TOWARDS BROWSING DISTANT METADATA WITH SEMANTIC SIGNATURES

by

Andrew Shek-Ting Choi
B.Sc., Simon Fraser University, 2002
B.Com., University of Alberta, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the School
of
Interactive Arts and Technology

SIMON FRASER UNIVERSITY

Summer 2005

APPROVAL

Name: Andrew Shek-Ting Choi

Degree: Master of Science

Title of Thesis: Towards Browsing Distant Metadata with Semantic Signatures

Examining Committee:

Chair: Dr. Rob Woodbury

Dr. Marek Hatala
Senior Supervisor

Dr. Vive Kumar
Supervisor

Dr. Tom Calvert
External Examiner

Date Defended/Approved: August 3rd, 2005

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

ABSTRACT

One of the benefits of an E-Learning network is to connect users to distributed learning repositories where they can be exposed to numerous learning resources. However, metadata of learning resources stored in different repositories are often annotated with concepts defined by different ontologies or classifications 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 similar learning resources. As such, using a keyword-based approach to find relevant information may not yield satisfactory results.

In this thesis, I describe a lightweight information integration solution for browsing federally distributed metadata without incurring expensive schema matching or semantic mapping. I present experiments on real-world data that validate the proposed solution. Finally, I discuss how this approach can simplify semantic mapping and enhance browsing experience in a distributed repository network.

To my dearest parents
for their utter love and infinite sacrifices. I am forever in their debt.

To my beloved wife April
for her unconditional love and support

ACKNOWLEDGEMENTS

I would like to thank my supervisory committee and my examiners. Your input to this work has greatly improved its quality.

I would like to thank Dr. Marek Hatala, my senior supervisor, for his support and advice during my research.

This thesis has benefited from many conversations with a number of people over the last year. I would like to thank everyone else in the Laboratory for Ontological Research at Simon Fraser University for all the stimulating discussion.

TABLE OF CONTENTS

Approval ii
Abstract iii
Dedication iv
Acknowledgements v
Table of Contents vi
List of Figures viii
List of Tables ix
Chapter 1: Introduction 1
1.1 Thesis Statement 2
1.2 Thesis Organization 3
Chapter 2: Related Works 5
2.1 Information Retrieval 5
2.1.1 Text Retrieval 6
2.1.2 Vector Space Model 7
2.2 Information Integration 9
2.2.1 Federated Collaborative Network 10
2.3 Ontology 13
2.3.1 Why are ontologies important? 15
2.4 WordNet 16
2.5 E-learning 18
2.5.1 What is metadata? 20
2.5.2 Is metadata always *better*? 22
Chapter 3: Challenges in Semantic Mapping 24
3.1 Research problem 24
3.1.1 Semantic mapping scenarios 24
3.2 Semantic mapping 26
3.3 Challenges in Semantic mapping 26
3.4 Review of current approaches 30
Chapter 4: Browsing with Semantic Signature 35
4.1 Semantic Signature Definition 36
4.2 Why use Semantic Signature? 37
4.3 Building Semantic Signatures with WordNet senses 40
4.3.1 General Methodology 40

4.3.2 Signature Generation in Action 42
4.4 Federated Concept Browsing in a Repository Network 50
4.4.1 Browsing distant metadata with semantic signature 51
Chapter 5: Experiment and Evaluation 54
5.1 Evaluation settings 54
5.2 Assumptions 57
5.3 Dataset Description 58
5.4 Metric 59
5.5 Limitations 60
5.6 Results 60
5.7 Interpretation 63
Chapter 6: Conclusion and Future Directions 66
6.1 Conclusion 67
6.2 Future Directions 68
Appendix A 70

LIST OF FIGURES

Figure 2.1 Document term matrix 7
Figure 2.2 Federated collaborative network 12
Figure 2.3 An example of simplified domain ontology 14
Figure 2.4 WordNet senses for word '*Web*' 17
Figure 2.5 Simplified portion of WordNet on term of '*turtle*' 18
Figure 4.1 Semantic vs. syntactic matching in different ontologies 38
Figure 4.2 Semantic Signature Generation Framework 42
Figure 4.3 Word term to WordNet sense mapping 44
Figure 4.4 Associative frequency calculation between word senses 47
Figure 4.5 Word sense generalization to immediate (1-k) parent 48
Figure 4.6 Aggregation of document signature to generate class signature 50
Figure 4.7 Inverted index by Semantic Signature 51
Figure 4.8 Integrated process of semantic-based browsing of metadata 52
Figure 5.1 Metadata distribution in simulated distributed data sources 56

