Identification of Ontological Relations in Domain Corpus Using Formal Concept Analysis

Mireya Tovar, David Pinto, Azucena Montes, Gabriel González and Darnes Vilariño

Abstract—In this paper we present an approach for the automatic identification of relations in ontologies of restricted domain. We use the evidence found in a corpus associated to the same domain of the ontology for determining the validity of the ontological relations. Our approach employs formal concept analysis, a method used for the analysis of data, but in this case used for relations discovery in a corpus of restricted domain. The approach uses two variants for filling the incidence matrix that this method employs. The formal concepts are used for evaluating the ontological relations of the target ontology. The performance obtained was about 96% for taxonomic relations and 100% for non-taxonomic relations.

Index Terms—Formal concept analysis, ontology evaluation, ontological relations

I. INTRODUCTION

THERE is a huge amount of information that is uploaded every day to the World Wide Web, thus arising the need for automatic tools able to understand the meaning of such information. However, one of the central problems of constructing such tools is that this information remains unstructured nowadays, despite the effort of different communities for giving a semantic sense to the World Wide Web. In fact, the Semantic Web research direction attempts to tackle this problem by incorporating semantic to the web data, so that it can be processed directly or indirectly by machines in order to transform it into a data network [1]. For this purpose, it has been proposed to use knowledge structures such as “ontologies” for giving semantic and structure to unstructured data. An ontology, from the computer science perspective, is “an explicit specification of a conceptualization” [2].

Ontologies can be divided into four main categories, according to their generalization levels: generic ontologies, representation ontologies, domain ontologies, and application ontologies. Domain ontologies, or ontologies of restricted domain, specify the knowledge for a particular type of domain, for example: medical, tourism, finance, artificial intelligence, etc. An ontology typically includes the following components: classes, instances, attributes, relations, constraints, rules, events and axioms.

Our approach attempts to find evidence of the relations to be evaluated in a reference corpus (associated to the same domain of the ontology) using formal concept analysis. To our knowledge, the use of formal concept analysis in the automatic discovery of ontological relations has nearly been studied in the literature. There are, however, other approaches that may be considered in our state of the art, because they provide mechanisms for discovering ontological relations, usually in the construction of ontologies framework.

In [3], for example, it is presented an approach for the automatic acquisition of taxonomies from text in two domains: tourism and finance. They use different measures for weighting the contribution of each attribute (such as conditional probability and pointwise mutual information (PMI)).

In [4] are presented two experiments for building taxonomies automatically. In the first experiment, the attribute set includes a group of sememes obtained from the HowNet lexicon, whereas in the second the attributes are a basically set of context verbs obtained from a large-scale corpus; all this for building an ontology (taxonomy) of the Information Technology (IT) domain. They use five experts of IT for evaluating the results of the system, reporting a 43.2% of correct answers for the first experiment, and 56.2% of correct answers for the second one.

Hele-Mai Haav [5] presents an approach to semi-automatic ontology extraction and design by usign Formal Concept Analysis combined with a rule-based language, such as Horn clauses, for taxonomic relations. The attributes are noun-phrases of a domain-specific text describing a given entity. The non-taxonomic relations are defined by means of predicates and rules using Horn clauses.

In [6] it is presented an approach to derive relevance of “events” from an ontology of the event domain. The ontology of events is constructed using Formal Concept Analysis. The
event terms are mapped into objects, and the name entities into attributes. These terms and entities were recovered from an corpus in order to build the incidence matrix.

From the point of view of the evaluation of the ontology, some of the works mentioned above perform an evaluation by means of gold standard ([3]) in order to determine the level of overlapping between the ontology that has been built automatically and the manually constructed ontology (called gold standard).

Another approach for evaluating ontologies is by means of human experts as it is presented in [4].

In our approach we used a typed dependency parser for determining the verb of a given sentence, which is associated to the ontological concepts of a triple from which the relation component require to be validated through a retrieval system. The ontological concepts together with their associated verbs are introduced, by means of an incidence matrix, to Formal Concept Analysis (FCA) system. The FCA method allow us to find evidence of the ontological relation to be validated by searching the semantic implicit in the data. We use several selection criteria to determine the veracity of the ontological relation.

We do not do ontological creation, but we use formal concept analysis to identify the ontological relation in the corpus and we evaluate it.

In order to validate our approach, we employ a manual evaluation process by means of human experts.

The remaining of this paper is structured as follows: Section II describes more into detail the theory of formal concept analysis. In section III we present the approach proposed in this paper. Section IV shows and discusses the results obtained by the presented approach. Finally, in Section V the findings and the future work are given.

