Construction automatique d'une ontologie de maladies par une approche descendante

Ihssen Belhadj[1, 2], Christian Jacquelinet [1, 2]

[1] Laboratoire d'informatique médicale et de bioinformatique, Faculté de médecine de l'université Paris 13, Bobigny,
[2] Agence de la biomédecine, Saint-Denis La Plaine, France
BelhadjIhssene(at)gmail.com

Abstract: This paper examines through an experimental approach the feasibility of an automatic disease ontology generation. The proposed method is based on a systematic specialization of concepts starting from a given model of knowledge and ontology of domain primitive attributes. The implementation of this method has permitted to generate test ontologies that were evaluated manually and by the use of Formal Concept Analysis auditing methods. This implementation has shown that it is possible to create automatically a multi-ontology concept hierarchy that highlights.

Keywords : Ontologies construction, top-down approach, disease concept

Résumé: Cet article examine expérimentalement la faisabilité d'une méthode de génération automatique d'une ontologie de maladies. La méthode proposée repose sur une spécialisation systématique des concepts à partir d'un modèle donné des connaissances et d'une ontologie des attributs primitifs. L'implémentation de cette méthode a permis de faire des générations d'ontologies qui ont été évalué manuellement et en utilisant des méthodes de Formal Concept Analysis. Cette implémentation a montré qu'il est possible de créer une ontologie de concept multi-hiérarchique selon différents points de vue à partir d'un modèle générique des connaissances d'un domaine et une ontologie de ses concepts primitifs.

Mots-clés : Construction d'ontologies, approche descendante, concept de maladie

.

1 Introduction

Agence de la biomédecine maintains a national information system (AB-IS) to support its missions concerning the evaluation of organ retrieval and transplantation activities, the management of the national waiting list and the epidemiology of end stage renal diseases in France (Strang et al, 2005. & Couchoud C et al, 2006., BenSaid et al, 2003). AB-IS comprises a contextual terminological server that was build according to sound ontological foundations (Jacquelinet et al 2003a., Jacquelinet et al, 2003b). A first component of this terminological server is an ontological editor that permits to describe the semantic structure of each new concept according to a knowledge model that is automatically inherited from its supertypes but manually specialized to provide the most accurate semantic structure.

A second component is a terminological resources manager that allows integrating any kind of terminological resources.

The need to provide ontology able to support the lateral integration of many terminologies emerged as a crucial requirement with the inclusion of the French ESRD REIN registry as a contributor the European ESRD registry using EDI technologies in the realm of the Nephro-Quest project (Jager & Zoccali, 2008**).** The new ERA-EDTA Thesaurus (ET) and the Thesaurus of the French Society of Nephrology (TSN) are both detailed terminologies but unfortunately organized around different axis so that the provision of smallest common supertypes for many concepts of the TSN that have no exact equivalent or no unique supertype. This led us to examine the feasibility of an automatic disease ontology generation by systematic specialization of concepts from a given knowledge model and according to the hierarchical structure of the attributes.

2 Materials et methods

2.1 Study framework

To generate concepts from a starting root concept, we use: The description knowledge model related to the root concept, the primitive pre-existing domain ontology that supports its constituents and all their subtypes, a set of specialization, differentiation and combining principles guiding the generation process.

2.2 Disease Description Model

In (Bertaud V, 2007) was analyzed how diseases are represented in existing terminological resources such as UMLS (Bodenreider, 2004). It has concluded that diseases are often defined as pathological entities resulting from pathological processes. This definition is of a little use once interesting to computerize diseases definitions because of the constant evolution of medical knowledge on pathological processes. To address this issue one can consider diseases as patterns of data

providing information about a patient that are useful for the diagnosis process. Hence, the question of defining a DDM can be dealt by the specification of a generic data structure considering that only discriminating knowledge is required to split, to abstract and to make a domain ontology.

In the Work of (Jacquelinet et al, 2003b) addressing the issue of retrieving from medical text information about patient, it was used conceptual graphs formalism (CGs) to represent diseases using a Description Model (DDM) as in Fig.1. Concepts related to terms are described with simple CGs that relates a concept to other concepts by the mean of conceptual relations.