LIST OF TABLES

Table 5.1 Source and Category of Metadata 58
Table 5.2 Comparison on precision, recall and F-measure on concept retrieval 62
Table 5.3 Comparison of similarity score using *1-k* parent generalization on remote1 62

CHAPTER 1: INTRODUCTION

With the advance of the Internet and rapid developments in E-learning, more and more institutions are joining to form distributed learning networks to provide their users with access to resources from different learning repositories. This creates pressure for institutions to provide an efficient way to organize the huge volume of learning materials located in different repositories. The classification has to be flexible and robust enough to deal with variation in conceptual frameworks of dispersed audiences 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. Metadata are descriptive indexing labels used to describe the characteristics and content of learning objects. They are used to facilitate searching, management and assembling of learning content. Ontology, on the other hand, is to define the set of vocabularies to describe the metadata for each particular concept. However, in a federated environment, keywords-based search on metadata elements could not guarantee the discovery of all relevant information. This is because linguistic variation in metadata makes direct querying with keywords sometimes ineffective even with the ontology to control the vocabularies used to describe the metadata. Obviously, it is almost impossible to expect that two institutions would use exactly the same keywords and classifications to describe the same learning resource.

Therefore, it is very unlikely that using a keyword-based retrieval system could return all relevant documents, or more precisely, semantically relevant documents in a federated learning network. In addition, descriptions used in metadata to define and classify learning resources may not be in the same standardized format across a learning network [1][2]. Hence, finding related information on a topic from heterogeneous sources is very challenging. This is a long-standing problem of information integration [3, 4]. Currently, active research is underway to provide efficient and effective solutions for a global view of information from distributed sources. Information integration aims to transform heterogeneous data sources into a single global homogeneous database and to provide a unified view of these data for future query processing purposes [5].

1.1 Thesis Statement

The research question of this thesis is to explore the use of a lightweight semantic mapping strategy to browse in a federated network of repositories for semantically relevant metadata with the use of WordNet. This research integrates a number of techniques in information retrieval, information integration and data mining to achieve semantic mappings among different metadata repositories in a cooperative network environment, in order to allow users to browse by semantically similar concepts.

The central thesis of this research is to explore the use of semantic signatures in WordNet terms to enhance relevance of federated browsing in a collaborative repository network. This ultimate goal encompasses a number of objectives. First, we would like to establish experimentally the benefit of using

WordNet as a mediated schema to construct semantic signatures for semantic mapping. Secondly, we would also like to investigate the use of linguistic heuristics to select the appropriate senses to construct a "good" semantic signature to represent a concept in the classification schema. Finally, we would like to demonstrate the relative merit contributed by semantic signature mapping by comparing the result of federated browsing using semantic signatures against another widely used keyword-based browsing.

1.2 Thesis Organization

The rest of this thesis is structured as follows:

Chapter 2 *Related Works*

This chapter summarizes the major concepts in Information Retrieval, Information Integration, Ontologies, WordNet, and E-learning metadata. It builds an intellectual foundation for subsequent chapters, and identifies works that this research relates to.

Chapter 3 *Challenges of Semantic Mapping*

This chapter illustrates the research problem, as a motivation, in semantic mapping with examples. It presents various research and technical challenges that we are facing in semantic mappings. It includes a brief review of several popular approaches in semantic mapping.

Chapter 4 *Browsing with Semantic Signature*

This chapter introduces the concept of a semantic signature to represent a concept in the classification schema or ontology and describes how it is

constructed with WordNet terms. It also further illustrates the idea of using semantic signatures to facilitate semantic mapping to enable concept browsing in a distributed collaborative learning network. A semantic signature indexing tool will also be displayed as a realization of the signature browsing approach.

Chapter 5 *Evaluation and Interpretation*

The utility of the semantic mapping using WordNet signature is evaluated in a specific verification domain, namely, that of the E-learning metadata browsing by concept. A set of assumptions, dataset description, experimental results and their evaluations will be discussed in detail.

Chapter 6 *Conclusion*

This chapter summarizes the major contributions of this work and provides future directions of this research.

CHAPTER 2: RELATED WORKS

This chapter will highlight the main concepts in Information Retrieval, Information Integration, Ontologies, WordNet, and E-learning. They provide important background information to understand the content and approaches adopted in this thesis.

2.1 Information Retrieval

Information retrieval is usually used as a generic term to cover the study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms. Specifically, it deals with the representation, storage, organization of, and access to information items [6]. Unlike data retrieval, information retrieval is not a database querying. Databases work with highly structured information. The data model of a specific database determines the possible queries a user can ask. Usually the form of the query will have to follow the data model. A database is used where exact matching is demanded. On the other hand, information retrieval models use highly ambiguous queries and include some amount of fuzziness as the user defines the search query.

