Informative Common Subsumers for Diseases Diagnosis

Simona Colucci
(Politecnico di Bari, Bari, Italy and
D.O.O.M. s.r.l., Matera, Italy
s.colucci@poliba.it)

Eugenio Di Sciascio
(Politecnico di Bari, Bari, Italy
disciascio@poliba.it)

Francesco M. Donini
(Università della Tuscia, Viterbo, Italy
donini@unitus.it)

Marcella Mastronardi
(Policlinico di Bari, Bari, Italy
marcellamastronardi@yahoo.it)

Abstract: This paper proposes an approach for automatically extracting symptoms associated to a given disease from semantic-based descriptions of health records of patients affected by an investigated pathology. The proposal implements non-standard reasoning services developed in Description Logics for the individuation of informative commonalities in concept collections and can make significantly easier the diagnosis process of rare and unknown diseases.

Key Words: Common Subsumers, Healthcare, Description Logics.
Category: M.4, M.7, H.3.3

1 Introduction

Healthcare is one of the knowledge domains in which semantic technologies have been traditionally most widely employed in literature. Such a domain is in fact characterized by a large amount of available information whose sharing is crucial in several circumstances.

Of course a multitude of information systems exist at local level to manage available data, with a consequent heterogeneity in data representation, which makes hard the integration process. By adopting ontological structures conveying a shared interpretation on the meaning surrounding the available data, a partial integration may be reached.

Most ontologies developed for healthcare domain aim at modeling human anatomy and known pathologies. GALEN [Rector et al., 1994] is probably the most known general ontology developed to provide a common terminology in

the medical domain. UMLS [Bodenreider, 2004] and SNOMED[1] represent further examples of ontologies working as vocabularies for healthcare. Several other approaches are focused on the proposal of semantic annotation schemes for clinical texts and health records ([Bodenreider, 2004], [Sheth et al., 2006]), in order to provide interoperability and information sharing. Nevertheless, only a few approaches cope with the issue of relating diseases with their symptoms in order to support diagnosis([Hadzic and Chang, 2005], [Minchin et al., 2006]).
It is also noteworthy that the new HL7[2] version has been somehow designed with an ontological formalization.

In this paper we present a semantic-based approach for the automated individuation of factors commonly related to a known disease. The approach exploits non-standard inference services [Colucci et al., 2008] proposed in Description Logics [Baader et al., 2003] to automatically extract shared characteristics in a set of concept descriptions. In particular, the approach uses patients health records modeled according to an ontology describing possible symptoms, treatments or behavioral factors affecting patients health. Such records represent the knowledge base to be investigated for discovering factors common to patients sharing a diagnosis. In other words health records of patients affected by the same disease work as training set for learning how to diagnose known pathologies. Obviously, the more the disease is rare and difficult to diagnose, the more our approach is helpful in suggesting affecting factors both explicitly recorded in health records and implied by elicited information.

The validity of the approach has been tested with real data collected in the Pediatric Surgery ward of the Policlinico general hospital in Bari, Italy. At the first stage of analysis, the health records selected to validate the approach are related to patients affected by appendicitis.

The paper is organized as follows: in the next section we shortly provide the background of the semantic-based approach we propose here. Section 3 shows instead how to implement the proposed approach for supporting diagnosis in a real data framework. Finally, conclusions close the paper.

2 Formalism and Inferences

The proposed approach models information in patients health records according to Description Logics (DLs) formalism, in order to take advantage from useful non-standard reasoning services specifically developed for the extraction of commonalities in collections of concept descriptions in DL. Due to the lack of space, we do not delve into details about DL formalism and standard reasoning (the interested reader may find them in [Baader et al., 2003]), but we shortly recall

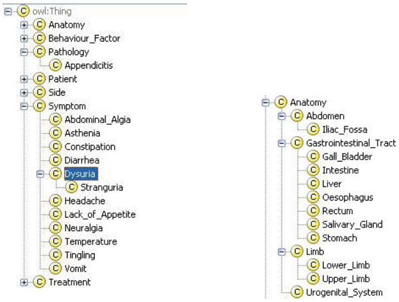
[1] http://www.snomed.org
[2] http://www.hl7.org/

only non-standard services fundamental in our approach. In particular, Least Common Subsumer (LCS) was defined [Cohen and Hirsh, 1994] to find features shared by *all* of the elements of a given collection.

