Extending an Ontology by Analyzing Annotation Co-occurrences in a Semantic Cultural Heritage Portal

Tomi Kauppinen[1], Heini Kuittinen[1], Katri Seppälä[1], Jouni Tuominen[1] and Eero Hyvönen[1]

Semantic Computing Research Group (SeCo)
Helsinki University of Technology
and University of Helsinki, Finland
http://www.seco.tkk.fi/
firstname.lastname@tkk.fi

Abstract. Ontologies aim to capture knowledge about things and their relationships. Publishing ontologies on the Semantic Web enables people and organizations to use shared ontologies in annotating e.g. photographs, videos, music, and other types of cultural objects. Search engines also use relationships provided by ontologies in semantic search, e.g. for query expansion or for view-based search. However, building ontologies is a time-consuming process, and it should be helped by automatic finding of interesting, possible relationships. Finding the correct concept for annotation purposes is helped by subsumption and partonomy hierarchies and associative relationships. In this paper we show how an analysis of co-occurrences of concepts in annotations can be used to provide interesting relationships for enriching ontological structures. We use association rule mining techniques and test the idea using a set of annotations of cultural objects in CULTURESAMPO portal and the Finnish General Upper Ontology YSO. The results are visualized in the ONKI SKOS browser to give an additional layer on top of the original relationships of the YSO ontology. An analysis shows that best ranked relationships should also be included in the ontology as subclassof or associative relationships.

1 Introduction

Ontologies consist of entities such as classes and individuals. The idea is that ontologies capture essential knowledge about the world, either upper-level, generic or domain-specific knowledge. This is done by identifying entities and then by relating these entities together by using properties and relationships. Publishing ontologies on the Semantic Web enables people and organizations to use these shared ontologies in annotating e.g. photographs, videos, music and other types of cultural objects.

The problem is that building and maintaining ontologies is both time-consuming and expensive. E.g. Surowiecki has explored the idea that “large groups of people are smarter than an elite few, no matter how brilliant” [10]. Association rule mining [1] has e.g. helped to analyze and build structure for folksonomies [9]. This means that ontology building could be facilitated by analyzing e.g. co-occurrences [9, 8] of terms in annotations by the help of association rule mining techniques.

In this paper we describe how an already existing ontology, the Finnish General Upper Ontology YSO [6] and a set of annotations made using YSO in the semantic portal CULTURESAMPO [5] were used for mining association rules. The YSO ontology already had a basic structure, so our purpose was to improve ontology by populating more relationships between its concepts rather to create new concepts or structure from scratch.

2 Method and Materials for the Study

In this section we describe how association rule mining techniques are applied to find and rank interesting relationships based on existing annotations and an existing ontology. The goal is to provide support for ontology engineering in populating new relationships between concepts. First we provide an overview of association rules applied to ontology-based annotations, then we provide an overview of the used ontology tthe Finnish General Upper Ontology YSO [6], and finally we describe the semantic portal CULTURESAMPO[5] that was used as a source for annotations.

2.1 Association Rules from Annotations

We are interested in finding those concepts that co-occur often in annotations. We therefore apply a method for mining association rules [1] and define $support(A \Rightarrow B) = S_{A \Rightarrow B} (= S_{B \Rightarrow A})$ as follows

$$S_{A \Rightarrow B} = \frac{number\ \ of\ \ annotations\ \ containing\ \ both\ \ A\ \ and\ \ B}{total\ \ number\ \ of\ \ annotations} \tag{1}$$

Similarly, $confidence(A \Rightarrow B) = C_{A \Rightarrow B}$ is defined as

$$C_{A \Rightarrow B} = \frac{number\ \ of\ \ annotations\ \ containing\ \ both\ \ A\ \ and\ \ B}{number\ \ of\ \ annotations\ \ containing\ \ just\ \ A} \tag{2}$$

The method consists of three phases. First one creates the candidate relationships. The second and third phases will prune out those relationships that already exist in an ontology. Hence they will ensure that concepts which already have a close relationship in an ontology will not get associated again.

1. In the first phase the method goes through all the annotations a_i that contain concepts from an ontology O_i to search for an initial set of association rules. For each concept used in at least one annotation the method creates a set of items. For example, one set z_A contains all those items that are annotated using the concept A. And similarly the set z_B contains another set of items annotated with B, accordingly. Next, to calculate $C_{A \Rightarrow B} = confidence(A \Rightarrow B)$ the method creates an intersection of the sets z_A and z_B. Finally, $C_{A \Rightarrow B}$ is calculated by dividing the number of items in the intersection with the number of items in the set z_A. The support $S_{A \Rightarrow B}$ is calculated in a similar manner.