The key notion of information retrieval is that relevance is defined in terms of similarity. This assumes that if a document is similar to a query it is relevant. Similarity in turn can be defined in several ways, depending on the type of the information. For text retrieval, similarity is usually measured in the overlap of the

words used in both query and document [7]. A document could be anything from a title, to an abstract, to a full-text paper. Information retrieval is a "broad interdisciplinary field of research that draws on many other disciplines such as cognitive psychology, linguistics, information science and computer science"[1]. In this thesis, information retrieval is limited to text retrieval as the experiment is carried on with text-based learning resource metadata.

2.1.1 Text Retrieval

Text retrieval can be separated into two distinct phases: indexing and matching [6]. The indexing phase is concerned with extracting keywords and estimating their relevance to the document in which they occur, and finally indexing the document with a set of representative keywords. The matching phase calculates the similarity of query terms against the index terms. A document is relevant to a query if they are similar. Many approaches determine the similarity of a query to a document based on the words that are used. A common assumption in the information retrieval community is that documents can be treated as a "bag-of-words". In this view, a document is treated as an unordered set of words. As such, determining whether a document is relevant to a given query is simply reduced to looking up the query words in the document index. The more query words that are overlapped with index words, the more relevant the document.

[1] Definition of information retrieval from Wikipedia
http://en.wikipedia.org/wiki/Information_retrieval

2.1.2 Vector Space Model

The vector space model is a popular data model to represent a document for computational manipulation. It enables logical operations to be performed on the documents. It is a commonly used technique in text retrieval [7, 8]. In this model, each document is broken down into a word frequency table representing n-dimensional weighted vectors. Each word is a dimension in Euclidean space. Let $T = \{t^1, t^2, \ldots, t^n\}$ denote the set of terms in the collection of documents. Then we can represent the terms d_j^T in document d_j as a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ with:

$$x_i = \begin{cases} t_j^i \text{ if } t^i \in d_j^T; \\ 0 \text{ if } t^i \notin d_j^T \end{cases}$$

where t_j^i represents the frequency of term t^i in document d_j. Combining all document vectors creates a document-term matrix. An example of such a matrix is shown in Figure 2.1.

Figure 2.1 Document term matrix

	d_1	d_2	d_3	.....	d_n
t_1	1	0	0	.....	3
t_2	2	0	0	.....	0
:				.....	
t_n	0	1	1	.....	0

Each dimension in the document vector corresponds to the term frequency of a key term. In addition, weight can be added to each term based on its importance to distinguish the document category. There are a number of term-weighting schemes; however, the most widely used one is called Term Frequency-Inverse

Document Frequency (TFIDF) [9]. This scheme assigns a weight to each term in a given document. The weight increases in proportion to the number of times the term occurs in the document, but is offset by a term itself, which devalues terms common in the overall corpus. Mathematically, it can be expressed as follows:

$$W_{ij} = tf_{ij} * log(\frac{N}{df_i})$$

where: W_{ij} = weight of term t_i in document d_j

tf_{ij} = frequency of term t_i in document d_j

N = number of documents in corpus

df_i = number of documents containing t_i

The TFIDF weighting scheme is a broadly recognized method to select most representative keywords to represent a category of documents in document classification [10]. The advantages of the vector space model include ranked results of the retrieved documents, the possibility to enter free text, and not requiring a strict matching of the documents. The ranked results are ordered using the distance of a document vector to the query vector. Such an ordering represents the similarity of a document to the query. Free text search eliminates the use of difficult query languages as in the Boolean model. The matching is not strict, in the sense that a query containing multiple words will also retrieve documents where not all words are present [11].

In order to retrieve document relevant to a user query, we calculate the similarity between the query vector and document vector based on a distance function. A common similarity measure, known as the cosine measure,

determines the degree of closeness between the document vectors and the query vector is frequently used [7]. Precisely, the similarity between a document vector d and a query vector q is defined as:

$$sim(d,q) = \frac{\sum_{i=1}^{t} d_i q_i}{\sqrt{\sum_{i=1}^{t} d_i^2 \times \sum_{i=1}^{t} q_i^2}}$$

where $\sum_{i=1}^{t} d_i q_i$ is the standard vector dot product between document vector and query vector while $\sqrt{\sum_{i=1}^{t} d_i^2 \times \sum_{i=1}^{t} q_i^2}$ is called the normalization factor to discard the effect of document length on the overall similarity score.

