

Towards Browsing Distant Metadata with Semantic Signatures

Andrew Choi and Marek Hatala

Simon Fraser University

School of Interactive Arts and Technology

Simon Fraser University, 2400 Central City, Surrey, BC, Canada

{aschoi, mhatala}@sfu.ca

Abstract

We present a light-weighted ontology mediation method allowing users to query distant repositories with local concepts. As repositories annotate their objects with different ontologies or taxonomies this hinders the search of objects by users familiar with their own local conceptual structures. The proposed method indexes concepts with signatures composed from WordNet senses. Instead of sending the local concepts to the distant repositories, the query contains their WordNet signatures that are matched against those of distant ontological concepts. The evaluation performed on the real world data in eight different domains demonstrates substantially increased precision and recall against keyword based retrieval.

1 Introduction

With the advance of the Internet and rapid development in E-learning, more and more institutions are joining to form distributed learning networks to allow users to access resources in different learning repositories [Stojanovic *et al.*, 2001]. This creates pressure for institutions to provide an efficient way to organize a huge volume of material located in different repositories. Currently, the use of metadata and ontologies to formalize semantics of concepts in the E-learning domain does not completely resolve the problem of interoperability in a federated environment [Hatala *et al.*, 2004]. This is because metadata of learning resources used in different repositories are very often annotated with concepts defined by different taxonomies or ontologies that are specific to their organizations. This makes finding information based on a local conceptual structure difficult. The well known problems of semantic ambiguity are at play: different organizations with different backgrounds and target audiences may use different terms with similar semantics to define and describe two similar learning resources. In addition to ontological differences, linguistic variations in metadata values and lack of standardized metadata format across learning networks make direct querying with keywords

sometimes ineffective to discover a conceptually similar metadata.

To help distributed learning repositories to organize and manage their metadata in compliance with a global semantic view, we create a semantic mapping strategy using WordNet [Miller *et al.*, 1990]. In our approach WordNet serves as a mediator providing a word sense disambiguation mechanism to generate semantic signature to represent ontological concepts with respect to the local context. The method was implemented into the research prototype. The prototype consumes XML-based metadata as an input, performs semantic analysis of metadata annotated with the same concept, and as a result generates a semantic signature in the form of WordNet senses. Finally, the signatures are used to create inverted index of ontological concepts and indirectly learning resources. The system architecture consists of three components: *Signature Generator*, *Signature Index Database*, *Concept Browser*.

2 Semantic Signature Generation

Semantic signature of an ontological concept can be defined as a logical grouping of representational word senses. In essence, it is a semantic representation of a the concept with important WordNet senses with respect to the local context that is in our case manifested by the metadata records.

To briefly explain, semantic signature of a concept is built from a set of important document senses from all documents (metadata records) annotated with a particular concept. In turn, document senses are generated from a collection of the best WordNet senses for all representational keywords for a particular document.

The system generates the semantic signatures in the three step process:

1. *Document preprocessing and word extraction.* Taking a collection of metadata records annotated with the same concept and specified stopword list as an input, this module returns a set of most significant nouns and binary noun phrases (called keywords).

2. *Document sensitization.* Taking a set of keywords as input this function returns corresponding WordNet word senses for each keyword.
3. *Senses Selection Strategy.* Given the set of retrieved word senses this function selects the best word sense to represent each keyword based on its context.

Once semantic information has been found for each metadata record the signatures for the ontological concepts are computed. The senses in the signatures annotated by the same ontological concept are aggregated using TFIDF weighting scheme to form the *concept signature*. The concept signatures are used to reverse-index the records that are annotated with corresponding concepts.

The third component is a concept browser that contains a signature similarity calculator. When user specifies the query referring to the local concept this is substituted with the concept's signature and the query is distributed in the network. The repositories receiving the query compute the similarity between received signature and their own signatures in the database using the cosine distance and return the records annotated with the most similar concepts.

At the end, the system provides an integrated approach to index concepts in the form of semantic signatures and enables the users to search distant repositories using their local ontological definitions.

3 Evaluation

The proposed method has been compared with keyword-based retrieval in three simulated databases.

Dataset. 2235 metadata records annotated with 8 different concepts were acquired from a number of different sources. Table 1 summarizes the dataset used for testing. The dataset was partitioned into training and testing groups. The 'local' repository stores the training dataset while two simulated remote repositories store the testing datasets. The concepts of all metadata records were known upfront.

Table 1 Source and category of metadata records

Category	Source	No. of metadata
<i>Accounting</i>	Business Source Premier Publications	382
<i>Biology</i>	Biological and Agricultural Index, BioMed Central Online Journals	315
<i>Computing Science</i>	Citeseer	320
<i>Economics</i>	American Economic Association's electronic database	353
<i>Education</i>	Educational Resource Information Center	307
<i>Geography</i>	Geobase	237
<i>Mathematics</i>	arXiv.org, MathSciNet	157
<i>Psychology</i>	PsycINFO, ERIC	164

Table 1. Comparison on precision and recall for concept retrieval

Category	Precision		Recall		F-measure	
	S	K	S	K	S	K
<i>Accounting</i>	1.00	0.67	1.00	0.75	1.00	0.71
<i>Biology</i>	0.75	0.75	0.75	0.75	0.75	0.75
<i>Computing Sci</i>	1.00	0.50	1.00	0.50	1.00	0.50
<i>Economic</i>	1.00	0.75	1.00	0.75	1.00	0.86
<i>Education</i>	1.00	0.50	1.00	0.75	1.00	0.45
<i>Geography</i>	0.75	0.50	0.75	0.50	0.75	0.50
<i>Mathematics</i>	0.67	0.33	0.67	0.50	0.67	0.40
<i>Psychology</i>	0.67	0.33	0.67	0.67	0.67	0.44
<i>Average</i>	0.86	0.54	0.86	0.65	0.86	0.58

S = Signature-based retrieval, K = Keywords-based

Results. As shown in Table 2, the use of semantic signature for indexing and browsing query can consistently improve retrieval relevance in terms of recall and precision. In all categories, the semantic based retrieval outperformed the keywords-based retrieval.

4 Conclusions

This project offers empirical evidence that the use of semantic signatures can enhance relevance in concept-based retrieval in distributed environment without prior knowledge of remote conceptual model. In other words, this approach enables a light-weighted semantic mediation between remote ontologies used for annotation of objects in the repositories. In comparison to other approaches to the semantic mapping our method does not require intensive user involvement to provide mapping information, nor resorts to complicated heuristics or rule-based machine learning. Although more rigorous evaluation especially in specific domains is needed the approach shows some promise.

References

- [Stojanovic et al., 2001] L. Stojanovic, S. Staab and R. Studer, "eLearning based on the SemanticWeb," in Proceedings of WebNet 2001, 2001, pp. 191-201.
- [Voorhees, 1993] E.M. Voorhees, "Using WordNet to disambiguate word senses for text retrieval," in SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 1993, pp. 171-180.
- [Hatala et al., 2004] Hatala, M., Richards, G., Eap, T., Willms, J.: The Interoperability of Learning Object Repositories and Services: Standards, Implementations and Lessons Learned. 13th World Wide Web Conference, Educational Track, New York, May 2004, pp.19-27.
- [Miller et al., 1990] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. and Miller, "Wordnet: an on-line lexical database," International Journal of Lexicography, vol. 3, 1990, pp. 235-244.