Towards Browsing Distant Metadata Using Semantic Signatures

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 based on local ontological concepts. With the use of WordNet, we develop a toolkit that indexes ontological concepts with WordNet senses for semantic browsing in order to integrate information in a distributed learning community. The effectiveness of the toolkit was validated with real-world data in a specific domain, namely E-learning metadata.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – information integration, retrieval models, search process

General Terms

Algorithms, Management, Experimentation, Verification

Keywords

Semantic Retrieval, Data Integration, Ontology Mediation

INTRODUCTION

As the advance of the Internet and rapid development in E-learning, more and more institutions are joining to form a distributed learning network to allow users to access resources from different learning repositories. 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 semantics of concepts in the E-learning domain does not completely resolve the problem of interoperability in a federated environment. This is because metadata in different repositories are very often annotated with concepts defined by different ontologies specific to their organizations or communities. That makes finding information based on a local conceptual framework difficult. Different organizations with different backgrounds and target audience 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 use of metadata standard across learning network makes direct querying with keywords sometimes ineffective to discover a conceptually similar metadata.

PROBLEM DESCRIPTION

The primary objective of this research is to explore the use of semantic signatures expressed in WordNet senses to provide mediation between different ontologies in order to enhance concept retrieval. Consider the scenario when the learner L_1 associated with the repository R_1 looking for learning resources related to the topic of how to find a good bass musical instrument, L_1 sends out a request "*search for bass*" to remote repositories R_2 and R_3 respectively in an E-learning network. However, the returned results from them are mixed with many irrelevant resources related to catching a bass (e.g. fish). Such a problem occurs frequently when the concepts are defined by different domain ontologies with different sets of vocabularies carrying different intended meanings. Imagine another case when the same learner L_1 sends out a distributed request for learning resources on the topic "*advance databases*". Since the topic is annotated by the concept "*database systems II*" in remote repositories, that is to say it is labelled differently. Therefore, in a concept-based label matching search, learning resources defined by the concept "*database systems II*" will not be returned for the request of "*advance databases*" even though the two concepts are actually semantically equivalent.

From these simple scenarios, one can easily see that without a proper semantic mapping between ontologies in heterogeneous data sources, even with the ontology to

define vocabulary used to describe metadata on learning resources, it is still challenging to find learning resources based on the local conceptual definition.

OVERVIEW OF ONTOLOGY MAPPING

Semantic or ontology mapping can be described as a mapping task that identifies common concepts and establishes semantic relationships between heterogeneous data models in the same domain of discourse [1]. Since semantics is mostly defined by ontological constructs in modern knowledge systems, we will use the term semantic mapping interchangeably with ontology mapping in this discussion. According to [10], ontology mapping between two ontologies O_1 and O_2, can be expressed as a mathematical structure: $O_1 = (C_1, A_1)$ to $O_2 = (C_2, A_2)$ by a function f: $C_1 \rightarrow C_2$ to semantically related concept C_1 to concept C_2 such that $A_2 |= f(A_1)$ whose all interpretations that satisfy axioms in O_2 also satisfy axioms in O_1. For example, if the concept *agent* (C_1) is defined in O_1 by a set of properties such as *<broker, travel agent* and *officer>* with axioms such as *<part-of agency, is-a individual, is-a organization* and *type-of communicator>* (ignoring other attributes and cardinality for the sake of simplicity), it is possible to map it to a concept *representative* (C_2) defined in O_2 with a set of properties such as *<government agent, client, spokesperson* and *advisor>* and having axioms such as *<part-of government, is-a person*, and *is-a expert>*. This assumes that all the semantic interpretations of C_1 will be respected by C_2 in the domain of discourse when executing logical inference operation on C_2.