2.2 Information Integration

While early databases were usually self-contained, it is now generally recognized that there is a great value in taking information from various geographically separated sources and making them work together as a whole [12]. This is particularly true in situations that call for high-level collaboration to share information such as in a research network and an E-learning network. We acknowledge that a vast amount of information is stored in distributed data sources. The physical distance is usually not a major problem but rather the difference in their logical representation (semantics). Indeed, a global view of cohesive information is not easy to obtain when information is not only stored in different databases operated with different DBMS, but also they usually come in different structures, represented in different data models and expressed with linguistic variations. These differences pose problems for distributed searching

using free text because of the variations in query keywords. Therefore, an effective information integration strategy is important to glue distributed information to form complete and coherent information that is consistent with a local users' semantics.

Information integration has long been identified as a central problem of distributed multi-database systems, which are required to provide interoperability among an array of information repositories [4]. Specifically, information integration refers "to the problem of merging, coalescing and transforming autonomous heterogeneous data sources into a single global homogeneous database and providing a unified view of these data for future query processing purposes" [2]. In order to perform semantic integration of heterogeneous information, it is necessary to form one or more integrated schemas expressed in some common data model. Detailed discussion on semantic mapping of heterogeneous sources can be found in Section 3.

2.2.1 Federated Collaborative Network

Distributed storage systems come in different flavours. Based on specific requirements and applications within organizations, these could be a tightly-coupled distributed database systems controlled by a centralized DBMS or it can be a loosely-coupled federated system in which each component database has high autonomy controlling its degree of participation in the federation [13]. In summary, two important aspects of federated systems can be noted:

[2] Source: http://www.cs.ubc.ca/~ycai/Academics/projects_files/infointeg.html

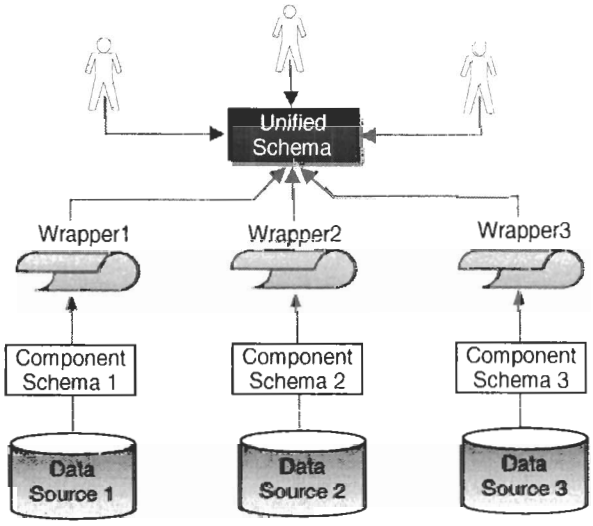
- *Heterogeneity*: Federated systems could have a high degree of differentiation in their various data sources. Each component data source may run on different hardware, use a different communication network, and have a different DBMS to manage their data repositories. They may also have different query languages, different query capabilities, and even different data models [14]. Apart from these structural heterogeneity differences, semantic heterogeneity may also occur in each component data source in which the intended use of same or related words would be different, or different words in fact carry the same semantic interpretation. This creates various technical difficulties when integrating information from heterogeneous data sources.

- *Autonomy*: Typically, a data source has existing organizational requirements to fulfil and target users to serve. It is important, therefore, that the operation of the source is not affected and it remains independent when it is brought into a federation [5, 14]. In particular, the way the data source processes requests for data should not be affected by the execution of global queries against the federated system. In addition, all the design and execution decisions will remain with local authority. A component data source can participate in more than one federated system.

In this thesis, we focus on a variant of a federated system called a federated collaborative network. It can be described as a federated network

where each participating data repository dedicates minimum resources to provide an extra layer of consolidation to ensure the semantic consistency and quality of delivered information. This is a working model adopted in several emerging areas ranging from the knowledge management communities, bioinformatic research networks to E-learning networks [15]. Under this model, a number of participating organizations would join to form a community or even a cluster of communities in order to share their resources and information with a goal to minimize development effort to provide richer content to users in each community. Although this thesis primarily focuses on the discussion of the E-learning repository network, we believe that the validity of the methodology described in this thesis can be easily extended to other collaborative networks with minimum modification. The detailed model of E-learning repository network will be discussed later in this section. Despite some differences between federated collaborative networks, they all share similar design as below.

