Refining Ontologies via Pattern-based Clustering

Francesca A. Lisi and **Floriana Esposito** [1]

Abstract. In this paper we consider the problem of finding subconcepts of a known concept (reference concept) in a given ontology in the light of new knowledge coming from a data source. These subconcepts are discovered by looking for frequent association patterns between the reference concept and other concepts also occurring in the existing ontology. As an illustration, we report preliminary results obtained from the refinement of an $\mathcal{ALC}$ ontology with respect to DATALOG data extracted from the on-line CIA World Fact Book.

1 INTRODUCTION

Ontology Refinement is the adaptation of an existing ontology to a specific domain or the needs of a particular user [8]. In this paper we consider the problem of finding subconcepts of a known concept C_{ref}, called *reference concept*, in the existing ontology Σ in the light of new knowledge coming from a data source Π. We assume that a *concept* $\mathcal{C}$ consists of two parts: an *intension* $int(\mathcal{C})$ and an *extension* $ext(\mathcal{C})$. The former is an expression belonging to a logical language $\mathcal{L}$ whereas the latter is a set of objects that satisfy the former. More formally, given

- a reference concept $C_{ref} \in \Sigma$,
- a data set $\mathbf{r} = \Sigma \cup \Pi$,
- a language $\mathcal{L}$

our Ontology Refinement problem is to find a directed acyclic graph (DAG) $\mathcal{G}$ of concepts $\mathcal{C}_i$ such that (i) $int(\mathcal{C}_i) \in \mathcal{L}$ and (ii) $ext(\mathcal{C}_i) \subset ext(C_{ref})$. Note that C_{ref} is among both the concepts defined in Σ and the symbols of $\mathcal{L}$. Furthermore $ext(\mathcal{C}_i)$ relies on a notion of satisfiability of $int(\mathcal{C}_i)$ w.r.t. **r**. Note that **r** includes Σ because in Ontology Refinement, as opposite to other forms of Ontology Learning such as Ontology Extraction (or Building), it is mandatory to consider the existing ontology and its existing connections.

A Knowledge Representation and Reasoning (KR&R) framework that turns out to be suitable for our problem is the one offered by the *hybrid system* $\mathcal{AL}$-log [2]. It allows for the specification of both relational and structural data: the former is based on DATALOG [1], the latter on $\mathcal{ALC}$ [11]. The integration of the two logical formalisms is provided by the so-called constrained DATALOG clause, i.e. a DATALOG clause with variables possibly constrained by concepts expressed in $\mathcal{ALC}$. Within this KR&R framework, the data set **r** is represented as a $\mathcal{AL}$-log knowledge base $\mathcal{B}$ and the language $\mathcal{L}$ contains expressions, called $\mathcal{O}$*-queries*, of the form

$$Q = q(X) \leftarrow \alpha_1, \ldots, \alpha_m \& X : C_{ref}, \gamma_1, \ldots, \gamma_n,$$

relating individuals of C_{ref} to individuals of other concepts (*task-relevant concepts*) also occurring in Σ. Thus, for a concept $\mathcal{C}$, $int(\mathcal{C})$ is an $\mathcal{O}$-query $Q \in \mathcal{L}$ and $ext(\mathcal{C})$ is the set $answerset(Q, \mathcal{B})$ of correct answers to Q w.r.t. $\mathcal{B}$. The DAG $\mathcal{G}$ is structured according to the *subset relation* between concept extensions.

The problem in hand can be considered as a case of tha form of unsupervised learning, known under the name of *Conceptual Clustering* [10], that aims at determining not only the clusters but also their descriptions expressed in some representation formalism. As a solution approach to the problem, we follow a recent trend in Cluster Analysis: using *frequent (association) patterns* as candidate clusters [13]. A frequent pattern is an intensional description, expressed in a language $\mathcal{L}$, of a subset of a given data set **r** whose cardinality exceeds a user-defined threshold (*minimum support*). Note that patterns can refer to multiple levels of description granularity (*multi-grained patterns*). In any case, a frequent pattern highlights a regularity in **r**, therefore it can be considered as the clue of a data cluster. In the context of Ontology Refinement these clues are called *emerging concepts* because they are concepts whose only extension is determined. In [4] it has been proposed to extend [6] in order to provide a pattern-based approach to Conceptual Clustering.

