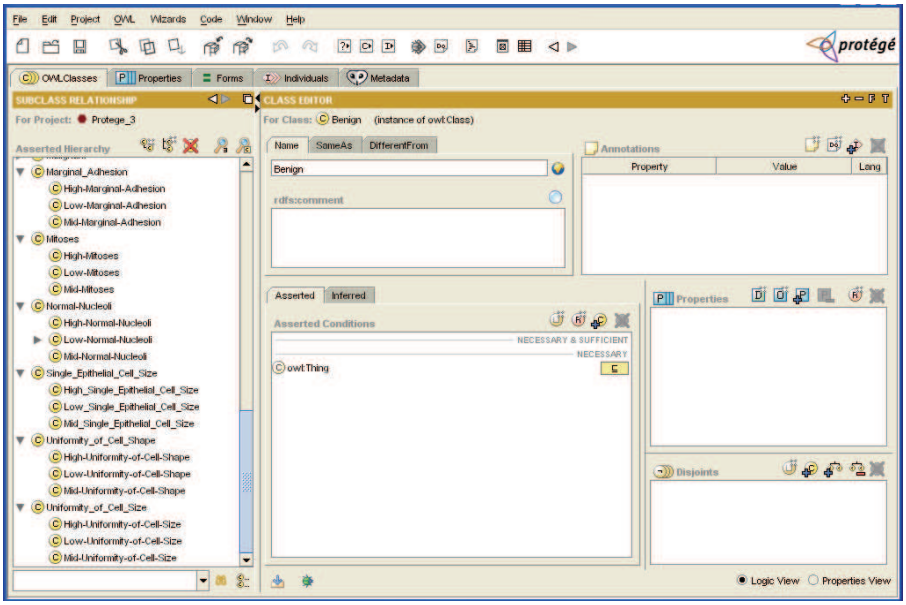


Fig. 9. A snapshot for some of the WBC ontology concepts

5. Conclusion

This chapter has presented a new approach to develop certain domain ontology. The proposed approach has integrated different data mining techniques to assist in developing a

set of representative consensual concepts of the underlying domain. The ontology development algorithm is proposed to transform a generated discovered ruleset to domain ontology. Learning classifier system has been used to generate a representative ruleset, and the Wisconsin Breast Cancer Dataset (WBC) has been selected as a test case. After applying the first phase of the proposed approach to WBC, the generated rulset from XCS contains 25 rules that mainly describe two concepts (Benign and Malignant). The results from phase two have produced WBC ontology with the description of more than concepts using subsumption relations, and the logical operators (and/or) without any human interaction.
While, this research has been focused on exploring the main concepts of the underlying domain, future work needs to consider the possibility of exploring the intrinsic and mutual properties of that domain. This may suggest enriching the process of ontology development and alleviating the complexity in understanding a shared and consensual domain knowledge agreed by a community.

6. References

Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics,* 15(6), pp. 510-520.

Bernaras, A., Laresgoiti, I. , and Corera, J. (1996). Building and reusing ontologies for electrical network applications, in: Proc. European Conference on Artificial Intelligence (ECAI'96), Budapest, Hungary, pp. 298-302.

Blake, C., & Merz, C. (1998). UCI Repository of Machine Learning Databases [online]. Irvine, CA: University of California, Department of Information and Computer Science. Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html

Breast Cancer Care. (2004). Breast Cancer Facts and Statistics [online]. Available from: http://www.breastcancercare.org.uk/content.php?page_id=1730

Bull, L. (2004)(ed) *Applications of Learning Classifier Systems*. Springer.

Cancer Research. UK, What Is Cancer? [Online]. Available from: http://www.cancerresearchuk.org /aboutcancer/whatiscancer/

Cheng, H., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-Aided Detection and Classification of Micro-calcifications in Mammograms: *A Survey Pattern Recognition*, 36 (12), pp 2967-2991.

Corcho, O., Fernández-López, M., and Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering,* 46(1), pp. 41–64.

Freitas, A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery. *Advances in Evolutionary Computing: Theory and Applications*. In: A. Ghosh and S. Tsutsui (eds). Natural Computing Series. Springer-Verlag, pp 819-845.

Gómez-Pérez, A., Fernández-López, M., de Vicente, A. (1996). Towards a Method to Conceptualize Domain Ontologies, in: ECAI96 Workshop on Ontological Engineering, Budapest, pp. 41-51.

Gómez-Pérez, A., Fernandez-Lopez, M., and Corcho, O. eds. (2004). *Ontological engineering: with examples from the areas of knowledge management, e-Commerce and the semantic web*. London: Springer-Verlag.

Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition,* 5(2), pp. 199-220.

Grüninger, M. , Fox, M.S. (1995)., Methodology for the design and evaluation of ontologies, in: Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal.