[Disease]-
(Has_Location) → **[Anatomical_Entity_to_Specify]**
(Has_Lesion) → **[Lesion_to_Quantify]**
(Has_Evolution) → **[Evolution_to_Specify]**
…
(Has_Attribute_ n) → **[Discriminating attribute_n]**

Fig. 1 –Disease Description Model

Thus, as example, one can represent the concept of "Chronic renal disease" as bellow (Fig.2):

[Chronic Kidney Disease]
(has_Location) → **[Kidney]**
(has_Lesion) → **[Lesion to quantify]**
(has_Evolution) → **[Chronicity]**

Fig. 2 –Description Model of chronic renal disease concept

In this example, [Chronic Kidney Disease] is defined by three attributes: **a location** which is the kidney, a **lesion** that is not specified and an evolution that is **chronic**. The concepts [Rein], [Lesion to quantify], [Chronicity] constitute the attributes that **define** and **discriminate** the concept [Chronic Kidney Disease]. Each attribute refers to an elementary pre-existing concept with a fixed position in a predefined concepts hierarchy. This hierarchy should reflect the underlying reality through a domain theory (Bodenreider & Burgun, 2005). Moreover, one can build a space of attributes-based domain concepts by an automatic specialization of DKM.

2.3 Generative Principles

We assume here that the location of a given concept in a hierarchy is given by its semantic structure: the order relation between two concepts is derived according to specialization/generalization operations that permit to transform a semantic structure so as it defines a new resulting to a new concept (Rassinoux et al, 1992).

2.3.1 Specialization operation

Let C, C', two concepts defined by two graphs: **C: D → (RC) → C1 (1), C': D → (RC) → C2 (2).**

If **C1** is a specialization of **C2**, then **C'** is a specialization of **C**. We note our specialization rule: **C1 < C2 => C < C' (3).** Our seed knowledge model is defined to comprise the supertypes of all specialized attributes of its specializations. As a result, (3) is the sole specialization we need to use. The specialization of concepts is performed by the acquisition of specialized attribute. This principle defines a useful rule to organize a hierarchy that is lattice multiple inheritance.

2.3.2 Combinatory restriction rules

Let C be a root concept related to its DDM. **C: D – (RC1)→C1, (RC2)→C2**. We experimented 3 different combinatory rules. Combinatory rule 1 assumes that all the subtypes of C1 can be combined with the subtypes of C2 with no restriction. Combinatory rule 2 restricts the generation of concept to those combining terminal specializations of DDM attributes. More interestingly, we introduced the possibility to add a exclusivity property to certain subtypes of the attributes related to the DDM. The combinatory restriction rule 3 imposes that a specialization of the DDM defining a generated concept comprises a maximum of 1 exclusive differentiating attribute, all other combined attributes be non exclusive specializations.

2.3.3 Siblings Opposition Principles

In the Primitive Domain ontology, the hierarchy of KM sub-types comprises intermediate concepts that support the siblings opposition principle and its unicity according to (Zweigenbaum et al, 1995). For example:

Existing Etiological_Process
  Etiological Process according to its acknowledgement <e>
    Unknown Etiological Process
    Known Etiological Process <e>
      Unspecified Etiological Process
      Specified Etiological Process <e>
        Etiological_Process according to its type <e>
          Hypofunction
          Dysfunction
          Hyperfunction
            Hyperfunction according to Anatomical Entity <e>
              Cardiac HyperFunction
              …….
Lesion to Quantify <e>
  Unquantified Lesion
  Quantified Lesion <e>
    Unexisting Lesion
    Existing Lesion
With combinatory restriction rule 1, we will generate some: Disease-
  (Has Evolution)→ [Quantified Lesion] <e>

(has Etiology)→ [Specified Etiological Process] <e>

Such a Concept will not exist with combinatory restriction rule 3 that will impose to specialize exclusive opposition principle weaving concepts before to hybridize with another exclusive opposition principle wearing concept.