II. FORMAL CONCEPT ANALYSIS

Formal Concept Analysis (FCA) is a method of data analysis that describes relations between a particular set of objects and a particular set of attributes [7]. FCA was firstly introduced by Rudolf Wille in 1992 [8] as an field of research based on a model of set theory to concepts and concept hierarchies which proposes a formal representation of conceptual knowledge [8]. FCA allows data analysis methods for the formal representation of conceptual knowledge. This type of analysis produces two kinds of output from the input data: a concept lattice and a collection of attribute implications. The concept lattice is a collection of formal concepts of the data, which are hierarchically ordered by a subconcept-superconcept relation. The attribute implication describes a valid dependency in the data. FCA can be seen as a conceptual clustering technique that provides intentional descriptions for abstract concepts. From a philosophical point of view, a concept is a unit of thoughts made up of two parts: the extension and the intension [9]. The extension covers all objects or entities belonging to this concept, whereas the intension comprises all the attributes or properties valid for all those objects.

FCA begins with the primitive idea of a context defined as a triple \((G, M, I)\), where \(G\) and \(M\) are sets, and \(I\) is a binary relation between \(G\) and \(M\) (\(I\) is the incidence of the context); the elements of \(G\) and \(M\) are named objects and attributes, respectively.

A pair \((A, B)\) is a formal concept of \((G, M, I)\), as defined in [3], iff \(A \subseteq G, B \subseteq M, A' = B\) and \(A = B'\). In other words, \((A, B)\) is a formal concept if the attribute set shared by the objects of \(A\) are identical with those of \(B\); and \(A\) is the set of all the objects that have all attributes in \(B\). \(A\) is the extension, and \(B\) is the intension of the formal concept \((A, B)\).

\(A'\) is the set of all attributes common to the objects of \(A\), \(B'\) is the set of all objects that have all attributes in \(B\). For \(A \subseteq G, A' = \{m \in M \mid \forall g \in A : (g, m) \in I\}\), and dually, for \(B \subseteq M, B' = \{g \in G \mid \forall m \in B : (g, m) \in I\}\).

The formal concepts of a given context are ordered by the relation of subconcept - superconcept defined by:

\[(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1\]

FCA is a tool applied to various problems such as: hierarchical taxonomies, information retrieval, data mining, etc., [7]. In this case, we use this tool for identifying ontological relations of restricted domain.

III. APPROACH FOR EVALUATING SEMANTIC RELATIONS

We employ the theory of FCA to automatically identify ontological relations in a corpus of restricted domain. The approach considers two variants in the selection of properties or attributes for building the incidence matrix that is used by the FCA method for obtaining the formal concepts.

The difference between the two variants is the type of syntactic dependencies parser used in the preprocessing phase for getting the properties.

The first variant uses the minipar tagger [10], whereas the second variant employs the Stanford tagger [11]. For each variant, we selected manually a set of dependency relations in order to extract verbs from each sentence of the corpus that contains an ontology concept. These verbs are then used as properties or attributes in the incidence matrix.

The Stanford dependencies are triples containing the name of the relation, the governor and the dependent. Examples of these triples are shown in Table I. For the purpose of our research, from each triple we have selected the governor (p=1), the dependent (p=2) or both (p=1,2) as attributes of the incidence matrix.

In the case of the minipar parser, we use the pattern C:i:V for recovering the verbs of the sentence. The grammatical categories that made up the pattern follows: \(C\) is a clause, \(i\) is an infeclional phrase, and \(V\) is a verb or verbal phrase. Some examples of triples recovered from the sentences are shown in Table II.

The approach proposed in this paper involves the following three phases:

1) Pre-processing stage. The reference corpus is split into sentences, and all the information (ontology and the sentences) are normalized. In this case, we use

<table>
<thead>
<tr>
<th>Triples</th>
<th>fn C:i:VBE be</th>
<th>inf C:i:V make</th>
<th>inf C:i:V function</th>
</tr>
</thead>
</table>

TABLE II

TRIPLES OBTAINED BY THE MINIPAR PARSER

(Advance online publication: 24 April 2015)
we have employed an ontology of the Artificial Intelligence (AI) domain\cite{14} for the experiments executed. In Table IV we present the number of concepts (C), taxonomic relations ($T R$) and non-taxonomic relations ($NT$) of the ontology evaluated. The characteristics of its reference corpus are also given in the same Table: number of documents (D), number of tokens (T), vocabulary dimensionality (V), and the number of sentences filtered (O) by the information retrieval system (S).

### A. Dataset

We have employed an ontology of the Artificial Intelligence (AI) domain\cite{14} for the experiments executed. In Table IV we present the number of concepts (C), taxonomic relations ($T R$) and non-taxonomic relations ($NT$) of the ontology evaluated. The characteristics of its reference corpus are also given in the same Table: number of documents (D), number of tokens (T), vocabulary dimensionality (V), and the number of sentences filtered (O) by the information retrieval system (S).
TABLE III
PATTERNS OR RELATION NAME USED BY EACH VARIANT