Guarino, N. (1998). Formal ontology and information systems. *In:* N. Guarino, ed. *Formal Ontology in Information Systems*. Amsterdam, Netherlands: IOS Press, pp. 3-15.

Hendler, J. (2001). Agents and the semantic web. *IEEE Intelligent Systems*, 16(2), pp. 30-36.
Highnam, R., & Brady, M. (1999). *Mammographic Image Analysis*. Kluwer Academic.
Holland J., (1976). Adaptation *in* R. Rosen and F. Snell *Progress in Theoretical Biology IV* Academic Press, pp.263-93.
Holland, J. (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press.
Holland, J. (1986). Escaping Brittleness: The Possibility of General-Purpose Learning Algorithms Applied to Rule-Based Systems. In: R.S.Mishalski, J.G. Carbonell, and T.M. Mitchell (eds). Machine Learning II. Kaufman, pp 593-623.
Holmes, J., Lanzi, P., Stolzmann, W., & Wilson, S., (2002). Learning Classifier Systems: New Models, Successful Applications. *Information Processing Letters,* 82 (1), pp. 23-30.
Jurisica, I., Mylopoulos, J., and Yu, E. (2004). Ontologies for knowledge management: An information systems perspective. *Knowledge and Information Systems-Springer-Verlag,* 6 (4), pp. 380-401.
Kharbat, F., (2006) *Learning Classifier Systems for Knowledge Discovery in Breast Cancer*, PhD thesis, University of the west of England, UK.
Lacroix, Z. and Critchlow, T. eds. (2003). *Bioinformatics: Managing scientific data*. Los Altos: Morgan Kaufmann.
Maedche A, Staab S. (2001). Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2).
Mangasarian, O., & Wolberg, H. (1990). Cancer Diagnosis via Linear-Programming *SIAM News,* 23 (5), pp 1-18.
Matsui, W., & Hopkins, J. (reviewers) (2002). Breast Cancer [online]. Available from: http://health.allrefer.com/health/breast-cancer.html
Milton, N., (2008*) Knowledge Technologies*. Polimetrica.
Noy, N. and McGuinness, D. (2001). *Ontology development 101: A guide to creating your first ontology*. Technical Report No. KSL-01-05, Stanford University.
Peńa-Reyes, C., & Sipper, M. (2000). Evolutionary Computation in Medicine: An Overview, *Artificial Intelligence in Medicine,* 19 (1), pp 1-23.
Pinto, H. S. and Martins, J. P. (2004). Ontologies: How can They be Built? Knowledge Information Systems. 6, 4,, pp. 441-464.
Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data and Knowledge Engineering,* 25, pp. 161-197.
Swartout, B., Ramesh, P., Knight, K., Russ, T.(1997). Toward Distributed Use of Large-Scale Ontologies, AAAI Symposium on Ontological Engineering, Stanford.
Uschold, M. and Grüninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(2), pp. 93-155.
Uschold, M. and Jasper, R. (1999). A framework for understanding and classifying ontology applications. *In: Proceedings of the IJCAI99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*, August 1999, Stockholm, Sweden. Available from http://sunsite.informatik.rwthaachen.de/Publications/CEUR-WS/Vol-18/.
Uschold, M., King, M. (1995). Towards a Methodology for Building Ontologies, in: IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal.
Van Heijst, G., Schreiber, A., and Wielinga, B. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, 46(2/3), pp. 183-292.
Wilson, S. (2001). Mining Oblique Data with XCS. In: Lanzi. P, Stolzmann. W, and S. Wilson (eds). *Advances in Learning Classifier Systems. Third International Workshop (IWLCS-2000),* pp 253-272, Berlin, Springer-Verlag .
Wilson, S., (1995). Classifier Fitness Based on Accuracy. *Evolutionary Computing*, 3, pp 149-175.

Data Mining in Medical and Biological Research

Edited by Eugenia G. Giannopoulou

ISBN 978-953-7619-30-5
Hard cover, 320 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

This book intends to bring together the most recent advances and applications of data mining research in the promising areas of medicine and biology from around the world. It consists of seventeen chapters, twelve related to medical research and five focused on the biological domain, which describe interesting applications, motivating progress and worthwhile results. We hope that the readers will benefit from this book and consider it as an excellent way to keep pace with the vast and diverse advances of new research efforts.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Faten Kharbat and Haya El-Ghalayini (2008). Building Ontology from Knowledge Base Systems, Data Mining in Medical and Biological Research, Eugenia G. Giannopoulou (Ed.), ISBN: 978-953-7619-30-5, InTech, Available from: http://www.intechopen.com/books/data_mining_in_medical_and_biological_research/building_ontology_from_knowledge_base_systems

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821