REVIEW OF OTHER APPROACHES

This section presents a brief overview of two approaches on semantic mapping. The two selected approaches are GLUE and MAFRA. The former is a system that employs machine-learning techniques to find ontology mappings with the use of probabilistic multiple learners while the latter uses a declarative representation of mappings as instances in a mapping ontology defining bridging axioms to encode transformation rules. With two domain ontologies, for each concept in an ontology GLUE claims to find the most similar concept in another ontology [7]. A number of features distinct GLUE from other similar mapping systems. First, unlike many mapping systems that only incorporate single similarity function to determine if two concepts are semantically related, GLUE utilizes multiple similarity functions to measure the closeness of two concepts based on the purpose of the mapping. The intuition behind the multiple similarity functions is to take advantage of the mapping requirement to relax or limit the choice of corresponding concepts. For instance, based on the requirement of the application the task of mapping the concept *"associate professor"* can be satisfied by similarity criteria "exact", "most-specific-parent" or "most-general-child" similarity criteria to find *"senior lecturer"*, *"academic staff"* or *"John Cunningham"* respectively. This

gives GLUE flexibility to find semantic mappings between ontologies. Second, GLUE applies a multi-strategy learning approach to use certain information discovered by different classifiers during the training process. This approach divides the classification process into two phases. First, a set of base classifiers is developed to classify instances of concepts on different attributes with different algorithms. Then, the prediction of these base classifiers, assigned with different weights representing their importance on overall accuracy, is combined to form a meta-learner. Finally, the classification is determined by the result from the meta-learner. As an instance, one base learner can exploits the frequency of words in the name property using a Naïve Bayes learning technique while another base learner can use pattern matching on another property using a Decision Tree Induction technique. At the end, the meta-learner will gather all the results to form the final prediction. Using multiple classifiers, GLUE intends to increase the accuracy of the overall prediction. Third, GLUE incorporates label relaxation techniques into the matching process to boost the matching opportunity based on features of the neighbouring nodes. Generally, the relaxation labelling iteratively makes use of neighbouring features, domain constraints and heuristic knowledge to assign the label of the target node.

MAFRA (Mapping FRAmework) is another ontology mapping methodology that prescribes "all phases of the ontology mapping process, including analysis, specification, representation, execution and evolution" [14]. It uses the declarative representation approach in ontology mapping by creating a Semantic Bridging Ontology (SBO) that contains all concept mappings and associated transformation rule information. In this model, given two ontologies (source and target), it requires domain experts to examine and analyze the class definitions, properties, relations and attributes to determine the corresponding mapping and transformation method. Then, all accumulated information will be encoded into concepts in SBO. Therefore, SBO serves as an upper ontology to govern the mapping and transformation between two ontologies. Each concept in SBO consists of five dimensions: they are *Entity*, *Cardinality*, *Structural*, *Constraint* and *Transformation*. During the process of ontology mapping, software agent will inspect the values from two given ontologies under these dimensions and execute the transformation process when constraints are satisfied.

Some recent approaches like INRIA[1] make use of OWL API to build a set of alignment APIs with built-in WordNet function for the purpose of ontology alignment or axioms generation and transformations. However, the details on the use of WordNet to generate the alignments are not well documented in the published literatures.

---

[1] http://co4.inrialpes.fr/align/index.html

WORDNET

WordNet is a widely recognized online lexical reference system, developed at Princeton University, whose design is inspired by "current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synsets (synonym sets), each representing one underlying lexical concept that is semantically identical to each other" [2]. Synsets are interlinked via relationships such as synonymy and antonymy, hypernymy and hyponymy (*Subclass-Of* and *Superclass-Of*), meronymy and holonymy (*Part-Of* and *Has-a*) [3]. Each synset has a unique identifier (ID) and a specific definition. A synset may consist of only a single element, or it may have many elements all describing the same concept. Each element in a particular synset's list is synonymous with all other elements in that synset. For example, the synset {World Wide Web, WWW, Web} represents the concept of computer network consisting of a collection of internet sites. In this context, 'World Wide Web', 'WWW' and 'Web' are all semantically equivalent. For cases where a single word has multiple meanings (polysemy), multiple separate and potentially unrelated synsets will contain the same word. For instance, the word 'Web' can have 7 multiple meanings defined in WordNet as computer network, entanglement, simply spider web and etc.

OUR APPROACH

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 as a mediator to provide word sense disambiguation and to generate semantic signature each representing learning resource category.