In order to deal with the problem of extracting commonalities in a *portion* of elements in a set, in [Colucci et al., 2008] four different common subsumers were proposed:
- **k-Common Subsumers**(k-CSs) formally represent characteristics shared by k concepts in a collection of n elements (with $k < n$);
- **Informative k-Common Subsumer** (IkCS) are defined as k-CSs adding informative content to the LCS computation, given that LCSs are also k-CSs, for every $k < n$;
- **Best Common Subsumers** (BCSs) represent features shared by the maximum number of elements in a collection, and makes sense when such a maximum is less than n(*i.e.*, when the LCS is equivalent to the universal concept);
- **Best Informative Common Subsumers** (BICS) are defined as BCSs adding informative content to the LCS computation, given that, when the LCS is not equivalent to the universal concept, it is of course the BCS of a collection.

3 Illustrative Example

In this section we show how to exploit inference services proposed in Section 2 to automatically extract commonalities in health records of patients sharing a final diagnosis. In particular we employed anonymous health records belonging to a sample of 100 patients affected by appendicitis hospitalized in the Pediatric Surgery ward of the general hospital in Bari, Italy. In order to show how the proposed approach works, we provide the data related to three patients in the sample, whose clinical status is introduced in the following:
Patient 1: Female, 12 years old, feeling abdominal algia to right iliac fossa, constipation, lower right limb neuralgia, asthenia, temperature and dysuria;
Patient 2: Female, 7 years old, feeling abdominal algia to right iliac fossa, alternate phenomena of constipation and diarrhea, vomit and temperature;
Patient 3: Male, 5 years old, feeling abdominal algia to right iliac fossa, stranguria, lack of appetite, constipation, headache, limb tingling and temperature.
We notice that the three patients have not behavioral factors and do not take treatments affecting health because of their young age, but w.l.o.g. we considered also such affecting elements in our modeling. In order to provide the vocabulary for health records description, we developed an ontology in the DL $\mathcal{ALEN}$ [Baader et al., 2003]; an excerpt of the main hierarchy is shown in Figure 1(a), together with the hierarchy of a specific class, `Anatomy`, in Figure 1(b). The ontology axioms involved in the representation of the three patients taken as case study are instead listed below:

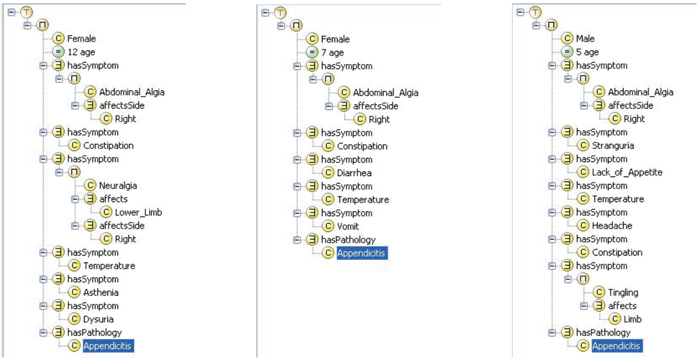


```
owl:Thing
  Anatomy
  Behaviour_Factor
  Pathology
    Appendicitis
  Patient
  Side
  Symptom
    Abdominal_Algia
    Asthenia
    Constipation
    Diarrhea
    Dysuria
      Stranguria
    Headache
    Lack_of_Appetite
    Neuralgia
    Temperature
    Tingling
    Vomit
  Treatment
```

```
Anatomy
  Abdomen
    Iliac_Fossa
  Gastrointestinal_Tract
    Gall_Bladder
    Intestine
    Liver
    Oesophagus
    Rectum
    Salivary_Gland
    Stomach
  Limb
    Lower_Limb
    Upper_Limb
  Urogenital_System
```

(a) (b)

