Towards Browsing Distant Metadata with Semantic Signature

Andrew Choi
Simon Fraser University
School of Interactive Arts and Technology
Surrey, BC, Canada
aschoi@sfu.ca

Marek Hatala
Simon Fraser University
School of Interactive Arts and Technology
Surrey, BC, Canada
mhatala@sfu.ca

Abstract

In this document, we describe a light-weighted ontology mediation method that allows users to send semantic queries to distant data repositories to browse for learning object metadata. In a collaborative E-learning community, member data repositories might use different ontologies to control a set of vocabularies describing topics in learning resources. This could hinder the search of learning resources when search queries are built according to local ontological concepts. With the use of WordNet, we develop a toolkit that index ontological concepts with WordNet senses. This enables semantic browsing of distant metadata in a distributed learning community. The effectiveness of the toolkit is validated with real world data in a specific domain, namely E-learning metadata.

1 Introduction

As the advance of the Internet and rapid development in E-learning, more and more institutions are joining to form distributed learning network to allow users to access resources from different learning repositories [1]. This creates pressure for institutions to provide an efficient way to organize a huge volume of materials located in different repositories, according to a consistent concept classification, in order to answer distributed retrieval requests. Currently, the use of metadata and ontologies to formalize semantic of concepts in the E-learning domain do not completely resolve the problem of interoperability in a federated environment [3]. This is because metadata of learning resources resided in different repositories are very often annotated with concepts defined by different ontologies specific to their organizations. That makes finding information based on a local conceptual framework difficult. 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 difference, linguistic variations in metadata values and the lack of a standardized metadata format across learning network makes direct querying with keywords sometimes ineffective to discover a conceptually similar metadata.

Can we find a way to integrate heterogeneous learning resources metadata semantically, in order to enable semantic concept browsing? In this poster, a research system prototype, called Learning Resource Concept Browser (LRCB), is presented. It takes XML-based metadata as the input and performs a semantic analysis to generate a semantic signature in WordNet senses to index learning resources categories. Its system architecture consists of three components: *Signature Generator*, *Signature Index database*, *Concept Browser*.

2 Learning Resource Concept Browser

The first and also the key component of LRCB is a *semantic signature generator*, which contains the following three functions:

- **Document preprocessing and word extraction**.
  Taking a collection of metadata and a specified stop-word list as the input, this function returns a set of the most significant nouns and binary noun phrases.
- **Document sensitization**
  Taking a set of important keywords as the inputs, this function returns corresponding WordNet word senses for all keywords.
- **Senses Selection Strategy**
  Given the set of retrieved word senses, this function selects the best word sense based on the local context to represent each keyword.

After all semantic information has been found for a document, the second component *signature index database* aggregates them to form the document signature. By the same token, all document signatures will finally be aggregated based on TFIDF [4] weighting scheme to form the concept signature. The signature will be stored in the database and used to index the concept.

The third component is a *concept browser* that contains a signature calculator. The signature calculator is used to compute similarity between signatures to determine “closeness” in concepts. This component also provides with user interface for federated concept browsing.

At the end, LRCB provides an integrated approach to index concepts in semantic signature and enable users to browse for similar concepts based on local ontological definition.

3 Evaluation and Result

To verify the utility of the prototype, a simulated distribute concept browsing is performed to compare semantic browsing to both classical keyword-based and label-matching based browsing. Three simulation databases are set up. They are called “*local*”, “*remote1*”, and “*remote2*”.

3.1 Dataset

Metadata are acquired through a number of different sources. Table 1 shows the category of metadata acquired and their respective sources. In total, 2235 metadata subdivided into the 8 different categories are acquired. The dataset is partitioned into training and testing groups. The local database stores the training dataset while remote1 and remote2 store the testing dataset. All metadata are known with their class label. Metadata are distributed randomly, using Microsoft Excel random generator.

Table 1 Source and Category of Metadata

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

3.2 Result

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 out perform both keywords-based retrieval and label-matching retrieval.

4 Conclusion

This project offers an empirical evidence to use semantic signature to enhance relevance in concept retrieval in distributed E-learning environment. In other words, this gives a new light-weighted semantic (ontology) mediation approach to enable cross platform concept browsing in a federated network. Unlike many current practices in semantic mapping that either requires intensive user involvement to provide mapping information, or resort to complicated heuristic or rule-based machine learning approach, this work shows an effective automatic mapping protocol that can allow federated concept browsing with semantic signature. It is evident from the experimental results that it is effective to use WordNet to provide semantic knowledge for metadata classification in the domain of E-learning. The merits include the provision of semantic representation of categorical data and increased semantic relevance in categorical browsing.

Table 2 Comparison on precision, recall on concept retrieval

Category	Precision			Recall			F-Measure		
	S	*K*	*L*	*S*	*K*	*L*	*S*	*K*	*L*
Acc	1	0.67	0.50	1	0.75	0.50	1	0.71	0.50
Bio	0.75	0.75	0.50	0.75	0.75	0.50	0.75	0.75	0.50
CS	1	0.50	0.33	1	0.50	0.33	1	0.50	0.33
Econ	1	0.75	0.67	1	0.75	0.67	1	0.86	0.67
Educ	1	0.50	0.50	1	0.75	0.50	1	0.45	0.50
Geo	0.75	0.50	0.50	0.75	0.50	0.50	0.75	0.50	0.50
Math	0.67	0.33	0.67	0.67	0.50	0.67	0.67	0.40	0.67
Psy	0.67	0.33	0.33	0.67	0.67	0.33	0.67	0.44	0.33

S = Signature-based retrieval
K = Keywords-based retrieval
L = Label-matching retrieval

5 Future Direction

We are planning to acquire a larger dataset for further testing. In addition, we would like to convert the current application into a web service to allow easy access, with more vigorous text and natural language processing.

Acknowledgments

This work was partially funded by Canada's LORNET NSERC Research Network.

References

[1] L. Stojanovic, S. Staab and R. Studer, "eLearning based on the SemanticWeb," in Proceedings of WebNet 2001, 2001, pp. 191-201.

[2] 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.

[3] G. Richards and M. Hatala, "Interoperability Framework for Learning Object Repositories," LORE., Simon Fraser University, Tech. Rep. May, 2003, 2003.

[4] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," Cornell University., 1987.