Semantic signature in the categorical browsing context can be defined as a logical grouping of representational word senses for a class of metadata. In essence, it is a semantic representation of a class label with important WordNet senses regarding context. To formalize the concept of semantic signature, it can be written as follows:

$$Sig(c) = \bigcup_{j=1}^{n} DS_j = \bigcup_{i=1}^{n} BS_{di} \quad BS_{di} = Max\{Fav(d_j, s), \{t \in T \mid s \in WS(t)\}\}$$

where $Sig(c)$ = semantic signature for class c

DS_j = set of document senses for class c

BS_{di} = set of best sense in document d_j

T = all keywords in document d_j

Fav = selection function to find best sense

$WS(t)$ = set of WordNet sense for term t_i

To briefly explain, semantic signature of a class of metadata is built from a set of important document senses from all documents (metadata records) belonging to a particular class. In turn, document senses are generated
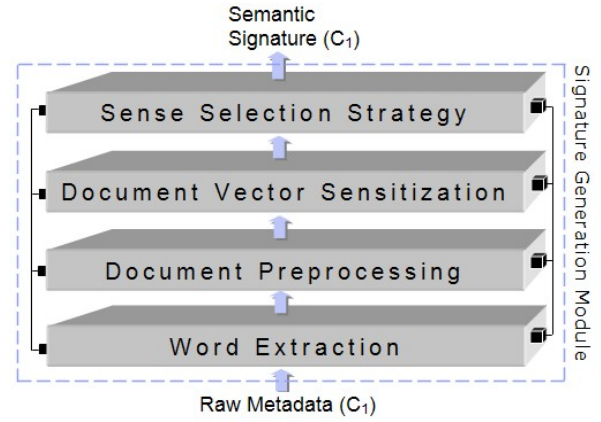

Figure 1. Semantic Signature Generation Framework

from a collection of the best WordNet senses for all representational keywords for a particular document.

The generation of a semantic signature for a class of metadata is divided into three distinct phases. In the rest of this section, the general architecture of the methodology is described while each phase is discussed in detail and as well as illustrated with examples.

System Design and Architecture

The methodology for creating semantic signature relies heavily on the assumptions that the aggregates of all semantic information from metadata records of a particular class are a good representation of the concept for that class. In fact, the metadata record is an instance of a concept in the ontological framework. Moreover, the methodology assumes that semantic information of a class can be approximated by a set of important word senses from all metadata records. Besides, semantic word senses specific to the context can be found based on important terms extracted from metadata through WordNet. Finally yet importantly, it assumes that the local semantic signature for a class of metadata is similar to signatures for metadata of semantically equivalent concepts in distant repositories. The methodology uses k-Nearest Neighbour (kNN) search algorithm to classify semantically relevant concepts in distant repositories based on local semantic signatures [11]. The instances (metadata) of concepts in local repository serve as the training dataset. Based on semantic features of the local metadata, semantic signatures for each class of concepts are formed. To find semantically relevant concepts in distant repositories, a distance function is defined and used to measure closeness between the query signature and semantic signatures for concepts in distant repositories. Eventually, k most similar concepts to the query signature will be retrieved from remote repositories.

Figure 1 shows the four phases of the semantic signature generation framework. In the Word Extraction phase representative features are extracted from each metadata document. The Document Preprocessing phase eliminates all irrelevant information as well as all non-noun words. In the Document Vector Sensitization phase all the

representative keywords are used as seeds to find the corresponding word senses from WordNet. Finally, in the Sense Selection phase several strategies are applied to select the best word sense is selected among all senses to represent each word term.

Signature Generation in Action

Phase I: Word Extraction

First, the input metadata are transformed to comply with the IEEE LOM standard[2] using XML transformer. Then, adapted from Edmundsonian paradigm [4], content from *<Title>* and *<Description>* elements is extracted to represent the whole metadata document. That presumes that the content from these two elements carry important weight as cue phrase to be able to represent the whole document [4]. This view seems reasonable in the case of learning object metadata because other elements like *publication date*, *ISBN* or *format* do not bear good semantic information to signify the category of the metadata.