<table>
<thead>
<tr>
<th>Variant</th>
<th>Pattern or relation name</th>
<th>Type of selection</th>
<th>Type of relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>minipar</td>
<td>C:i:V</td>
<td>All verbs</td>
<td>taxonomic, non-taxonomic</td>
</tr>
<tr>
<td>stanford1</td>
<td>root, cop</td>
<td>Only the verbs to be and include</td>
<td>taxonomic</td>
</tr>
<tr>
<td>stanford2</td>
<td>nsubj, prep, root, dobj, acomp, advcl, agent, aux, auxpass, cop, csubj, csubjpass, dobj, expl, iobj, cop, nsubjpass, parataxis, pcomp, prepc, prt, tmod, vmod</td>
<td>All verbs recovered</td>
<td>non-taxonomic</td>
</tr>
<tr>
<td>stanford3</td>
<td>nsubj, prep, root, dobj, acomp, advcl, agent, aux, auxpass, cop, csubj, csubjpass, dobj, expl, iobj, cop, nsubjpass, parataxis, pcomp, prepc, prt, tmod, vmod</td>
<td>Only the verbs present in the ontological relations</td>
<td>non-taxonomic</td>
</tr>
</tbody>
</table>

TABLE IV
DATASETS

<table>
<thead>
<tr>
<th>Domain</th>
<th>Ontology</th>
<th>Reference corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>TR</td>
<td>NT</td>
</tr>
<tr>
<td>D</td>
<td>T</td>
<td>V</td>
</tr>
<tr>
<td>O</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AI</td>
<td>276</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>11,370</td>
<td>1,510</td>
</tr>
<tr>
<td></td>
<td>475</td>
<td>415</td>
</tr>
</tbody>
</table>

TABLE V
ACCURACY OF THE AI ONTOLOGY, AND QUALITY OF THE SYSTEM PREDICTION FOR TAXONOMIC RELATIONS

<table>
<thead>
<tr>
<th>Variation</th>
<th>Accuracy</th>
<th>Q(Ei1)</th>
<th>Q(Ei2)</th>
<th>Q(Ei3)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>minipar</td>
<td>0.96</td>
<td>0.90</td>
<td>0.85</td>
<td>0.94</td>
<td>0.90</td>
</tr>
<tr>
<td>stanford1</td>
<td>0.61</td>
<td>0.57</td>
<td>0.56</td>
<td>0.60</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table VI shows the results obtained by the approach when the non-taxonomic relations are evaluated.

TABLE VI
ACCURACY OF THE AI ONTOLOGY AND QUALITY OF THE SYSTEM PREDICTION FOR NON-TAXONOMIC RELATIONS

<table>
<thead>
<tr>
<th>Variation</th>
<th>Accuracy</th>
<th>Q(Ei1)</th>
<th>Q(Ei2)</th>
<th>Q(Ei3)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>minipar</td>
<td>0.93</td>
<td>0.81</td>
<td>0.86</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>stanford2</td>
<td>0.97</td>
<td>0.87</td>
<td>0.92</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>stanford3</td>
<td>0.92</td>
<td>0.85</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The results presented here were obtained with a subset of sentences associated to the ontological relations because of the great effort needed for manually evaluate their validity. Therefore, in order to have a complete evaluation of the two type of ontological relations, we have calculated their accuracy, but in this case considering all the sentences associated to the relations to be evaluated. Table VII shows the variants used for evaluating the ontological relations and the accuracy assigned to each type of relation (Accuracy).

As can be seen, the approach obtained a better accuracy for non-taxonomic relations than for taxonomic ones. This result is obtained because the approach is able to associate the verbs that exist in both, the relation and the domain corpus, by means of the FCA method. Therefore, when non-taxonomic relations are evaluated, the approach has more opportunity to find evidence of their validity.

V. CONCLUSION

In this paper we have presented an approach based on FCA for the evaluation of ontological relations. In summary, we attempted to look up for evidence of the ontological relations to be evaluated in a reference corpus (associated to the same domain of the ontology) by using formal concept analysis. The method of data analysis employed was tested by using two types of variants in the selection of properties or attributes for building the incidence matrix needed by the FCA method in order to obtain the formal concepts. The main difference between these two variants is the type of syntactic dependency parser used in the preprocessing phase when obtaining the data properties (Stanford vs. minipar). The Stanford variant was more accurate than the minipar one; actually, the minipar variant obtained a good accuracy for the two types of relations evaluated (taxonomic and non-taxonomic), whereas the Stanford variant obtained the best results for the non-taxonomic relations. The minipar variant, on the other hand, is quite fast in comparison with the Stanford one.

According to the results presented above, the current approach for evaluating ontological relations obtains an accuracy of 96% for taxonomic relations, and 100% for non-taxonomic relations. Even if these results determine the evidence of the target ontological relations in the corresponding reference corpus, the same results should be seen in terms of the ability of our system for evaluating ontological relations. In other words, the results obtained by the presented approach show, in some way, the quality of the ontology.

We have observed that the presented approach may have future in the evaluation of ontologies task, but we consider
that there still more research that need to be done. For example, as future work, we are interested in analyzing more into detail the reasons for which the approach does not detect 100% of the ontological relations that have some kind of evidence in the reference corpus.

REFERENCES


(Advance online publication: 24 April 2015)