Patterns Used to Identify Relations in Corpus Using Formal Concept Analysis

Mireya Tovar[1,2(✉)], David Pinto[1], Azucena Montes[2,3], Gabriel Serna[2], and Darnes Vilariño[1]

[1] Faculty Computer Science, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico
{mtovar,dpinto,darnes}@cs.buap.mx
[2] Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Mexico
{gabriel,amontes}@cenidet.edu.mx
[3] Engineering Institute, Universidad Nacional Autónoma de Mexico, Mexico City, Mexico

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 two ontologies. The performance obtained was about 96 for taxonomic relations and 100 % for non-taxonomic relations, in the first ontology. In the second it was about 92 % for taxonomic relations and 98 % for non-taxonomic relations.

Keywords: Formal concept analysis · Ontology · Semantic relations

1 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].

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2015, LNCS 9116, pp. 236–245, 2015.
DOI: 10.1007/978-3-319-19264-2_23

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.

In this paper we are interested in the process of discovering and evaluating ontological relations, thus, we focus our attention on the following two types: taxonomic relations and/or non-taxonomic relations. The first type of relations are normally referred as relations of the type "is-a" (hypernym/hyponymy or subsumption).

There are plenty of research works in literature that addresses the problem of automatic construction of ontologies. The major of those works evaluate manually created ontologies by using a gold standard, which in fact, it is supposed to be manufactured by an expert. By using this approach, it is assumed that the expert has created the ontology in a correct way, however, there is not a guarantee of such thing. Thus, we consider very important to investigate a manner to automatically evaluate the quality of this kind of resources, which are continuously been used in the framework of the semantic web.

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: Sect. 2 describes more into detail the theory of formal concept analysis. In Sect. 3 we present the approach proposed in this paper. Section 4 shows and discusses the results obtained by the presented approach. Finally, in Sect. 5 the findings and the future work are given.

2 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 beloging 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 $A \subseteq G$, $A' = \{m \in M | \forall g \in A : (g, m) \in I\}$, and dually, for $B \subseteq M$, $B' = \{g \in G | \forall m \in B : (g, m) \in I\}$

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

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq B_1)$$

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.

3 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 1. 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 inflectional phrase, and V is a verb or verbal phrase. Some examples of triples recovered from the sentences are shown in Table 2.

Table 1. Dependency relations obtained using the Stanford dependency parser

Relation name	p	Meaning	Example
nsubj	1	Nominal subject	nsubj(specialized, research)
prep	1	Prepositional modifier	prep_into(divided, subfields)
root	2	Root of the sentence	root(ROOT, give)
acomp	1	Adjectival complement	acomp(considered, feasible)
advcl	1,2	Adverbial clause modifier	advcl(need, provide)
agent	1	Agent complement of a passive verb	agent(simulated, machine)
aux	1,2	Auxiliar verb	aux(talked, can)
auxpass	1,2	Passive auxiliar	auxpass(used, is)
cop	1,2	Copula	cop(funded, is)
csubj	2	Clausal subject	csubj(said, having)
csubjpass	1,2	Clausal passive subject	csubjpass(activated, assuming)
dobj	1	Direct object of a verbal phrase	dobj(create, system)
expl	1	Expletive	expl(are, there)
iobj	1	Indirect object	iobj(allows, agent)
nsubjpass	1	Passive nominal subject	nsubjpass(embedded, agent)
parataxis	2	Parataxis	parataxis(Scientist, said)
pcomp	2	Prepositional complement	pcomp(allow, make)
prepc	1	Prepositional clausal modifier	prepc_like(learning, clustering)
prt	1,2	Phrasal verb particle	prt(find, out)
tmod	1	Temporal modifier	tmod(take, years)
vmod	2	Reduced non-finite verbal modifier	vmod(structure, containing)

Table 2. Triples obtained by the Minipar parser

Triples
fin C:i:VBE be
inf C:i:V make
fin C:i:V function

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 the TreeTagger PoS tagger for obtaining the lemmas [12]. An information retrieval system is employed for filtering those sentences containing information referring to the concepts extracted from the ontology. The ontological relations are also extracted from the ontology[1]. Thereafter, we apply the syntactic dependency parser for each sentence associated to the ontology

[1] We used Jena for extracting concepts and ontological relations (http://jena.apache.org/).

concepts. In order to extract the verbs from these sentences, we use the patterns shown in Table 3 for each syntactic dependency parser, and each type of ontological relation.