Phase II: Document Preprocessing

The condensed metadata with only the <Title> and <Description> elements are subjected to cleaning to remove all stopwords, punctuation information, numerical values and irregular symbols. Next, all non-noun words are removed using part-of-speech tagger except some commonly used phrasal words which carry specific meaning. For example, the word "artificial" in the phrase "artificial intelligence" will be preserved to retain the special meaning of the binary phrase in the branch of computer science. The reason why this approach only uses nouns as the base keyword is explained in [5] where it is said that long phrases are not easily disambiguated comparing to a single word term or a binary word term. The accuracy to use a phrase as a distinguishing feature for a document classification in effect will be lower through previous experiments demonstrated in [6]. On the other hand, it has been shown that the use of noun word terms carry the most salient expression to serve as distinguishing feature for doing text classification [7].

Phase III: Document Vector Sensitization

Supposing that all irrelevant information has been eliminated, the physical metadata documents are projected onto the vector space model. The document vector becomes a logical representation of the physical metadata record. Then, using TFIDF weighting scheme we select most significant terms across all document vectors to represent a category of metadata [12]. After that, each word term with the TFIDF score higher than the threshold is sent to WordNet to retrieve the corresponding word senses and its definition. The threshold is determined by trial and error approach with a test run. A single word term can have

---

multiple word senses retrieved. For example, the word "search" can be mapped to WordNet senses as <hunting, hunt>, <lookup> and <investigation>. Because of this, the mapping information of a single noun word term can be denoted by a triple construct in the form <T, S, D> where T is the original word term, S is the synset of T and D is the definition of T. When a noun term can be mapped to multiple senses, there will be multiple triples. Take the word term "search" as an example. After the sensitization, it becomes <search – {hunting. hunt} = "the activity of looking thoroughly in order to find something or someone" (TFIDF 0.623101)> in triple construct. The triple construct format is used to substitute the original word term in the master document vector. Then again, recall that since a single word term could be mapped to possible different word senses through WordNet. Each word sense is represented in synset which may have multiple synonymous terms. Because of this, the length of the document vector in word sense will grow considerably. This problem is addressed in the next phase.

Phase IV: Sense Selection Strategy (S^3)

This is the last, and the most crucial phase in the method. It chooses the best word sense among all retrieved word senses from WordNet to represent the word term. As stated, a word term can be mapped to multiple WordNet senses. In such a case, the dimensionality of the vector grows significantly after the sensitization procedure. Imagine that a word term "light" can be mapped to 15 WordNet noun senses "visible light", "light source", "luminosity", "lighting", etc. The growth ratio is 15 times in this case. Such a high dimension not only negatively affects the efficiency of the similarity computation, but more seriously, the many senses are noise which does not carry actual meaning of the word in the *context* of the document. Included irrelevant senses will distort the semantic representation of the signature and lower the accuracy in similarity calculation when finding similar classes of metadata using signature matching. On the other hand, from the semantic knowledge standpoint, WordNet senses only provide the lexical information of the word term, but not the contextual information to determine how the meaning is clarified in a specified context [8]. Without that, the semantic signature is just a bigger collection of keywords and would have small use in identifying the classes of metadata based on the semantic relevance of the signature. Therefore, it is necessary to find a way to reduce the dimension and only select the sense that conveys the main idea of the word in the current context. To select the best sense representing a word term, a contextual-based Senses Selection Strategy (S^3) is applied to retrieved word senses. The strategy is based on the assumption that the local contextual information of a document serves as a good hint to tell which sense represents the actual meaning of the word term best. The S^3 approach can be summarized in the following algorithm:

Steps of algorithm (Calculate the best senses for class C1):

For each metadata document $D \in C_1$

  Get the list of synsets for each word term $T_1 \in D$

  For each synset Syn_1 of the word term T_1

   For each sense term $S_i \in Syn_1$

    1. Compute associative frequency af for S_i to other senses S_k $\in Syn_k$, $Syn_k \subseteq T_k$ and $T_1 \neq T_k$

      1.1 Find the sense Sl with highest score Max(af)

      1.2 If (Max(af) < 1) then go to 2 otherwise stop and return S_l

    2. Compute associative frequency af for S_i to k-order parent senses $PS_k \in P(Syn_k)$, $P(Syn_k) \subseteq T_k$ and $T_1 \neq T_k$