2.4 Auditing Method

Formal Concept Analysis (FCA) can be described as a method for attributes exploration (Granter & Wille, 1999). It has been applied for building, completion and evaluation of ontologies (Baader et al, 2007, Jiang et al, 2003, Jiang et al, 2009a.)

The work (Jiang et al, 2009b) describes an FCA auditing method that inspired us to adapt it to the framework of our study. In fact, FCA can be used to represent semantically a concept definition by formal contexts which could be visualized by a lattice diagram.

The figure 3, the lattice diagram corresponding to this context is illustrated in the figure below. Therefore, we thought to make possible to convert the whole and/or part of generated concepts in a FCA formal one valued context.

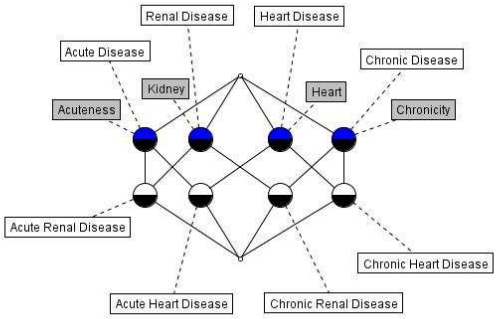

Fig. 3 –Using FCA for ontology auditing

The objects in FCA will refer to the generated objects and the attributes to all of the relation signatures involving the concept except subsumption relations. In other terms we hided "is-a" relations so that the formal context refers to only to its attributes relations. Otherwise, applying these methods required to develop accessorily a module that allowing exporting generated concept to an XML file compatible with Concept Explore (http://conexp.sourceforge.net/) software implementing FCA toolbox.

2.5 Experimentation

2.5.1 Manual Building of attributes ontology

Using protégé, 3 sample primitive ontologies were built, ranging from a small to larger coverage domain ontology built in group involving epidemiologists and medical computer scientists by abstracting definitions of sound disease related concepts. These sample ontologies served as test input for the generation algorithm.

The simplest sample input ontology comprised 21 concepts according to three axes [Anatomical location], [Evolution], [Lesion].

The DDM was built with Existential quantifiers of disease concept. The use of the input and/or output ontology required the development of an import / export module to and from OWL format ontology. This enables our algorithm to use a large panel of Protégé tools and plug-ins.

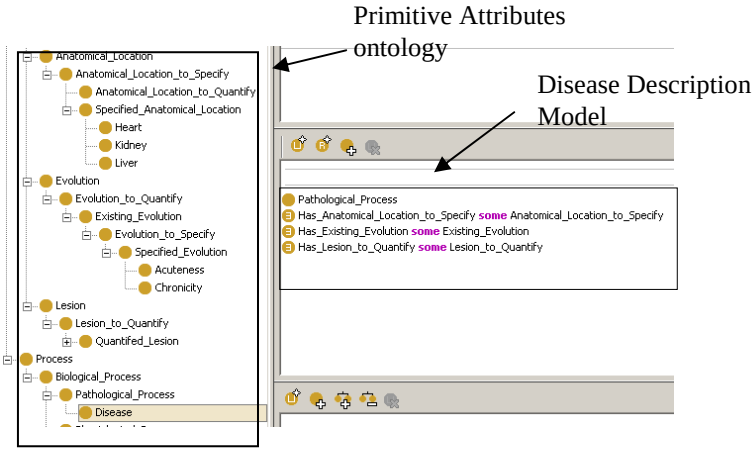

Fig. 4 –The input model built with Protégé

The exploitation of this ontology is done via an ad-hoc database format supported by a MySQL database system. In this format, a main table contains a list of all concepts while another contains signatures of relations instances according to the following formalism: (head Concept-Relation- tail Concept). The name of each concept is standardized by a module which generates its name based on its attributes. This allows a unique identification of concepts and to request about duplicate concepts or supertypes.