Figure 2.2 Federated collaborative network



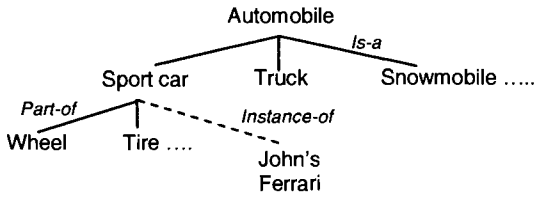
In this simplified view, a federated collaborative network is a kind of loosely coupled federated system that explicitly requires the use of wrappers to provide an intermediary between different component data sources. In the network, participating data sources have total control over the data they manage. They have access control rules to allow partial and controlled sharing of their data. Yet there is no centralized control in a federated system, it usually provides users with a single global interface to access data sources in the federation. The whole federated system is transparent to users who are treated as local users who can indirectly access a distant database.

2.3 Ontology

In theory, ontology can be defined as a specification of a conceptualization [16]. More formally, ontology is a formal representation of a set of concepts, properties of concepts, and relations between concepts that are possible in a specified domain of knowledge [17]. Practically, ontology provides a vocabulary whose terms are precisely defined by a body of knowledge or facts in some domain. However, it is important to remember that "it is not the vocabulary as such that qualifies as an ontology, but the conceptualization that the terms in the vocabulary are intended to capture" [17]. Further, ontology also defines semantic relationships like "*type-of*" and "*is-a*" between terms using formal modelling techniques, in general taken from logic-based specification formalisms such as description logics or predicate calculus. Ontology starts from precisely defined simple concepts that are universally accepted such as "Thing" or "Entity" and then leads to concepts with narrower scope and more specifications. Concepts in

ontology are interconnected by means of a set of semantic relationships. A general structure of a portion of an ontology shown in figure 2.3 defines that "*Automobile*" that subsumes "*Snowmobile*" because it logically implies that snowmobile is a kind of automobile. Three kinds of semantic relationships are commonly used to create ontology: they are specialization (*is-a*), instantiation (*instance-of*), and component membership (*part-of*).

Figure 2.3 An example of simplified domain ontology



The *is-a* relation is used to represent specialization. A concept represented by C_j is said to be a specialization of the concept represented by C_i if C_j is a kind of C_i. The *instance-of* relation denotes concept instantiation. If an instance I_j is a type of concept C_i, the interrelationship between them corresponds to an *instance-of* denoted by a dotted line. For example, "*John's Ferrari*" is an instance of "*Sport car*". A concept represented by C_j is *part-of* a concept represented by C_i if C_i has a C_j (as a part) or C_j is a part of C_i. For example, "*Wheel*" is part of concept "*Sport car*". These semantic relationships permit the construction of ontologies with richer structure than plain hierarchy commonly found in taxonomies. The ontologies enable programs to deduce knowledge by combining different concepts and examining their semantic relationships. Ontology can be

constructed in two ways, domain dependent and generic. The former provides a small number of fine grain concepts while the latter provides a large number of coarse concepts. WordNet is an example of a generic ontology that will be discussed shortly.

2.3.1 Why are ontologies important?

Given their solid foundation built from logical formalism, ontologies find their roles in many areas of artificial intelligence applications, information systems, knowledge engineering and computational linguistics. First, ontology provides us with a set of logical axioms to account for the intended semantics of a vocabulary used to describe facts, beliefs, hypotheses, and phenomena about the world or in a specific domain [18]. The set of axioms is usually stated in the form of first-order logic where vocabulary terms are the predicates while the object and relations are the variables. As an example, if $G(x, y)$ is the predicate representing "x *greater than* y", then the sentence "9 is greater than 6" can be expressed as $G(9,6)$. Recall from figure 2.3 that the ontology describes a hierarchy of concepts related by "*Type of*" subsumption relationships; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and to constrain their intended interpretation [19]. Formal axioms can clarify conceptual confusion by limiting the intended meaning of a vocabulary, and the linked relations between concepts in a domain. As such, factual knowledge in a relevant domain can be represented in logical symbols and be understandable by computers. Computers can then make logical inferences by operating on the existing facts and axioms. Moreover, with a

