Rating the Naturalness of Ontology Taxonomies

Yoo Jung An, Kuo-chuan Huang and James Geller

Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102
{ya8, kh8}@njit.edu, geller@oak.njit.edu

Abstract

The quality of ontologies (QoO) is increasingly becoming a research issue on the Semantic Web. Ontology users may have difficulties locating the proper concepts in large ontologies, due to low quality. To quantify these problems, we use the notion of *naturalness.* In this paper we evaluate several existing important ontologies (WordNet, UMLS, etc.) to get numeric measures of naturalness. We concentrate on the question to what degree concept pairs connected by IS-A relationships are natural and therefore comprehensible to users.

1. INTRODUCTION

In Gruber's (1993) original work on ontologies, it was stressed that ontologies are about knowledge sharing. In two recent papers (Lee and Geller 2005; An et al. 2006) we have raised the question whether existing ontologies are constructed so that they may succeed at this task. Specifically in this paper, we are concentrating on the concept pairs used in the IS-A hierarchy of such an ontology.

We present an evaluation methodology for ontologies focusing on the *naturalness* of their IS-A hierarchies. Secondly, we propose an ontology maintenance model implied by this evaluation methodology.

If an ontology contains the relationship X IS-A Y then we would consider this a natural statement if we find many Web documents that contain both X and Y. If we find very few documents that contain both X and Y then we would consider X IS-A Y as an unnatural relationship. Obviously, if X and Y appear together, then they may stand in the relationship Y IS-A X or in some other relationship (part-of, connectedness, etc.) or they might not be related at all. However, if X and Y *rarely* occur together in documents then it is highly unlikely that they stand in an IS-A relationship. Furthermore, we are *not* mining IS-A relationships from Web pages. Rather, we use correct IS-A relationships, taken from popular ontologies, and check whether those IS-A relationships are *natural*. Thus, saying that an Accountant IS-A Professional appears intuitively natural, while saying that an Accountant IS-A Mammal appears unintuitive, even though it is biologically correct. Thus, we consider it more likely to find *many* Web documents containing both Professional and Accountant (5,930,000, according to Google) as opposed to containing both Mammal and Accountant (66,300).

Research (Brewster et al. 2003) has found that co-occurrences of two related concepts are not frequently observed in domain specific texts, so they suggested the Internet as partial textual source for constructing an ontology. Thus, our analysis is based on the number of search results of concept pairs reported by Google. For example, let X and Y be two concepts shown in an ontology in which X and Y are in an IS-A relationship. We obtain $Google\#(X \cap Y)$, called Concept Pair Google Number (CPGN) in this paper, indicating the search results when we query Google for both X and Y in one search. From each ontology that we are investigating, we randomly sample concepts X_i and extract their parents Y_j which are in IS-A relationships with X_i .

An Initial Experiment: We are assuming that if X_i and Y_j are indeed in an IS-A relationship then they are closely related and tend to co-occur in Web pages. In order to support our assumption, one simple experiment was conducted in which concept pairs in IS-A relationships and non-related concept pairs were compared with respect to their Google numbers. For each concept pair in an IS-A relationship (X_i, Y_j), we generated a non-related concept pair as $((X_i \ or \ Y_j), Z_k)$ where Z_k is a randomly selected concept. A group of these non-related concept pairs is the control group to verify the assumption that the naturalness of an IS-A relationship can be approximated by a Google search. The result of the comparison shows that submitting concept pairs with IS-A relationships to Google results in conspicuously higher numbers of frequency results than for non-IS-A-related concept pairs. So, we conclude that CPGN can be used as one measurement to distinguish pairs connected by IS-A relationships from random pairs. We are using the following ontologies/terminologies in our research: (1) UMLS Semantic Network (U.S. National Library of Medicine 2006), (2) WordNet (Princeton University 2006), (3) OpenCyc Class (Cycorp 2005), and (4) UMLS Metathesaurus (U.S. National Library of Medicine 2006).

For our statistical analysis, SAS software was used. In this paper we use, for example, the t-statistic (pooled and Sattherwaite) to determine whether significant differences exist between pairs or groups of ontologies.