The paper is organized as follows. Section 2 illustrates our approach to the problem. Section 3 reports a preliminary empirical evaluation of the approach. Section 4 concludes with final remarks and directions of future work.

2 PATTERN-BASED CLUSTERING

When faced with a pattern-based approach to Conceptual Clustering, the Ontology Refinement problem stated in Section 1 is decomposed in two subproblems:

I. discovery of frequent patterns in data
II. generation of clusters from frequent patterns

In particular, the subproblem I is actually a variant of frequent pattern discovery which aims at obtaining descriptions of the data set **r** at different levels of granularity [3]. Here **r** typically encompasses a taxonomy $\mathcal{T}$. More precisely, the problem of *frequent pattern discovery at* l *levels of description granularity*, $1 \leq l \leq maxG$, is to find the set $\mathcal{F}$ of all the frequent patterns expressible in a multi-grained language $\mathcal{L} = \{\mathcal{L}^l\}_{1 \leq l \leq maxG}$ and evaluated against **r** w.r.t. a set $\{minsup^l\}_{1 \leq l \leq maxG}$ of minimum support thresholds by means of the evaluation function $supp$. In this case, $P \in \mathcal{L}^l$ with support s is frequent in **r** if (i) $s \geq minsup^l$ and (ii) all ancestors of P w.r.t. $\mathcal{T}$ are frequent in **r**.

The method proposed for solving one such decomposed problem extends the *levelwise search* method [9] for frequent pattern discovery with an additional post-processing step to solve the subproblem II. This method searches the space $(\mathcal{L}, \succeq)$ of patterns organized according to a generality order $\succeq$ in a breadth-first manner, starting

[1] Dipartimento di Informatica, Università degli Studi di Bari, Via Orabona 4, 70125 Bari, Italy, email: {lisi, esposito}@di.uniba.it

from the most general pattern in $\mathcal{L}$ and alternating candidate generation and candidate evaluation phases. The underlying assumption is that $\succeq$ is a quasi-order monotonic w.r.t. *supp*. For $\mathcal{L}$ being a multi-grained language of $\mathcal{O}$-queries, *supp* supplies the percentage of individuals of C_{ref} that satisfy an $\mathcal{O}$-query Q and $\succeq$ is based on the $\mathcal{B}$*-subsumption* relation [6]. It has been proved that $\succeq_\mathcal{B}$ is a quasi-order that fulfills the condition of monotonicity w.r.t. *supp* [6]. Also the search for patterns is depth-bounded (up to $maxD$).

The subproblem II concerns choosing a description for each cluster. In [5] it has been proposed a criterion obtained by combining two orthogonal biases: a language bias and a search bias. The language bias allows the user to define conditions on the form of $\mathcal{O}$-queries to be accepted as concept intensions. In particular, it is possible to state which is the minimum level of description granularity and whether (all) the variables must be ontologically constrained or not. The search bias allows the user to define a preference criterion based on $\mathcal{B}$-subsumption. In particular, it is possible to state whether the *most general description (m.g.d.)* or the *most specific description (m.s.d.)* w.r.t. $\succeq_\mathcal{B}$ has to be preferrred. Since $\succeq_\mathcal{B}$ is not a total order, it can happen that two patterns P and Q, belonging to the same language $\mathcal{L}$, can not be compared w.r.t. $\succeq_\mathcal{B}$. In this case, the m.g.d. (resp. m.s.d) of P and Q is the union (resp. conjunction) of P and Q.

Note that this method for Conceptual Clustering is *top-down* and *incremental* due to the features of the levelwise search. Also it is not hierarchical because it returns a DAG instead of a tree of concepts.

3 PRELIMINARY EXPERIMENTS

As an illustration, we report the results of four experiments conducted on the $\mathcal{AL}$-log knowledge base $\mathcal{B}_{\texttt{CIA}}$ that has been obtained by adding DATALOG facts[2] extracted from the on-line 1996 CIA World Fact Book[3] to an $\mathcal{ALC}$ ontology $\Sigma_{\texttt{CIA}}$ concerning the concepts Country, EthnicGroup, Language, and Religion. The parameter settings are: C_{ref} = MiddleEastCountry, $maxD = 5$, $maxG = 3$, $minsup^1 = 20\%$, $minsup^2 = 13\%$, and $minsup^3 = 10\%$. Thus each of them started from the same set $\mathcal{F}$ of 53 frequent patterns out of 99 candidate patterns.

