A Redundancy-based Method for Relation Instantiation from the Web

Viktor de Boer and **Maarten van Someren** and **Bob J. Wielinga** [1]

Abstract. The Semantic Web requires automatic ontology population methods. We developed an approach, that given existing ontologies, extracts instances of ontology relations, a specific subtask of ontology population. We use generic, domain independent techniques to extract candidate relation instances from the Web and exploit the redundancy of information on the Web to compensate for loss of precision caused by the use of these generic methods. The candidate relation instances are then ranked based on co-occurrence with a seed set. In an experiment, we extracted instances of the relation between artists and art styles. The results were manually evaluated against selected art resources.

1 INTRODUCTION

The ongoing project of the Semantic Web calls for (semi-)automatic methods for the construction of ontologies (ontology learning) and knowledge bases (ontology population)[3]. In this paper, we describe a method for *relation instantiation*, a subtask of ontology population.

We define a (partly) populated ontology as a set of labeled classes (the domain concepts) $C_1, ..., C_n$, hierarchically ordered by a subclass relation. Non-hierarchical relations between concepts are also defined ($R : C_i \times C_j$). We also have a knowledge base containing instances of the ontology concepts. The task of relation instantiation is to identify for a single instance i of C_i for which instances j of C_j, the relation $R(i, j)$ is true given the information in the corpus. In this paper we assume that R is not a one-to-one relation (The instance i is related to multiple instances of C_j). We also assume that we know all instances of C_j and have a method available that recognizes these elements in the documents in our corpus. For a textual corpus such as the Web, this implies that the instances must have a textual label.

2 THE REDUNDANCY METHOD

Current approaches for Information Extraction or Question Answering tasks could also be used for ontology population. However, the performances of methods such as [2] depend on the specific structure or domain of the corpus. We designed our method to be structure- and domain-independent. Also, methods that use some form of supervised Machine Learning assume a large number of tagged example instances to be able to learn patterns for extracting new instances and this is a serious limitation for large scale use. Our method requires only a small amount of examples that are used as a seed set.

Our approach incorporates generic methods that do not rely on assumptions about the domain or the type of documents in the corpus. By using these general methods for the extraction, we will lose in precision since the general methods are not optimized for a specific corpus or domain. However, since we use more generic methods, we are able to extract information from a greater number of sources. The main assumption behind our method is that because of the redundancy of information on the Web and because we are able to combine information from heterogeneous sources, we can compensate for this loss of precision.

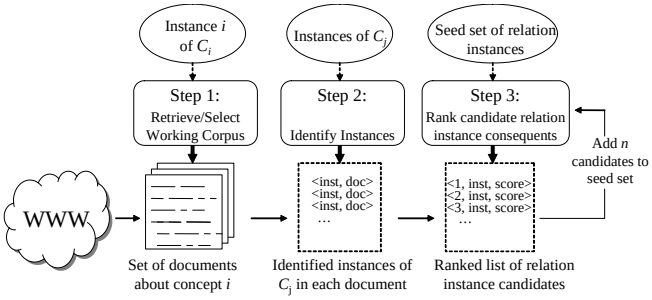

Figure 1. Outline of the method

The method consists of three steps, shown in Figure 1. We first construct a 'working corpus' by feeding the label(s) of the instance i to the Google search engine. The size of this working corpus is a parameter of the method.

In step 2, we identify the instances of the concept C_j in the documents of the working corpus. For this, we use a Named Entity Recognition module and match the results to the instances of C_j in our populated ontology, this yields our candidate relation instances.

In step 3, the method combines the evidence from the different documents to produce a ranking for these candidates. We base this ranking on the assumption that on average in individual web pages, a target relation is either well represented (the web page contains a number of correct right-hand side instances) or not represented (it contains few or none of these instances).

We therefore calculate a Document Score DS for each document. This is the probability that for all candidates in that document the relation R holds, according to the seed set. This is equal to the number of identified instances that are in the

[1] Human-Computer Studies Laboratory, Informatics Institute, Universiteit van Amsterdam, email: {vdeboer,maarten,wielinga}@science.uva.nl

seed set divided by the total number of candidate instances in that document. We then combine all evidence for each of the candidate instances by taking the average of DS over all used documents in the corpus resulting in an Instance Score IS for each candidate instance.

We then add the candidate with the highest value of IS to the seed set and iterate by recalculating all DS and IS, based on the expanded seed set. The method iterates up to a threshold on the number of iterations or a drop in the Instance Scores. In Section 3, we explore the effects of these thresholds.

3 EXTRACTING ART STYLE-ARTISTS RELATION

We tested our method in the cultural heritage domain. We used two well-known art thesauri as our partly populated ontologies: the AAT[4] and the ULAN[5]. In this experiment we extracted the instances of the relation 'has_artist' between `aat:Art Style` and `ulan:Artist`. We tested the method for nine art styles[2].