2. In the second phase we prune out all those association rules $C_{A \Rightarrow B}$ where one of the following two conditions is true: i) A and B have an associative relation, ii) A and B are siblings (i.e. they both are a subclass of the same concept K).
3. In the third phase the transitive closure is inferred for the ontology O_i by using the relationships expressing subsumption. This is called an inferred model I_i. For example, if *university libraries* is a subclass of *academic libraries* and if *academic libraries* is a subclass of *libraries* then *university libraries* becomes also a subclass of *libraries*.
4. The inferred model I_i is used to prune out all such associative rules $C_{A \Rightarrow B}$ where the following condition is true: iii) A is a direct or inferred subclass of B.

2.2 The Finnish General Upper Ontology YSO

The Finnish General Upper Ontology YSO [6] is based on the Finnish General Thesaurus YSA maintained by the National Library of Finland. The first printed version of the YSA thesaurus was published in 1988 and today it is widely used for indexing purposes. YSO (ca. 20,000 concepts) is intended to be the main ontology in Finland, interlinking domain and instance ontologies.

The top ontology uses the idea of the DOLCE model [2]. The ontology includes relationships for explicating subsumption, partonomy and associative relationships. The existing subclass of relations inherited from the original thesaurus YSA have been checked and new relations added. Since the original thesaurus includes ambiguous words, in many cases new concepts have been added in order to disambiguate the different senses of the words. During the work, almost 1,000 concepts have been added to the ontology and the number of subclass of relations increased by nearly 6,000. The ontology work has been based on Finnish concepts, but the material also includes translated Swedish and English equivalents.

2.3 Cultural objects of semantic portal CULTURESAMPO

The materials analyzed in this research were annotations of cultural objects in the CULTURESAMPO [5]. The material consists of heterogeneous cultural content which originates from over ten cultural heritage organizations and are annotated using various ontologies including YSO.

CULTURESAMPO contains over 50,000 objects, for example paintings, photographs, et cetera from almost 100 different collections (e.g. from Finnish National Gallery). Over 20,000 of these objects are annotated using YSO and together 4,700 different YSO concepts are used. Other ontologies used in CULTURESAMPO annotations include e.g. SAPO (The Finnish Time-Location Ontology) [7], HISTO (History Ontology) [4] and ICONCLASS Ontology [3].

In this paper our research is limited to annotations where at least two concepts from YSO are used to describe the content of an object in a $keyconcept$-field. There are over 13,000 objects annotated this way.

3 Results

We applied the method for populating relationships between concepts in YSO ontology based on annotations of items in CULTURESAMPO. The results were ranked by using 1) only the support $S_{A\to B}$, or confidence $C_{A\to B}$ or by 2) multiplying them together in two different ways: $C_{A\to B} \cdot C_{B\to A}$ and $S_{A\to B} \cdot C_{A\to B} \cdot C_{B\to A}$. The hypothesis behind e.g. multiplication $C_{A\to B} \cdot C_{B\to A}$ was that those concepts that co-occur many times but do not occur that many times alone are very related.

The main ontology engineer behind the development of YSO checked the 50 best relationship candidates of all these different rankings and marked those that she thought should be included in YSO. She also told which type of relationship should be used, e.g. a subClassOf or an associativeRelation.

The concept relationships suggested by rank and agreed upon by human are depicted in Table 1. According to this test the formula $S_{A\to B} \cdot C_{A\to B} \cdot C_{B\to A}$ gave the best ranking as 23 out of 50 relationships it suggested should be added to the ontology.

	$S_{A\to B}$	$C_{A\to B}$	$C_{A\to B} \cdot C_{B\to A}$	$S_{A\to B} \cdot C_{A\to B} \cdot C_{B\to A}$
subClassOf	3	2	0	4
associativeRelation	9	11	22	19
Total	12	13	22	23

Table 1. The number of relationships (out of 50) an ontology engineer accepted as new relationships in the YSO ontology based on candidates found by different methods.

The method found many potentially useful relations, for example between “musical instruments” and “folk music” and between “works of art” and “art of painting”. “Musical instruments” was used in 385 annotations, “folk music” in 125 annotations and together they were used in 125 annotations. “Works of art” and “art of painting” were used together 58 times, “works of art” independently 109 times and “art of painting” independently 66 times. The ontology engineer suggested an associativeRelation for both of these cases.

The ONKI SKOS browser [11] was used to visualize the relationship extensions for an ontology engineer, see Figure 1. The user can also select which type relationship she wants to add between concepts and generate it with the tool.