Case for $l \geq 2$. The first two experiments both require the descriptions to have all the variables ontologically constrained by concepts from the second granularity level on. When the m.g.d. criterion is adopted, the procedure of graph building returns the following twelve concepts:

$\mathcal{C}'_0 \in \mathcal{F}^1_1$
q(A) ← A:MiddleEastCountry
{ARM, BRN, IR, IRQ, IL, JOR, KWT, RL, OM, Q, SA, SYR, TR, UAE, YE}

$\mathcal{C}'_1 \in \mathcal{F}^2_3$
q(A) ← believes(A,B) &
A:MiddleEastCountry, B:MonotheisticReligion
{ARM, BRN, IR, IRQ, IL, JOR, KWT, RL, OM, Q, SA, SYR, TR, UAE}

$\mathcal{C}'_2 \in \mathcal{F}^2_3$
q(A) ← speaks(A,B) &
A:MiddleEastCountry, B:AfroAsiaticLanguage
{IR, SA, YE}

$\mathcal{C}'_3 \in \mathcal{F}^2_3$
q(A) ← speaks(A,B) &
A:MiddleEastCountry, B:IndoEuropeanLanguage
{ARM, IR}

$\mathcal{C}'_4 \in \mathcal{F}^2_5$
q(A) ← speaks(A,B), believes(A,C) &
A:MiddleEastCountry,
B:AfroAsiaticLanguage, C:MonotheisticReligion
{IR, SA}

$\mathcal{C}'_5 \in \mathcal{F}^2_5$
q(A) ← believes(A,B), believes(A,C) &
A:MiddleEastCountry,
B:MonotheisticReligion, C:MonotheisticReligion
{BRN, IR, IRQ, IL, JOR, RL, SYR}

$\mathcal{C}'_6 \in \mathcal{F}^3_3$
q(A) ← believes(A,'Druze') & A:MiddleEastCountry
{IL, SYR}

$\mathcal{C}'_7 \in \mathcal{F}^3_3$
q(A) ← believes(A,B) &
A:MiddleEastCountry, B:JewishReligion
{IR, IL, SYR}

$\mathcal{C}'_8 \in \mathcal{F}^3_3$
q(A) ← believes(A,B) &
A:MiddleEastCountry, B:ChristianReligion
{ARM, IR, IRQ, IL, JOR, RL, SYR}

$\mathcal{C}'_9 \in \mathcal{F}^3_3$
q(A) ← believes(A,B) &
A:MiddleEastCountry, B:MuslimReligion
{BRN, IR, IRQ, IL, JOR, KWT, RL, OM, Q, SA, SYR, TR, UAE}

$\mathcal{C}'_{10} \in \mathcal{F}^3_5$
q(A) ← believes(A,B), believes(A,C) &
A:MiddleEastCountry,
B:ChristianReligion, C:MuslimReligion
{IR, IRQ, IL, JOR, RL, SYR}

$\mathcal{C}'_{11} \in \mathcal{F}^3_5$
q(A) ← believes(A,B), believes(A,C) &
A:MiddleEastCountry,
B:MuslimReligion, C:MuslimReligion
{BRN, IR, SYR}

organized in the DAG $\mathcal{G}'_{\texttt{CIA}}$. They are numbered according to the chronological order of insertion in $\mathcal{G}'_{\texttt{CIA}}$ and annotated with information of the generation step. From a qualitative point of view, concepts $\mathcal{C}'_2$[4] and $\mathcal{C}'_9$ well characterize Middle East countries. Armenia (ARM), as opposite to Iran (IR), does not fall in these concepts. It rather belongs to the weaker characterizations $\mathcal{C}'_3$ and $\mathcal{C}'_8$. This proves that our procedure performs a 'sensible' clustering. Indeed Armenia is a well-known borderline case for the geo-political concept of Middle East, though the Armenian is usually listed among Middle Eastern ethnic groups. Modern experts tend nowadays to consider it as part of Europe, therefore out of Middle East. But in 1996 the on-line CIA World Fact Book still considered Armenia as part of Asia.