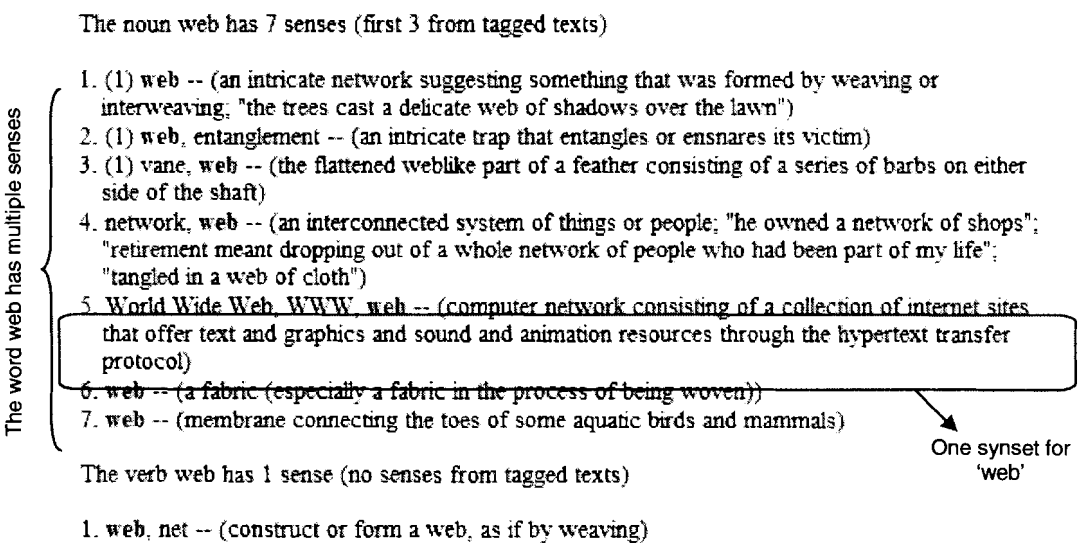
domain ontology that specifies the intended semantics of concepts, the ontology enables knowledge sharing and reuse with others who share similar needs for knowledge representation in that domain. By extension, ontology also facilitates the information integration process from heterogeneous sources in a particular domain in which specific concepts are defined by well-formed logical syntax and by their semantic category. In summary, the merit of ontologies can be attributed to their capability to provide an explicit specification of shared conceptualization of knowledge in a world that we wish to represent for some purpose, and to facilitate communication between people, organizations, or between information systems.

2.4 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" [17, 20, 21]. 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*) [21]. Each synset has a unique identifier (ID) and a specific definition. A synset may consist of only a single element (sense term), 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 a computer network consisting of a collection of internet sites. In this context, 'World Wide Web', 'WWW' and 'Web' are considered carrying the same meaning in English. For cases where a single word has multiple meanings (polysemy), multiple separate and potentially unrelated synsets will contain the same word. Considering that the word 'Web' can have 7 multiple meanings as a computer network, entanglement, spider web and etc, they all contain the word "web". In Figure 2.4, we would see the word 'Web' will appear in different synset lists offered by WordNet to characterize the concept 'Web'. From an ontological perspective, WordNet is not only a lexical dictionary but also a generic linguistic ontology that carries both semantic and syntactic information of words as well as organizing them in a hierarchical taxonomic structure linked by semantic

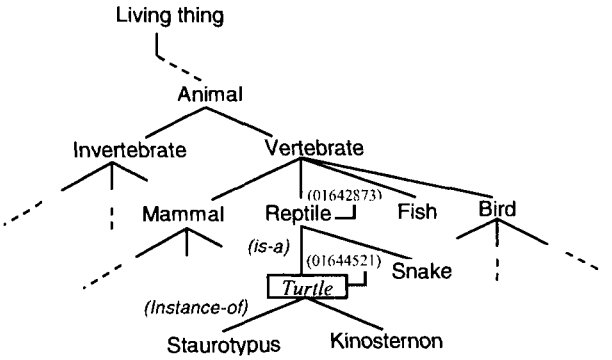
Figure 2.4 WordNet senses for word '*Web*'



relationships [20]. In WordNet, noun synsets are related to each other through hypernymy (generalization), hyponymy (specialization), holonymy (whole of) and

meronymy (part of) relations. As mentioned previously, these semantic relations are used to link concepts together in an ontology for a domain. As shown in Figure 2.5, the word 'turtle' is semantically related to other words in the biology domain in WordNet. In summary, WordNet has evolved from an original idea to

Figure 2.5 Simplified portion of WordNet on term of '*turtle*'



create an online dictionary and thesaurus that is machine readable and reasonable, to become a major language and concept lexical database that provides support for cross disciplinary research from natural language processing tasks, such as information retrieval, information extraction, word sense disambiguation to text mining, document summarization and others.