4 Discussion and Related Work

The annotations in the CULTURESAMPO originate from many organizations. This means that many people have been participating in the annotation process. Thus the annotation set used in this experiment represents the wisdom of crowds (collective intelligence). There are differences between annotations because different organizations have their own customs and because annotation is a manual process. However, the final decision about which concepts are selected is made by a person. We used annotations made by this crowd for populating new relations to an ontology. To get results of how useful the

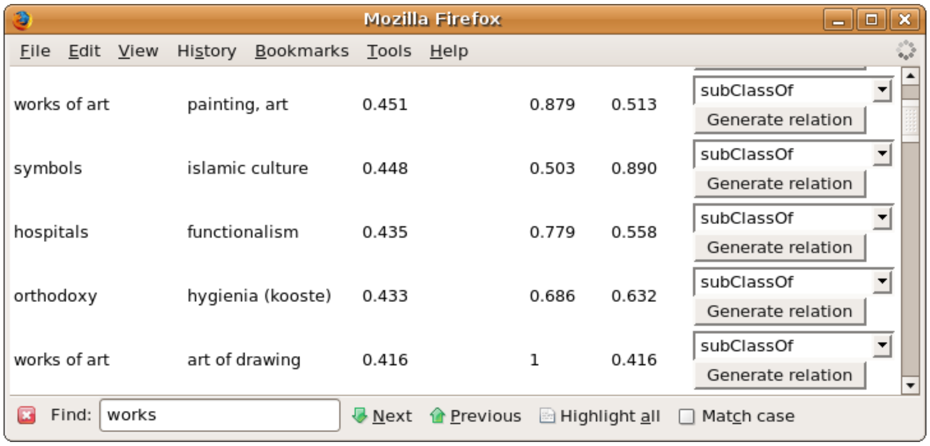

Fig. 1. The results visualized in ONKI SKOS browser.

method would be we let the main ontology engineer responsible for developing YSO ontology to analyze which of the different ranking methods produced the most interesting set of top-ranked relationships.

Schmitz et al. have examined how association rule mining can be adopted to analyze and to structure folksonomies, and how the results can be used for ontology learning and supporting emergent semantics [9]. They mentioned three purposes of use for association rules. First one was using them for helping the user to choose the tags which are most helpful in retrieving the resource later. Second one was using them to create a taxonomic structure. Unlike Schmitz we used association rule mining for analyzing annotations made using ontologies not by optional keywords. In our case an ontology already have a basic structure so our purpose is to improve an ontology by populating more relations between concepts. We also make use of ontology-based inference as a part of the method. The third purpose of using association rules was finding resources that have something in common.

5 Conclusions

In this paper we showed how an existing ontology, The Finnish General Upper Ontology YSO was populated by analyzing co-occurrences of concepts in annotations in the semantic portal CULTURESAMPO. According to the main developer of YSO a significant portion of the top ranked relationships should be added to the ontology. We also provided a visualization of relationships in the ONKI SKOS browser for further analysis.

Acknowledgments

We thank all colleagues in the Semantic Computing Research Group (SeCo) for fruitful discussions. We also thank anonymous reviewers and Agnieszka Ławrynowicz for valuable comments. Our research is done in the National Semantic Web Ontology Project in Finland [1] (FinnONTO) 2003–2007, 2008-2010 funded mainly by the Finnish Funding Agency for Technology and Innovation (Tekes) and in an EU project SmartMuseum[2] supported within the IST priority of the Seventh Framework Programme for Research and Technological Development.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., FebruaryJune–FebruaryAugust˜ 1993.
2. A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. In *Proceedings of EKAW 2002*, Siguenza, Spain, Jan. 2002.
3. C. Gordon. An introduction to Iconclass. In *Terminology for Museums. Proceedings of an International Conference*, pages 233–244, September 1988.
4. E. Hyvönen, O. Alm, and H. Kuittinen. Using an ontology of historical events in semantic portals for cultural heritage. In *Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, November 12 2007.
5. E. Hyvönen, T. Ruotsalo, T. Häggstöm, M. Salminen, M. Junnila, M. Virkkilä, M. Haaramo, E. Mäkelä, T. Kauppinen, and K. Viljanen. CultureSampo–Finnish culture on the semantic web: The vision and first results. In *In: K. Robering (ed.): Information Technology for the Virtual Museum. LIT Verlag, Berlin.*, 2008.
6. E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä. Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In *Proceedings of the European Semantic Web Conference ESWC 2008*. Springer, June 1-5 2008.
7. T. Kauppinen, J. Väätäinen, and E. Hyvönen. Creating and using geospatial ontology time series in a semantic cultural heritage portal. In *S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 ESWC 2008, LNCS 5021, Tenerife, Spain*, pages 110–123, June 1-5 2008.
8. P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, 2005.
9. C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In *Proceedings of the 10th International Federation of Classification Societies Conference*, pages 261–270. Springer, 2006.
10. J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday Publishing, New York, May 2004.
11. K. Viljanen, J. Tuominen, and E. Hyvönen. Publishing and using ontologies as mash-up services. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*, June 1-5 2008.

[1] http://www.seco.tkk.fi/projects/finnonto/
[2] http://smartmuseum.eu/