      2.1 Find the sense S_p with highest score Max(af)

      2.2 If (Max(af) < 1) then go to 3 otherwise stop and return S_p

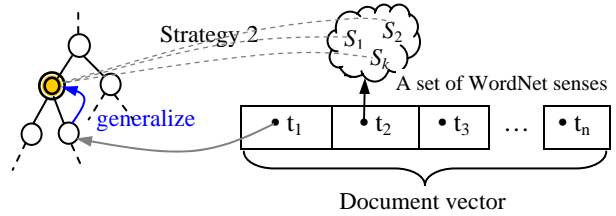   3. Return the most popular sense S_w offered by WordNet

  Return the Best Sense to represent word term T_1

Aggregate all sense from all important word terms to represent signature of the document D

The algorithm works in the following way. For each word sense of a word term, it first computes the associative frequency (af) of each sense term in a synset to other sense terms in other synset of other word terms in the same document. From this, the most occurred word sense will be used to substitute the semantic representation of the word term.

Next, if the word sense of a word term cannot be discriminated by Strategy 1, the algorithm generalizes the word term to the k-order parent senses. In this approach, the value of k is 1. Hence, it generalizes to its immediate parent word sense. Referring to Figure 2, Strategy 2 will use the immediate parent sense to compute the associative frequency against other senses from other word terms in the document vector. As such, in this example the word term t_1 will be rolled up to its immediate parent through hypernym (is-a) relation in the WordNet hierarchy. Then, the parent's synset is used to calculate the associative frequency to other word senses for other word terms. Unlike other generalization approaches [7, 13], we generalize the sense to its most-specific parent only. The reason why it uses immediate parent senses (k=1) to compute the associative frequency is given in [9] where the most specific parent in a hierarchical terminology has a higher distinctive power to classify the topic. Essentially following the intuition that if a word sense is generalized to higher order parent sense than k=1, the generalized sense may be too general and becomes incoherent to local context, and would become noise when used to classify metadata.

Finally, as arranged by WordNet, the word senses retrieved from WordNet for a particular word are a partial order set ranked by popularity in English usage. If the previous two strategies can not find the best sense to represent the word term, then the most popular sense offered by WordNet will be adopted in Strategy 3.

Figure 2. Compute associative frequency between immediate parent with other word sense

The rationale behind sequencing three strategies is based on observations and hypothesis that the local context is the most specific and relevant candidate to provide contextual meaning for the word term sense. Therefore, a word sense for a particular term can most likely be disambiguated by other local senses (Strategy 1). If it could not be resolved by step 1, then it compares the immediate parent sense to the other word senses to check if the parent sense is a frequently occurring sense for the underlying word term. At last, the most popular sense is adopted to represent the semantic meaning for a word term when the two strategies above could not resolve the ambiguity of the word term.

Following the above procedures, a set of senses becomes a semantic signature of a document. In order to generate the final semantic signature for a class of documents referring to particular concept, TFIDF scheme is applied again to each word sense in all document signatures for a particular class. Based on the score, the most relevant senses for characterizing the class of metadata are aggregated to form the final signature for the class.
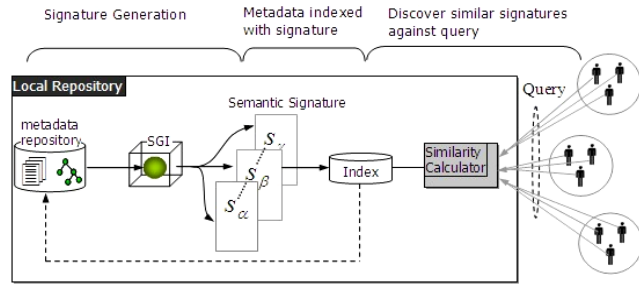
Concept browsing in heterogeneous ontologies

In our application, the generated semantic signatures are used to index the actual classes of metadata for fast distributed browsing. We developed a tool called Signature Generation Indexer (SGI) that supports the methodology described in the previous section. Focusing on the efficiency, the design of SGI is to allow repository operators to produce semantic signatures for classes of learning object metadata easily without tedious human interaction, or complicated implementation.