When the m.s.d. criterion is adopted, the intensions for the concepts $\mathcal{C}'_4$, $\mathcal{C}'_6$ and $\mathcal{C}'_7$ change as follows:

$\mathcal{C}'_4 \in \mathcal{F}^2_5$
q(A) ← speaks(A,B), believes(A,C) &
A:MiddleEastCountry,
B:ArabicLanguage, C:MuslimReligion
{IR, SA}

[2] http://www.dbis.informatik.uni-goettingen.de/Mondial/mondial-rel-facts.flp

[3] http://www.odci.gov/cia/publications/factbook/

[4] $\mathcal{C}'_2$ is less populated than expected because $\mathcal{B}_{\texttt{CIA}}$ does not provide facts on the languages spoken for all countries.

$\mathcal{C}_6' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,'Druze'), believes(A,B),
        believes(A,C), believes(A,D) &
        A:MiddleEastCountry, B:JewishReligion,
        C:ChristianReligion, D:MuslimReligion
{IL, SYR}
```

$\mathcal{C}_7' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,B), believes(A,C), believes(A,D) &
        A:MiddleEastCountry, B:JewishReligion,
        C:ChristianReligion, D:MuslimReligion
{IR, IL, SYR}
```

In particular $\mathcal{C}_6'$ and $\mathcal{C}_7'$ look quite overfitted to data. Yet overfitting allows us to realize that what distinguishes Israel (IL) and Syria (SYR) from Iran is just the presence of Druze people.

Case for $l \geq 3$. The other two experiments further restrict the conditions of the language bias specification. Here only descriptions with variables constrained by concepts of granularity from the third level on are considered. When the m.g.d. criterion is adopted, the procedure for graph building returns the following nine concepts:

$\mathcal{C}_0'' \in \mathcal{F}_1^1$

```
q(A) ← A:MiddleEastCountry
{ARM, BRN, IR, IRQ, IL, JOR, KWT, RL, OM, Q, SA, SYR, TR, UAE, YE}
```

$\mathcal{C}_1'' \in \mathcal{F}_3^3$

```
q(A) ← speaks(A,B) &
        A:MiddleEastCountry, B:ArabicLanguage
{IR, SA, YE}
```

$\mathcal{C}_2'' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,'Druze') & A:MiddleEastCountry
{IL, SYR}
```

$\mathcal{C}_3'' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,B) &
        A:MiddleEastCountry, B:JewishReligion
{IR, IL, SYR}
```

$\mathcal{C}_4'' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,B) &
        A:MiddleEastCountry, B:ChristianReligion
{ARM, IR, IRQ, IL, JOR, RL, SYR}
```

$\mathcal{C}_5'' \in \mathcal{F}_3^3$

```
q(A) ← believes(A,B) &
        A:MiddleEastCountry, B:MuslimReligion
{BRN, IR, IRQ, IL, JOR, KWT, RL, OM, Q, SA, SYR, TR, UAE}
```

$\mathcal{C}_6'' \in \mathcal{F}_5^3$

```
q(A) ← speaks(A,B), believes(A,C) &
        A:MiddleEastCountry,
        B:ArabicLanguage, C:MuslimReligion
{IR, SA}
```

$\mathcal{C}_7'' \in \mathcal{F}_5^3$

```
q(A) ← believes(A,B), believes(A,C) &
        A:MiddleEastCountry,
        B:ChristianReligion, C:MuslimReligion
{IR, IRQ, IL, JOR, RL, SYR}
```

$\mathcal{C}_8'' \in \mathcal{F}_5^3$