2.5.2 Implementation of generation method

The design of the implementation of our method was based on the idea that every concept shares multiple points of view that are determined by its structure and need to be saturated by attributes. The saturation process is done by acquisition of new attributes, leading to a concept hierarchy alternating concepts with abstract attributes wearing the differentiating points of view and concepts whose attributes are direct and concrete specializations of the differentiating points of view. To illustrate the generation process, we can by summarizing the model described in figure 2, make a first level generation: **[Diseases according to its evolution], [Diseases according to lesion to quantify], [Diseases according to its location].**We can for example start

from [Kidney Disease], a concrete specialization of **[Diseases according to its location].** kidney disease according to their evolution, we can make a second level generation as showed in figure 1. This generation use **[Specified Evolution]** abstract sub-hierarchy to generated new concepts. Therefore, generated concepts starting from **[Disease According To Its Evolution]** are: **[Chronic Kidney Disease] and [chronic Kidney Disease]**. These concepts saturate **[Specified evolution]** point of view. This concept itself will become a genius and tend to make subtypes because there still exist points of view that are not saturated. In this example, at this step, an unsaturated point of view is **[Lesion to quanitfy].**

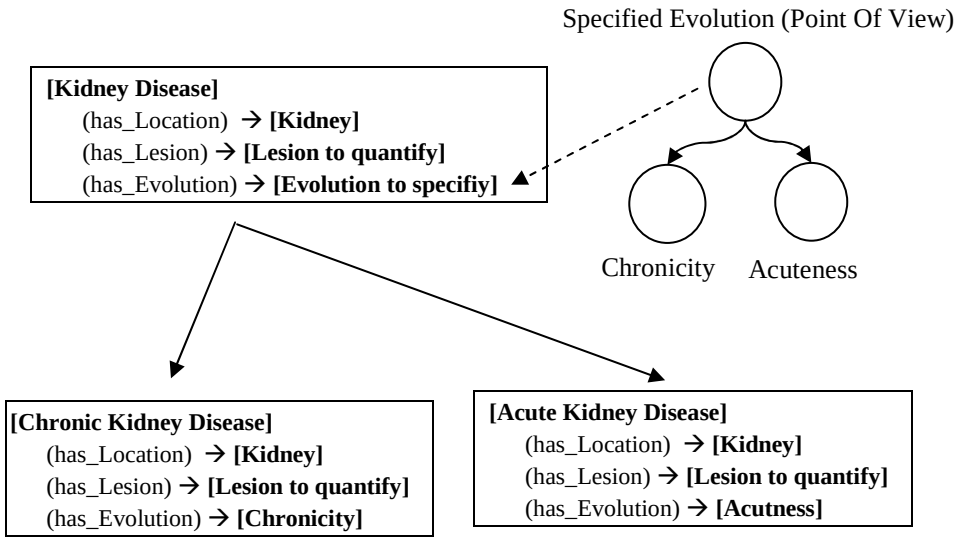

Fig. 5 –New concepts generation by specialization

Every time a new attribute is added, a new concept is generated after verifying it complies with integrity rules we can personalize by adding conditional clauses. The concept is then added to knowledge base and the new acquired attribute is instantiated as a new relation signature.

And so on, the algorithm still runs while unsaturated concepts are still existing and until deploying all of the points of view. The visualization of generated ontology is done by using a Java interface specially developed for this purpose. It helps to browse the subsumption hierarchy of concepts as well as viewing attributes, supertypes and subtypes.

3 Results

A first test permitted to generate a total of 157 concepts starting from a primitive conceptual hierarchy of 21 concepts and a DDM including three attributes. In this test, we examined manually a sample of subsumption relationships between the generated ontology and the lattice diagram. All the relations examined in FCA are concordant with the generated ontology subsumption relations. A first result is that FCA identifies automatically subsumption relations.

A second test aimed to observe the combinatory explosion while increasing each time a primitive a more finely conceptual hierarchy. The last one comprises 110 concepts and generated ontology of 12000 concepts according to a DDM including four attributes.

A third test consisted of repeating the previous test but after adding clauses referring to restriction rules defined in the method above. We can say, then, that combinatory restrictions rules reduced considerably the combinatory explosion. Thinner

4 Discussion – Conclusion

We have described and implemented a method of automatic generation of diseases definitions ontology based on a systematic combination of attributes and exclusively by a top-down approach. The implementation of this method showed that it is possible to create a multi-hierarchical concept ontology according different views from a starting domain concepts model and ontology of primitive domain concepts. This can be of a major methodological interest in building contextual domain reference ontologies. Future works will focuse not only on the generation method but also on the quality and usability of the generated ontology.