Figure 1: (a) Ontology Main Hierarchy (b)Anatomy Class Hierarchy



```
⊤
  ⊓
    Female
    12 age
    ∃ hasSymptom
      ⊓
        Abdominal_Algia
        ∃ affectsSide
          Right
    ∃ hasSymptom
      Constipation
    ∃ hasSymptom
      ⊓
        Neuralgia
        ∃ affects
          Lower_Limb
        ∃ affectsSide
          Right
    ∃ hasSymptom
      Temperature
    ∃ hasSymptom
      Asthenia
    ∃ hasSymptom
      Dysuria
    ∃ hasPathology
      Appendicitis
```

```
⊤
  ⊓
    Female
    7 age
    ∃ hasSymptom
      ⊓
        Abdominal_Algia
        ∃ affectsSide
          Right
    ∃ hasSymptom
      Constipation
    ∃ hasSymptom
      Diarrhea
    ∃ hasSymptom
      Temperature
    ∃ hasSymptom
      Vomit
    ∃ hasPathology
      Appendicitis
```

```
⊤
  ⊓
    Male
    5 age
    ∃ hasSymptom
      ⊓
        Abdominal_Algia
        ∃ affectsSide
          Right
    ∃ hasSymptom
      Stranguria
    ∃ hasSymptom
      Lack_of_Appetite
    ∃ hasSymptom
      Temperature
    ∃ hasSymptom
      Headache
    ∃ hasSymptom
      Constipation
    ∃ hasSymptom
      ⊓
        Tingling
        ∃ affects
          Limb
    ∃ hasPathology
      Appendicitis
```

Figure 2: Patients Health Record Descriptions

$\texttt{Abdominal\_Algia} \sqsubseteq \texttt{Symptom} \sqcap \forall\texttt{affects.Iliac\_Fossa}$

$\texttt{Constipation} \sqsubseteq \texttt{Symptom} \sqcap \forall\texttt{affects.Gastrointestinal\_Tract}$

$\texttt{Vomit} \sqsubseteq \texttt{Symptom} \sqcap \forall\texttt{affects.Gastrointestinal\_Tract}$

$\texttt{Diarrhea} \sqsubseteq \texttt{Symptom} \sqcap \forall\texttt{affects.Gastrointestinal\_Tract}$

$\texttt{Dysuria} \sqsubseteq \texttt{Symptom} \sqcap \forall\texttt{affects.Urogenital\_System}$

$\texttt{Stranguria} \sqsubseteq \texttt{Dysuria} \sqcap \exists\texttt{causesPain.}$

Figure 2 shows the modeling of the three patients health records according to the proposed ontology.

In order to automatically individuate symptoms usually shared by patients affected by appendicitis, we employ a general approach originally proposed in [Colucci et al., 2008] for extracting commonalities in concept collections.

Of course, we are not only interested in "obvious" symptoms, but the effort in modeling knowledge related to patients health records should be exploited to discover unknown associations between symptoms and pathologies. In other words, our approach aims at finding symptoms that are non necessarily felt by *all* of the patients affected by a given pathology, but also by a *relevant* number of patients.

Notice that the concept descriptions representing health records use the full expressiveness of $\mathcal{ALEN}$. In [Colucci et al., 2008] we proposed an algorithm for computing partial common subsumers in $\mathcal{ELHN}$, showing also its extensibility to $\mathcal{ALEN}$. The algorithm relies on the computation of the LCS proposed in [Küsters and Molitor, 2005]. In this paper we instead adopt the novel tableaux-based method for LCS computation proposed in [Donini et al., 2009], which extends to the full $\mathcal{ALEHIN}_{R^+}$. In the following we show the results of common subsumers computation on the collection of concept descriptions S={**Patient 1, Patient 2, Patient 3**}, commenting on their impact on disease diagnosis. First of all, we compute the LCS of the collection, representing the portion of health records common to all of the three patients:

$L = LCS(S) = \exists\mathtt{hasPathology.Appendicitis} \sqcap \exists\mathtt{hasSymptom.(Abdominal\_Algia} \sqcap \exists\mathtt{affectsSide.Right)} \sqcap \exists\mathtt{hasSymptom.Constipation} \sqcap \exists\mathtt{hasSymptom.Temperature}$

Of course, LCS represents symptoms which are well known for appendicitisbut we are interested in less common symptoms, frequently associated to appendicitis; to this aim we search for portions of health records shared by at least two patients out of the three in the collection and start computing the set $\underline{CS}_2$ of 2-CSs of the collection.