```
q(A) ← believes(A,B), believes(A,C) &
        A:MiddleEastCountry,
        B:MuslimReligion, C:MuslimReligion
{BRN, IR, SYR}
```

organized in a DAG $\mathcal{G}_{\texttt{CIA}}''$ which partially reproduces $\mathcal{G}_{\texttt{CIA}}'$. Note that the stricter conditions set in the language bias cause two concepts occurring in $\mathcal{G}_{\texttt{CIA}}'$ not to appear in $\mathcal{G}_{\texttt{CIA}}''$: the scarsely significant $\mathcal{C}_1'$ and the quite interesting $\mathcal{C}_3'$.

When the m.s.d. condition is chosen, the intensions for the concepts $\mathcal{C}_2''$ and $\mathcal{C}_3''$ change analogously to $\mathcal{C}_6'$ and $\mathcal{C}_7'$.

4 CONCLUSIONS AND FUTURE WORK

Ontology Refinement can be considered as the problem of contextualizing an input ontology. In our case, context is conveyed by task-relevant concepts and is attached to the reference concept by discovering strong associations between the reference concepts and the task-relevant concepts w.r.t. the input ontology. The idea of applying association rules in Ontology Learning has been already investigated in [7]. Yet there are several differences: [7] is conceived for Ontology Extraction instead of Ontology Refinement, uses generalized association rules (bottom-up search) instead of multi-level association rules (top-down search), adopts propositional logic instead of First Order Logic. Also our work has contact points with Vrain's proposal [12] of a top-down incremental but distance-based method for Conceptual Clustering in a mixed object-logical representation.

For the future we plan to extensively evaluate our approach. Experiments will show, among the other things, how emerging concepts depend on the minimum support thresholds set for the stage of frequent pattern discovery.

References

[1] S. Ceri, G. Gottlob, and L. Tanca, *Logic Programming and Databases*, Springer, 1990.

[2] F.M. Donini, M. Lenzerini, D. Nardi, and A. Schaerf, '$\mathcal{AL}$-log: Integrating Datalog and Description Logics', *Journal of Intelligent Information Systems*, **10**(3), 227–252, (1998).

[3] J. Han and Y. Fu, 'Mining multiple-level association rules in large databases', *IEEE Transactions on Knowledge and Data Engineering*, **11**(5), (1999).

[4] F.A. Lisi, 'A Pattern-Based Approach to Conceptual Clustering in FOL.', in *Conceptual Structures: Inspiration and Application*, eds., H. Schärfe, P. Hitzler, and P. Øhrstrøm, volume 4068 of *Lecture Notes in Artificial Intelligence*, 346–359, Springer, (2006).

[5] F.A. Lisi and F. Esposito, 'Two Orthogonal Biases for Choosing the Intensions of Emerging Concepts', in *ECAI 2006*. IOS Press, (2006).

[6] F.A. Lisi and D. Malerba, 'Inducing Multi-Level Association Rules from Multiple Relations', *Machine Learning*, **55**, 175–210, (2004).

[7] A. Maedche and S. Staab, 'Discovering Conceptual Relations from Text', in *Proceedings of the 14th European Conference on Artificial Intelligence*, ed., W. Horn, pp. 321–325. IOS Press, (2000).

[8] A. Maedche and S. Staab, 'Ontology Learning', in *Handbook on Ontologies*, eds., S. Staab and R. Studer, Springer, (2004).

[9] H. Mannila and H. Toivonen, 'Levelwise search and borders of theories in knowledge discovery', *Data Mining and Knowledge Discovery*, **1**(3), 241–258, (1997).

[10] R.S. Michalski and R.E. Stepp, 'Learning from observation: Conceptual clustering', in *Machine Learning: an artificial intelligence approach*, eds., R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, volume I, Morgan Kaufmann, San Mateo, CA, (1983).

[11] M. Schmidt-Schauss and G. Smolka, 'Attributive concept descriptions with complements', *Artificial Intelligence*, **48**(1), 1–26, (1991).

[12] C. Vrain, 'Hierarchical conceptual clustering in a first order representation.', in *Foundations of Intelligent Systems*, eds., Z.W. Ras and M. Michalewicz, volume 1079 of *Lecture Notes in Computer Science*, 643–652, Springer, (1996).

[13] H. Xiong, M. Steinbach, A. Ruslim, and V. Kumar, 'Characterizing pattern based clustering', Technical Report TR 05-015, Dept. of Computer Science and Engineering, University of Minnesota, Minneapolis, USA, (2005).