2.5 E-learning

E-learning has been a topic of increasing interest in recent years. Promoted by many E-learning experts, information technologists and even the governments, it is seen as a plausible solution to meet the educational challenges in today's growing demand for high quality education by providing learners with learning resources anywhere and anytime. Originated from

computer-based training (CBT), E-learning attempts to provide a dynamic and non-linear learning environment where learners can control their pace of learning and organize their learning processes to adapt to their own needs in the subject area. E-learning is often perceived as a group effort, where content authors, instructional designers, multimedia technicians, teachers, trainers, database administrators, and people from various other areas of expertise come together to serve a community of learners[3]. Compared to traditional learning in which the instructor plays the intermediate role between the learner and the learning material, the learning scenario in E-learning is completely different: instructors no longer completely control the delivery of material and learners have a possibility to combine learning material from various sources on their own. Thus, the content of the learning materials is independent from the course curriculum. However, despite the time or expense put into creating advanced E-learning materials, they are useless, unless they can be searched and discovered easily. This is especially true as the volume and types of learning content increase. If the learning resources delivered do not meet learner's expectations and requirements, this could lead to frustration in learners and reduce the number of E-learning users. Without a broad base of learners, it is difficult to justify the large investment in E-learning technology. On the other hand, lacking an interoperable platform to enable institutions to share learning resources, even in a collaborative network, each member must invest a great amount of cost and effort to develop useful learning resources. This could be a problem for many institutions as they struggle to allocate economic resources between traditional learning and E-

[3] http://ifets.ieee.org/discussions/discuss_march2003.html

learning development. In order to achieve the vision of E-learning society, we need to create a cost effective E-learning environment. In other words, learning resources must be easily searchable by learners based on their learning requirements, and discovered by instructors for re-development and adaptation for their changing needs. In this way we can lower the cost as well as reduce the time of learning resource development, and more importantly be able to deliver learning resources to target users in a most efficient way. This is the reason to prompt the growth of collaborative learning environments. Collaborative learning is a full-featured and, flexible cooperative learning system (network) established to enable teachers to reuse the learning resources, deliver and manage e-learning resources easily and efficiently across different locations in the network. Very often, a collaborative learning network is formed by several learning institutions that share similar learners and teaching goals. To facilitate the communications between institutions and to support the search for learning resources while maintaining autonomy in developing their own learning resources and curriculum, metadata is frequently used to describe learning resources for the purpose of easy searching and indexing.

2.5.1 What is metadata?

Metadata is data about data or information about information. It is structured data that describes the characteristics of a resource. A typical metadata record is an XML file consisting of a number of pre-defined elements, information like keywords, category, title, description, author, location, page-length, ISBN, and so on, representing specific attributes of a resource. To avoid

confusion and enable exchange, we use a metadata schema to define the syntax and semantics of each metadata element. Furthermore, metadata are often annotated by ontology or taxonomy. With ontology, the classifications of metadata are controlled by concepts or classes. With taxonomy, the classifications of metadata are organized by categories or nomenclatures. In the E-learning domain, metadata is the information-age term for information (e.g. index card) that librarians traditionally have used to classify learning resources and other print documents. The main purpose of the use of metadata in E-learning is for managing and accessing learning resources in a systematic way. In addition, the use of metadata promotes sharability of learning resources in a distributed environment because it provides a common nomenclature enabling learning resources to be described in a common way[4]. It also serves as a good foundation to aggregate learning resources from distributed sources. Until recently, three popular metadata standards were used to describe E-learning resources: they are IEEE LOM, ARIADNE and IMS. However, IEEE LOM has became the metadata standard in the E-learning community. One of the major purposes of all meta-models is to define how learning materials can be described in an interoperable way [1]. All the metadata elements necessary to describe a resource can be classified into several categories, each offering a distinct view on a resource. For example, the IEEE LOM standard specifies the following metadata categories and elements:

- general - groups all context-independent features;
- lifecycle - groups features related to the lifecycle of the resource;

[4] Learning Technology Standards Observatory: http://www.cen-ltso.net/Users/main.aspx?put=322

- meta-metadata - groups the data elements describing the metadata;
- technical - groups data elements describing the technical features;
- educational - groups educational and pedagogic data elements;
- rights - groups data elements pertaining to the conditions of use;
- relation - groups data elements that describe the linkage between resources;
- annotation - groups data elements that allow comments on the educational use;
- classification - groups data elements that describe the position of the resource in an existing classification system.

2.5.2 Is metadata always *better*?

The purpose of metadata is to describe resources in a standardized structural format for managing and searching resources systematically. In reality, with different metadata standards, different elements may be used or the same element would carry different meaning in the metadata. Coupled with that, most standards lack a formal semantics to explicitly control the meaning of the elements. This obviously will create problems of incompatibility between disparate and heterogeneous metadata descriptions or schemas across domains [1]. To illustrate, two different authors may describe semantically identical concepts in different terms according to their different points of view. Author X may use terms like *network*, *protocol* or *web* to describe the concept "Internet" while author Y would describe the same concept using other terms like *link*, *http* or *WWW*. Many organizations solve this problem by developing ontology to specify the meaning of vocabularies used in the metadata, and limit the mappings from terms of the domain vocabularies to concepts in the ontology. However, this only solves the problem of semantic discrepancy in metadata description in a single institution. This is because when metadata are widely distributed across a collaborative network in which each learning resources