The ultimate goal is to achieve semantic search based on E-learning topics defined by heterogeneous ontologies in a federated network. In a collaborative learning environment, users expect to be able to access all the learning resources within the learning network. To fulfill this anticipation, it is important to assume that all participant repositories in the collaborative network employ the same strategy to index learning resources metadata with WordNet semantic signature.

In this way, when users launches a query by selecting a specific topic (concept) from the local ontology (e.g. via user interface), the corresponding semantic signature representing the topic is retrieved from local database. The signature is then sent across the network to participating

Figure 3. Integrated process of semantic-based browsing of metadata

learning repositories. The query in the form of semantic signature is the input of the Similarity Calculator in distant repositories. The Similarity Calculator is used to compute the similarity of signatures in each of the learning repositories. The similarity calculator uses the cosine similarity function, thereby the more matched elements in the signature, the higher the score is. In calculating the similarity score, different weights are assigned to senses from <Title> and <Description> in which the match in the title sense gets higher contribution to overall score than the one from the description tag.

In order to ensure the global accuracy of the result, results from participating remote repositories are merged and sorted in the descending order based on the cosine similarity score. Then, the top k (k=5) topics of the metadata are offered as the answer to the local query. The overall operation of the semantic-based browsing of learning resources metadata is shown in Figure 3.

IMPLEMENTATION

The SGI is implemented in the C# programming language. The current version is a desktop application, but it can be easily extended to a web service. The goal of SGI is to integrate signature generation, document indexing and browsing capability. The signature indexes are stored in an inverted index database (e.g. MS Access). The similarity calculator is a separated module implemented in C# as well and connected to the index database. Figure 4 shows the browsing interface of SGI to illustrate how to search distant concept semantically.
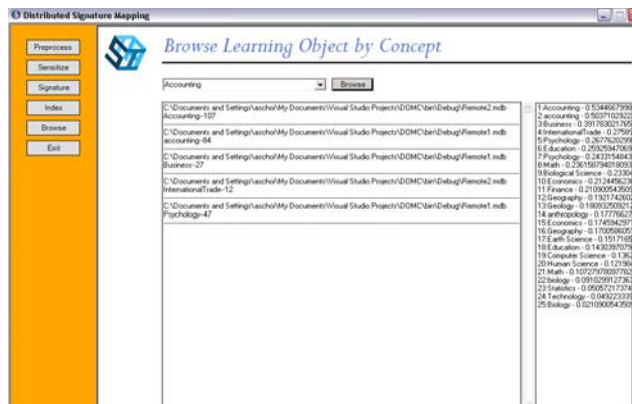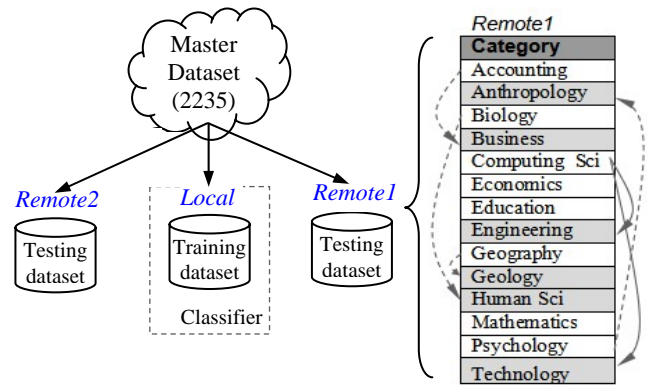
Figure 4. Browsing interface of SGI

Figure 5. Dataset distributions into training and testing data

EVALUATION

In order to test the hypothesis of using semantic signatures to enable distributed semantic browsing and to improve relevance we have simulated the distributed concept retrieval and compared the results with the traditional keyword-based and label-matching method. To replicate the distributed repositories in a collaborative E-learning network, the three independent databases are set up. As shown in Figure 5, they are called "local", "remote1" and "remote2" where the local, of course, denotes a local data source and both remote1 and remote2 simulate distant data sources. A single master set of metadata in 8 different categories is distributed evenly in number and randomly into the three simulated repositories.