$\underline{CS}_2 = \{L,\ L \sqcap \mathtt{Female},\ L \sqcap \exists\mathtt{hasSymptom.(}\exists\mathtt{affects.Limb)} \sqcap \exists\mathtt{hasSymptom.Dysuria}\}$

Among the three retrieved 2-CS, we are particularly interested in the ones adding informative content to the LCS, which we called I2CS. The set $\underline{ICS}_2$ of such concepts is given in the following:

$\underline{ICS}_2 = \{\ L \sqcap \mathtt{Female},\ L \sqcap \exists\mathtt{hasSymptom.(}\exists\mathtt{affects.Limb)} \sqcap \exists\mathtt{hasSymptom.Dysuria}\}$

Given that the LCS of the collection is not equivalent to the universal concept, computing the set of BCSs makes no sense(see Section 2). The set of BICSs (see Section 2) is instead equivalent to $\underline{ICS}_2$: if one more patient is added to the set of those sharing features expressed by I2CS, the latter necessarily reverts to the LCS of the collection.

The proposed tiny case study shows how, thanks to the computation of Informative Common Subsumers, we discover that two patients out of three are affected by some symptoms to the limbs and dysuria. We also found out that two patients are female. Even though introduced here in a limited set of three health records for the sake of simplicity, our approach has been evaluated with reference to a sample of 100 patients. Produced results confirm *e.g.*, that dysuria and symptoms to limbs affect a significant portion of patients affected by appendicitis. On the contrary, the prevalence of female patients has not been retrieved in results from the general sample (intuitively, such a not significant result is due to the very limited arity of the case study set and to the binary nature of patients sex information).

4 Conclusions

We have presented an approach supporting diagnosis through a semantic-based individuation of symptoms shared by a significant portion of patients affected by the disease to diagnose using informative common subsumers. We implemented our approach in a case study employing real data related to patients affected by appendicitis and the results have been validated by pediatric surgeons working in the general hospital of Bari. Even though appendicitis is a very frequent disease, apparently not needing support to diagnosis, we chose it to test our approach, in order for the evaluation to be free from interpretation faults. Nevertheless, our proposal is designed to make easier the diagnosis of rare and not well known diseases: initial difficulties in recognizing such pathologies may be overcome through a backward investigation on health records of affected patients and this is the objective of ongoing investigations.

Acknowledgments

We wish to acknowledge partial support of PS_092 and PS_121.

References

[Baader et al., 2003] Baader, F., Calvanese, D., Mc Guinness, D., Nardi, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook*. CUP.

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

[Cohen and Hirsh, 1994] Cohen, W. and Hirsh, H. (1994). Learning the CLASSIC description logics: Theorethical and experimental results. In *Proc. of KR'94*.

[Colucci et al., 2008] Colucci, S., Di Sciascio, E., Donini, F. M., and Tinelli, E. (2008). Finding informative commonalities in concept collections. In *Proc. of CIKM 2008*.

[Donini et al., 2009] Donini, F. M., Colucci, S., Di Noia, T., and Di Sciascio, E. (2009). A tableaux-based method for computing least common subsumers for expressive description logics. In *Proc. of IJCAI 2009*. AAAI. to appear.

[Hadzic and Chang, 2005] Hadzic, M. and Chang, E. (2005). Ontology-based support for human disease study. In *HICSS-2005*. IEEE Computer Society.

[Küsters and Molitor, 2005] Küsters, R. and Molitor, R. (2005). Structural Subsumption and Least Common Subsumers in a Description Logic with Existential and Number Restrictions. *Studia Logica*, 81:227–259.

[Minchin et al., 2006] Minchin, R., Porto, F., Vangenot, C., and Hartmann, S. (2006). Symptoms ontology for mapping diagnostic knowledge systems. *Computer-Based Medical Systems, IEEE Symposium on*, 0:593–598.

[Rector et al., 1994] Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., and Rossi-Mori, A. (1994). The GALEN CORE model schemata for anatomy: Towards a reusable application-independent model of medical concepts. In *MIE-94*.

[Sheth et al., 2006] Sheth, A., Agrawal, S., Lathem, J., Oldham, N., Wingate, H., Yadav, P., and Gallagher, K. (2006). Active semantic electronic medical record. In *ISWC-06*, volume 4273 of *LNCS*, pages 913–926. Springer.