References

BAADER F., GANTER B., SATTLER U. & SERKAYA B. (2007) Completing description logic knowledge bases using formal concept analysis, Proceedings of the twentieth international joint conference on artificial intelligence (IJCAI-07), AAAI Press (2007).

BEN SAID, M., SIMONET, A., GUILLON, D., JACQUELINET, C., GASPOZ, F., DUFOUR, E., MUGNIER, C., JAIS, J.P., SIMONET, M. & LANDAIS, P. (2003) A dynamic Web application within an n-tier architecture: a Multi-Source Information System for end-stage renal disease. Stud Health Technol Inform. 2003; 95:95-100.**)**

BERTAUD V., BELHADJ I., DAMERON O., GARCELON N., HENDAOUI L., MARIN F. & DUVAUFERRIER R. (2006) Computerizing the radiological sign. J Radiol 2007; 88:27-37.

BODENREIDER O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(database issue):D267–D270.

BODENREIDER O., BURGUN A. (2005). Biomedical ontologies. In: Chen H, Fuller S, Friedman C, Hersh W, editors. Medical informatics: Knowledge management and data mining in biomedicine, Springer, 2005, p 211–235.

COUCHOUD, C., STENGEL, B., LANDAIS, P., ALDIGIER, J.C., DE CORNELISSEN, F., DABOT, C., MAHEUT, H., JOYEUX, V., KESSLER, M., LABEEUW, M., ISNARD, H. & JACQUELINET, C. (2006)

The renal epidemiology and information network (REIN): a new registry for end-stage renal disease in France. *Nephrol Dial Transplant* 2006 Feb: 21(2):411-8.**).**

GRANTER B & WILLE R. (1999) Formal concept analysis: mathematical foundations. Berlin: Springer; 1999: ISBN: 3-540-62771-5.

JACQUELINET C, BURGUN A, DELAMARRE D, STRANG WN, DJABBOUR S, BOUTIN B, LE BEUX P. (2003) Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation. Int J Med Inform. Jul; 70(2-3):317-28.

JACQUELINET C, BURGUN A, DJABBOUR S, DELAMARRE D, CLERC P, BOUTIN B, LE BEUX P. (2003). A contextual coding system for transplantation and end stage diseases. *Stud Health Technol Inform* 2003:95:457-62.

JAGER, K.J. & ZOCCALI, C. (2008). Quality of care in end-stage renal disease: the importanceof comparing 'apples with apples'. *Nephrol Dial Transplant*; 23: 1116**.)**

JIANG G. & CHUTE CG. (2009a) Auditing the semantic completeness of the SNOMED CT using formal concept analysis, J Am Med Inform Assoc 16 (1) (2009), pp. 89–102. JIANG G., OGASAWARA K., ENDOH A. & SAKURAI T. (2003) Context-based ontology building support in clinical domains using formal concept analysis, Int J Med Inform 71 (1) (2003), pp. 71–81.

JIANG G., PATHAK J. & CHUTE CG. (2009b) Formalizing ICD coding rules using Formal Concept Analysis, J Biomed Inform. 2009 Jun; 42(3):504-17.

RASSINOUX A M, BAUD RH. & SCHERRER JR. Conceptual Graph model extension for knowledge representation of medical texts. In Proceedings of MEDINFO 92. Lun KC, Degoulet P.0, Piemme TE. & Reinhoff O, eds, Elsevier, North-Holland Publ Comp, Amsterdam, 1992; 1368-74.

STRANG, W.N., TUPPIN, P., ATINAULT, A. & JACQUELINET, C. The French Organ Transplant Data System. *Stud Health Technol Inform* 2005: 116:77-82.

ZWEIGENBAUM P., BACHIMONT B., BOUAUD J., CHARLET J. & BOISVIEUX J.F,. (1995) Issues in Structuring and Acquisition of an Ontology for Medical Language Understanding. Meth Inf Med. 1995; 34:15-24.