The metadata have been transformed to conform to the IEEE LOM format. After the distribution, the local database contains the metadata that represents the set of the training data for the classifier. During the training phase, the kNN classifier uses the instance of the local metadata to learn the features to identify the class of the metadata. It starts by extracting keyword terms from each category of metadata and projecting them into the vector space model. Next, after running through the signature generation module, each category of metadata is represented and indexed by a semantic signature in the database.

The dataset in both remote1 and remote2 is controlled to model the situation of potentially different ontological classification in a distributed environment. To simulate the effect of varied concept labelling, the original 8 categories of metadata are expanded to 14 categories in remote1. The 6 derived categories are labelled with different class names from their respective sources and described with the metadata taken out from source categories. Each newly derived category contains metadata belonging to the same class. To illustrate, a part of the metadata from the category "computing science" is distributed to the derived categories "technology" and "engineering" in remote1. Thereby, the metadata for concept "computing science" is now grouped

15

into "computing science", "technology" and "engineering". Essentially, this simulates the situation when a concept "computing science" could be categorized differently into concepts like "technology" and "engineering" in different ontology. The same distribution principle is applied to remote2 database which includes 13 categories of which 7 are derived categories.

Similar to the local database, each category of the metadata in remote1 and remote2 is mapped to a semantic signature in WordNet senses and stored in the local database as an index. To test semantic-based search, semantic signature representing a local concept is sent to query the remote repositories. The semantic similarity is compared between the query signature and the distant signature based on the similarity function. Finally, the result of the k most similar concept signatures from the remote databases are studied based on the relevance metric.

Dataset

Since there is no publicly available dataset of learning resources metadata, the experiment metadata were 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. As mentioned, 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, to train and test the group. After distribution, the local database contains 667 training records while remote1 and remote2 contain 1568 testing records.

Results

In order to gauge the effectiveness of the proposed mediation method between different E-learning ontologies, three standard metrics for information retrieval are used in the evaluation of the system performance: they are Recall, Precision and F-measure. Table 2 shows that the use of semantic signature can consistently improve retrieval relevance in terms of recall and precision. In all categories, the semantic based retrieval out perform both keywords-

Table 2. Comparison on precision, recall and F-measure on concept retrieval

Category	Precision			Recall			F-Measure		
	S	K	L	S	K	L	S	K	L
Acc	1	0.6	0.5	1	0.75	0.5	1	0.6	0.5
Bio	0.6	0.6	0.5	0.75	0.6	0.5	0.6	0.6	0.5
CS	1	0.5	0.3	1	0.5	0.3	1	0.5	0.3
Econ	1	1	0.6	1	0.75	0.6	1	0.6	0.6
Educ	0.6	0.5	0.5	0.75	0.75	0.5	0.6	0.45	0.5
Geo	0.6	0.5	0.5	0.75	0.5	0.5	0.6	0.5	0.5
Math	1	0.3	0.6	0.6	0.5	0.6	0.7	0.36	0.6
Psy	1	0.3	0.3	0.6	0.6	0.3	0.7	0.4	0.3

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

Table 1. Source and Category of Metadata

Category	Source	No. of records
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

based retrieval and label-matching retrieval.

As oppose to the classic or traditional keywords-based representation, semantic-based indexing with WordNet senses can include more lexicon information than simple syntactic approach. This implies that more features will be added to the class signature representation. Since more features are added, that may also mean that more noise is included as well.

Intuitively, the increased relevance of retrieval can be attributed to the expansion of features in class representation. However, different from what we expected, the precision does not decreased. It is suspected that due to the relatively small size of the dataset and 1-k hypernym generalization, the senses included in the signature are 'good' in terms of classification. Therefore, combined with a good contextual-based sense selection strategy, WordNet as a mediatory can provide source for ambiguity resolution and semantic information for the process of semantic browsing. Coupled with that, the selection of kNN algorithm as the classifier also contributes to the performance of the system.

kNN is an instance-based classifier. The performance of instance-based classifiers is more dependent on the sufficiency of the training set rather than other machine learning classification algorithms. Thus, it is a disadvantage for kNN to have a small dataset for training and testing. A smaller training set implies more terms or term combinations important for content identification may be missing from the training sample documents. This negatively affects the performance of a classifier. Nevertheless, the ontology (e.g. WordNet) guided approach seems to somewhat reduce the negative influence of this problem. The replacement of child concepts with parent concept through hypernym relationship appears to be able to discover an optimum concept set without adversely affecting performance. Therefore, an important term, which resides low in the concept hierarchy may be mapped to a parent concept and included in the signature for class comparison, even if this term is not included in the training set.