By using this information together with the ontology concepts, we construct the incidence matrix that feed the FCA system.

2. FCA system. We used the sequential version of FCALGS[2] [13]. The input for this system is the incidence matrix with the concepts identified as objects and the verbs identified as attributes. The output is the formal concepts list.
3. Identification of ontological relations. The concepts that made up the triple in which the ontological relation is present are searched in the formal concepts list obtained by the FCA system. The approach assigns a value of 1 (one) if the pair of concepts of the ontological relation exists in the formal concept, otherwise it assigns a zero value. We consider the selection criteria shown in the third column of Table 3 for each type of ontological relation.

   As can be seen, in the Stanford approach we have tested three different selection criteria based on the type of verbs to be used. In "stanford$_1$", we only selected the verbs "to be" and "include" that normally exists in lexico-syntactic patterns of taxonomic relations [14]. On the other hand, in "stanford$_3$ we only selected the verbs that exist in the ontological relation.
4. Evaluation. Our approach provides a score for evaluating the ontology by using the accuracy formulae: Accuracy(ontology) $= \frac{|S(R)|}{|R|}$, where $|S(R)|$ is the total number of relations from which our approach considers that exist evidence in the reference corpus, and $|R|$ is the number of semantic relations in the ontology to be evaluated. For measuring this approach, we compare the results obtained by our approach with respect to the results obtained by human experts.

4 Experimental Results

In this section we present the results obtained in the experiments carried out. Firstly, we present the datasets, the results obtained by our approach aforementioned follow; finally, the discussion of these results are given.

4.1 Dataset

We have employed two ontologies, the first is of the Artificial Intelligence (AI) domain and the second is of the standard e-Learning SCORM domain (SCORM)[3] [15] for the experiments executed. In Table 4 we present the number of concepts (C), taxonomic relations (TR) and non-taxonomic relations (NT) of the ontologies evaluated. The characteristics of their 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).

[2] http://fcalgs.sourceforge.net/.

[3] The ontologies together with their reference corpus can be downloaded from http://azouaq.athabascau.ca/goldstandards.htm.

Table 3. Patterns used by each variant

Variant	Pattern	Type of selection	Type of relation
minipar	C:i:V *	All verbs recovered	taxonomic, non-taxonomic
stanford_1	root(*,*), cop(*,*)	Only the verbs *to be* and *include*	taxonomic
stanford_2	nsubj(*,-), prep(*,-), root(*,*), dobj(*,-), acomp(*,-), advcl(*,*), agent(*,-), aux(*,*), auxpass(*,*), cop(*,*), csubj(-,*), csubjpass(*,-), dobj(*,-), expl(*,-), iobj(*,-), cop(*,*), nsubjpass(*,-), parataxis(-,*), pcomp(-,*), prepc(*,-), prt(*,*), tmod(*,-), vmod(-,*)	All verbs recovered	non-taxonomic
stanford_3		Only the verbs present in the ontological relations	non-taxonomic

Table 4. Datasets

Domain	Ontology			Reference corpus				
	C	TR	NT	D	T	V	O	S
AI	276	205	61	8	11,370	1,510	475	415
SCORM	1,461	1,038	759	36	34,497	1,325	1,621	1,606

4.2 Obtained Results

As we mentioned above, we validated the ontology relations by means of human expert's judgements. This manual evaluation was carried out in order to determine the performance of our approach, and consequently, the quality of the ontology.

Table 5 shows the results obtained by the approach presented in this paper when the ontologies are evaluated. We used the accuracy criterion for determining the quality of the taxonomic relations. The second column presents two variants for identifying the taxonomic relations. The last three columns indicate the quality (Q) of the system prediction according to three different human experts (E_1, E_2 and E_3). The third column shows the quality obtained by the approach for each type of variant. Table 6 shows the results obtained by the approach when the non-taxonomic relations are evaluated.