2. RESULTS

Followings are the symbols used for ontologies: "W" = WordNet,"US" = UMLS Semantic Network, "UM" = UMLS Metathesaurus, and "OCC" = OpenCyc Class. In addition, we defined an additional set of pairs (X, Y), where X is derived from the concepts of the UMLS Metathesaurus and Y from the Semantic Types of the UMLS Semantic Network. We use UMS for these pairs.

Symbols used for statistical measurements include "M," the mean value of the number of search results for a concept, "SD," the standard deviation, "R," the Range (the difference between the maximum and the minimum), and "N," the sample size.

Table 1. Descriptive Statistics: Pair Occurrence

	W	US	UM
N	3091	135	5,000
M	86,643	323,777	3,752
SD	1,094,492	1,774,338	57,133
R	55,099,969	17,699,975	2,500,000

	UMS	OCC
N	6,249	7,787
M	1,410	7,867
SD	62,138	184,993
R	4,500,000	14,400,000

Table 1 shows the descriptive statistics. The variable is the reported number of search results when we send a concept pair to Google. The ontology with the highest mean value, 323,777 is the Semantic Network. On the other hand, the relationship between the Semantic Network and the Metathesaurus has the lowest mean value, 1,410.

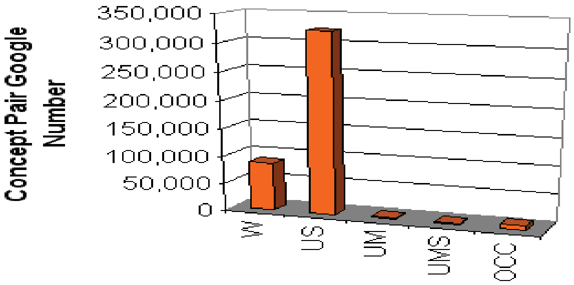

Figure 1. The naturalness of concept pairs by IS-A relationships

Figure 1 shows how natural each ontology is, with respect to concept pairs connected by IS-A relationships. Several t-tests were conducted to test whether the difference of means of two groups of ontologies is statistically significant at the 0.001 significance level with respect to the naturalness of concept pairs. (a) WordNet has a significantly higher naturalness than OpenCyc Class. (b) However, there is no significant difference between the naturalness of WordNet and the Semantic Network of the UMLS. (c) There is no significant difference between the naturalness of OpenCyc Class and the Metathesaurus. (d) WordNet has a significantly higher naturalness than the associations between Metathesaurus and Semantic Network (UMS). (e) OpenCyc Class has a significantly higher naturalness than the associations of UMS.

The evaluation model presented can be applied to ontology maintenance. If an ontology has a lower mean value than another ontology in the same domain, the quality of the ontology can be improved by repairing the concept pairs whose naturalness is below a specific threshold value. This may be done in different ways, including placing intermediate concepts and IS-A links in the hierarchy. Thus, to use an extreme example for the purpose of demonstration, we might place "Human" and "Professional" between "Mammal" and "Accountant." This procedure can be simplified when the necessary synonyms, hypernyms or hyponyms are derived from WordNet or a domain-specific 'gold standard' ontology if such an ontology exists (Castano & Antonellis 1999).

References

An, Y. J.; Huang, K.C.; and Geller, J. 2006. Naturalness of Ontology Concepts for Rating Aspects of the Semantic Web. *Communications of the International Information Management Association* 6(3): 63–76.

Brewster, C.; Ciravegna, F.; and Wilks, Y. 2003. Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance. In *Proceedings of the Semantic Web Workshop (SIGIR)*.

Castano, S., and Antonellis, V. D. 1999. A Discovery-Based Approach to Database Ontology Design. *Distributed and Parallel Databases - Special Issue on Ontologies and Databases* 7(1): 67–98.

Cycorp. 2005. OpenCyc. Retrieved June 16, 2006, from http://www.opencyc.org/releases/

Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2): 199–220.

Lee, Y., and Geller, J. 2005. Semantic Enrichment for Medical Ontologies. *Journal of Biomedical Informatics* 39(2): 209–226.

U.S. National Library of Medicine. 2006. Unified Medical Language System. Retrieved June 13, 2006, from http://umlsinfo.nlm.nih.gov/

Princeton University. 2006. WordNet 2.1. [Computer Program]. Retrieved June 13, 2006, from http://wordnet.princeton.edu/