DISCUSSION

The improvement on concept retrieval by using semantic signature is not uniform across different categories. For example, the improvement on retrieval of *"Psychology"* and *"Accounting"* metadata is more than improvement on *"Biology"* and *"Geology"*. We believe that for some classes of metadata like *"Biology"*, which are characterised by a set of specific keywords, the use of semantic signatures does not add extra useful information into the representation model to help in classifying metadata. On the other hand, using 1-k hypernym generalization on such a highly specialized domain may in fact introduce more noise to reduce the matching possibility in similarity calculations. In addition, with a small size of dataset, over-fitting on classification model may also result. Therefore, further experimentation and analysis are needed to fully understand the impact of WordNet signature with sense generalization in classification of metadata.

CONCLUSION

This project offers two important contributions. First, it gives a new light-weighted semantic (ontology) mapping approach to enable cross platform concept browsing in a federated network. Unlike many current practices in semantic mapping that either require 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 for the experimental results that establish the merit of using 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.

By using immediate parent sense generalization during sense selection process, it does not only successfully reduce the dimension in semantic signature, but more importantlly introduces flexibility in the sense selection and increases the opportunity to find a better sense without compromising the relevance in the search result. This creates incentive to explore the use of other sense selection strategy.

REFERENCES

[1] Robin Dhamankar, Yoonkyong Lee, AnHai Doan, Alon Halevy, Pedro Domingos, *"iMap: Discovering Complex Semantic Matches between Database Schemas"*, Proceedings of the ACM SIGMOD Conference on Management of Data. (2004)

[2] George A. Miller, *Wordnet: An Online Lexical Database*, International Journal of Lexicography (1993).

[3] Asuncion Gomez-Perez, *Ontological Engineering with Examples from the areas of Knowledge Mangement, e-Commerce and the Semantic Web*, Springer-Verlag London (2004).

[4] H. P. Edmundson, *New Methods in Automatic Extracting*, Journal of the ACM (1969).

[5] Ching Kang Cheng, Xiaoshan Pan and Franz Kurfess, *"Ontology-based Semantic Classification of Unstructured Documents"*.

[6] Khaled M. Hammouda and Mohamed S. Kamel, *"Phrase-based Document Similarity Based on an Index Graph Model"*, Proceedings of IEEE International Conference on Data Mining (2002).

[7] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy *"Learning to map between ontologies on the semantic web"*, Proceedings of WWW2002 conference (2002).

[8] Ching Kang Cheng, Xiaoshan Pan and Franz Kurfess, *"Ontology-based Semantic Classification of Unstructured Documents"*, Proceedings of 1st International Workshop on Adaptive Multimedia Retrieval (2003).

[9] Martin Ester, Hans-Peter Kriegel and Matthias Schubert, *"Web Site Mining : A new way to spot Competitors, Customers and Suppliers in the World Wide Web"*, Proceedings of 4th International Conference on Knowledge Discovery and Data Mining (2002).

[10] Yannis Kalfoglou and Marco Schorlemmer, *Ontology mapping: the state of the art*, The Knowledge Engineering Review (2003).

[11] Mineau, G.W, *"A simple KNN algorithm for text categorization"*, Proceedings of IEEE International Conference on Data Mining (2001).

[12] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).

[13] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, *Semantic matching*, In 1st European semantic web symposium (ESWS'04) (2004).

[14] A. Maedche, B. Motik, N. Silva and R. Volz, "MAFRA - A MApping FRAmework for Distributed Ontologies", in EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, pp. 235-250, 2002.