Table 5. Accuracy of the ontologies, and quality of the system prediction for taxonomic relations

Domain	Variation	*Accuracy*	$Q(E_1)$	$Q(E_2)$	$Q(E_3)$	*Average*
AI	minipar	0.96	0.90	0.85	0.94	0.90
	stanford_1	0.61	0.57	0.56	0.60	0.58
SCORM	minipar	0.89	0.65	0.75	0.64	0.68
	stanford_1	0.64	0.49	0.47	0.50	0.49

Table 6. Accuracy of the ontologies and quality of the system prediction for non-taxonomic relations

Domain	Variation	*Accuracy*	$Q(E_1)$	$Q(E_2)$	$Q(E_3)$	*Average*
AI	minipar	0.93	0.80	0.86	0.89	0.85
	stanford_2	1.00	0.87	0.92	0.95	0.91
	stanford_3	0.95	0.82	0.91	0.91	0.87
SCORM	minipar	0.96	0.85	0.89	0.94	0.89
	stanford_2	0.99	0.87	0.89	0.97	0.91
	stanford_3	0.90	0.83	0.83	0.90	0.85

Table 7. Accuracy given to the ontologies

Domain	*Relation type*	Variante	*Accuracy*
AI	Taxonomic	minipar	96.59 %
		stanford_1	73.17 %
	Non-taxonomic	minipar	95.08 %
		stanford_2	100.00 %
		stanford_3	96.72 %
SCORM	Taxonomic	minipar	92.49 %
		stanford_1	64.45 %
	Non-taxonomic	minipar	96.18 %
		stanford_2	98.95 %
		stanford_3	91.44 %

The results presented here were obtained with a subset of sentences associated to the ontological relations for the AI ontology because of the great effort needed for manually evaluate their validity. In the case of SCORM ontology, we only evaluate the 10 % of the ontological relations and a subset of sentences associated to these. 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 7 shows the variantes 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.

5 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 reference corpora (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) in AI ontology, 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 of the AI ontology. In the case of the SCORM ontology, our approach obtains an accuracy of 92 % for taxonomic relations, and 98 % 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 ontologies.

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

1. Solís, S.: La Web Semántica. Lulu Enterprises Incorporated (2007)
2. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In: Guarino, N., Poli, R. (eds.) Formal Ontology in Conceptual Analysis and Knowledge Representation. Kluwer Academic Publishers, Deventer (1993)
3. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. **24**(1), 305–339 (2005)

4. Li, S., Lu, Q., Li, W.: Experiments of ontology construction with formal concept analysis. In: ren Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prevot, L., (eds.) Ontology and the Lexicon, pp. 81–97. Cambridge University Press, New york (2010). Cambridge Books Online
5. Haav, H.M.: A semi-automatic method to ontology design by using FCA. In: Snsel, V., Belohlvek, R., (eds.) CLA. CEUR Workshop Proceedings, vol. 110. CEUR-WS.org (2004)
6. Xu, W., Li, W., Wu, M., Li, W., Yuan, C.: Deriving event relevance from the ontology constructed with formal concept analysis. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 480–489. Springer, Heidelberg (2006)
7. Belohlávek, R.: Introduction to formal context analysis. Technical report, Department of Computer Science. Palacḱ y University, Olomouk, Czech Republic (2008)
8. Wille, R.: Concept lattices and conceptual knowledge systems. Comput. Math. Appl. **23**(6–9), 493–515 (1992)
9. Wolf, E.K.: A first course in formal concept analysis. In: Faulbaum, F. (ed.) SoftStat 1993 Advances in Statistical Software 4, pp. 429–438. Gustav Fischer Verlag (1993)
10. Lin, D.: Dependency-based evaluation of minipar. In: Proceedings of Workshop on the Evaluation of Parsing Systems. Granada (1998)
11. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure trees. In: LREC (2006)
12. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)
13. Krajca, P., Outrata, J., Vychodil, V.: Parallel recursive algorithm for FCA. In: Proceedings of the Sixth International Conference on Concept Lattices and Their Applications, vol. 433, pp. 71–82. CEUR-WS.org, Olomouc (2008)
14. Tovar, M., Pinto, D., Montes, A., González, G., Vilariño, D., Beltrán, B.: Use of lexico-syntactic patterns for the evaluation of taxonomic relations. In: Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Olvera-Lopez, J.A., Salas-Rodríguez, J., Suen, C.Y. (eds.) MCPR 2014. LNCS, vol. 8495, pp. 331–340. Springer, Heidelberg (2014)
15. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M., (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)