We first populated the seed set with three well-known artists associated with that art style. Then in Step 1, 1000 pages were extracted as a working corpus by querying Google with the labels of the art style instances. In Step 2, we used the Person Name Extractor from the tOKO toolkit[1] and matched the results to the ULAN. In Step 3, the DS and IS scores were calculated and for each art style. We evaluated the results of 40 iterations by having two annotators manually score each of the 40 retrieved relation instances as 'correct' or 'incorrect' for each art style. The annotators were allowed to consult a fixed set of encyclopedic web sources. Inter-annotator agreement (Cohen's Kappa) = 0.83. Finally, consensus was reached to calculate precision.

We first illustrate the results for a single art style: 'Neue Sachlichkeit' ('New Objectivity'). Figure 2 shows the IS for the top artists and value of precision for all 40 iterations.

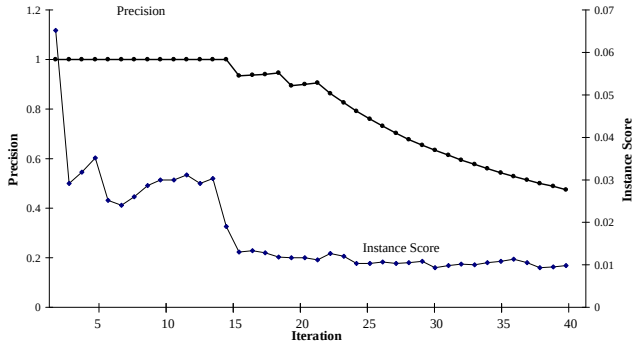

Figure 2. IS and precision for 'Neue Sachlichkeit'

The drop in precision co-occurs with a drop in IS. We can use this drop in IS as a threshold. We stop adding relation instances to the knowledge base if the value if IS of the next candidate instance is less than some drop factor, DF, multiplied by the maximum of the Instance Scores up to that iteration. We also stop adding instances after an absolute maximum number of iterations has been reached (Max). In the above example, setting DF to 0.2 and Max to 40, leads to a precision of 0.933, with 15 correct relations added.

[2] Art Deco, Art Nouveau, Cubism, Dada, Expressionism, Impressionism, Neo-Impressionism, Neue Sachlichkeit and Surrealism

Table 1. Average precision (prec) and total number of correct extractions (ex) for the nine Art Styles

	Max							
	10		20		30		40	
DF	prec	ex	prec	ex	prec	ex	prec	ex
0	0.856	77	0.806	145	0.722	195	0.650	234
0.1	0.856	77	0.806	145	0.721	193	0.648	228
0.2	0.856	77	0.799	137	0.776	179	0.746	197
0.3	0.865	73	0.842	117	0.830	138	0.810	144
0.4	0.857	62	0.834	96	0.826	114	0.824	120
0.5	0.902	55	0.878	86	0.868	103	0.866	109
0.6	0.924	46	0.896	67	0.882	81	0.880	87

In Table 1, we list both the average precision and the total sum of the number of correct relation instances extracted for the nine art styles for 24 combinations of the two threshold parameters DF and Max. The lowest value for precision is 0.65. This occurs at DF=0 (the drop in the Instance Score is not used to set the threshold) and Max=40. In that case, for the nine art styles, all 360 (9×40) extractions are added to the knowledge base, of which 234 are evaluated correct. The highest precision, 0.924, is reached at DF=0.6 and Max=10, with only 46 correct relation instances added to the knowledge base. We observe relatively high values for precision and a tradeoff between precision and number of correct extractions comparable to that of the traditional precision/recall tradeoff.

4 CONCLUSIONS

Considering the method uses very generic methods and intuitive ranking scores, the results are encouraging but also suggest that further processing of the results could improve the relation instantiation. Analysis of the working corpus showed the documents were highly heterogeneous in structure and language. How much redundancy helped is a topic for further research. Improvement in the Person Name Extraction module or combining different Person Name Extractors could improve the extraction. Also, other measures for DS and IS could be considered. An obvious direction for further research is to test this method on other relations in other domains.

ACKNOWLEDGEMENTS

This research was supported by the MultimediaN project (www.multimedian.nl) funded through the BSIK programme of the Dutch Government.

REFERENCES

[1] A. Anjewierden, B.J. Wielinga, and R. de Hoog, 'Task and domain ontologies for knowledge mapping in operational processes', *Metis Deliverable 4.2/2003, University of Amsterdam.*, (2004).

[2] N. Kushmerick, D. Weld, and R. Doorenbos, 'Wrapper induction for information extraction', in *in Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, p. 729737, (1997).

[3] A. Maedche and S. Staab, 'Ontology learning for the semantic web', *IEEE Intelligent Systems*, **13**, 993, (2001).

[4] The Getty Foundation, 'Aat: Art and architecture thesaurus', *http://www.getty.edu/research/tools/vocabulary/aat/*.

[5] The Getty Foundation, 'Ulan: Union list of artist names', *http://www.getty.edu/research/tools/vocabulary/ulan/*.