repository has total autonomy to use its own ontology to annotate metadata, it is very difficult to coerce other members to adopt each other's ontology to annotate their metadata. Therefore, at the end interoperability requires repositories to develop various ontology mapping strategies in order to have a shared-understanding about the meaning of each other's metadata and in turn to allow them to find semantically related learning resources. This not only makes the search process very inefficient but also requires important procedures to be performed to create semantic mappings between ontologies before distributed metadata can be shared between institutions. Therefore, using metadata does not completely solve the interoperability problem without a true semantic understanding of those metadata in a distributed environment.

CHAPTER 3: CHALLENGES IN SEMANTIC MAPPING

This chapter illustrates the research problem, as a motivation, with examples in situations where semantic mapping is demanded. It presents a brief overview of semantic mapping techniques and reviews the research challenges that many semantic researchers are facing. A set of current approaches in semantic mapping will be discussed.

3.1 Research problem

Regardless of the progress of information and communication technologies, the challenges remain on how to integrate information from heterogeneous data sources [22]. Semantic mapping is one of the most active research topics in information integration to help provide an interoperability policy for people to share information in a distributed environment [23, 24]. In fact, "semantics-based technologies will be an essential part of all interoperability solutions in the very near future" [25]. The scenarios that follow will reveal situations where information communication will not be possible without robust semantic mapping between distributed data sources.

3.1.1 Semantic mapping scenarios

Scenario 1:

Imagine a 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. However, the returned results from R_2 and R_3 are mixed with many irrelevant resources related to catching a bass (fish). This problem occurs frequently when the concepts are defined by different domain ontologies with different sets of vocabularies carrying different intended meanings.

Scenario 2:

Imagine that the same learner L_1 sends out a distributed request for learning resources on topic "*advanced database*"; however, since the same topic is in remote repositories annotated by the concept "*database system II*" in remote repositories so it is labelled differently. Therefore, in a concept-based search, learning resources defined by concept "*database system II*" will not be returned for request of "*advanced database*" 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, it is still difficult to find learning resources based on local conceptual definitions. This problem cannot be solved even with the fact that institutions have their own ontologies to define vocabularies used to describe metadata of learning resources. This research focuses on providing a light-weighted semantic mapping between distant metadata from heterogeneous repositories in a collaborative E-learning network, and enabling learners to browse for semantically related learning resources by topics.

3.2 Semantic mapping

Semantic mapping can be described as the mapping task to identify common concepts and to establish semantic relationships between heterogeneous data models in the same domain of discourse [26]. Since in modern knowledge systems semantics is mostly represented in ontological constructs, we will use the term semantic mapping interchangeably with ontology mapping in this discussion. To express ontology mapping in a mathematical expression, it can be written as mapping ontology $O_1 = (C_1, A_1)$ 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 $A_2 \models f(A_1)$ of which all interpretations that satisfy axioms in O_2 also satisfy axioms in O_1 [24]. For example, if the concept *agent* (C_1) is defined in O_1 by a set of properties as <*broker*, *travel agent* and *officer*> with axioms 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 as <*government agent*, *client*, *spokesperson* and *advisor*> and having axioms 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.

3.3 Challenges in Semantic mapping

Based on the scenarios above, the challenges in semantic mapping can be briefly summarized as follows:

(1) Conceptual incongruence

As pointed out in the previous chapter and exemplified again with scenarios in section 3.1, different ontologies are constructed to specify different conceptualization of their domains of discourse. Since we know that an ontology is an abstract model of how people think about things in a particular domain, it is natural to expect that the set of properties and attributes that are used to define the concept during the process of building the ontology will largely be affected by not only the domain knowledge, conceptual bias and personal viewpoint but also the purpose of the application's use of the ontology [27]. These various objectives and factors would often influence the development of the ontology resulting in different ontologies describing the same domain of discourse. Therefore, it is hard to rely on descriptions or properties of concepts defined locally to find semantically equivalent resources in a distant location. The "*bass*" may be defined by properties and attributes to codify a concept for musical instrument in local ontology; however, the concept with the same name could very well be defined as a kind of fish in a distant ontology. Thus, if we solely rely on the match of the ontology labels to retrieve semantically related resources, the relevance of the result will sometimes be lowered.

(2) Linguistic variation

The linguistic variations that humans used to describe information make it difficult for a computer to recognize the intended meaning of the text. This is because of the ambiguous nature of human language. The ambiguity

can be attributed to several levels of linguistic variation. Here they are limited to lexical and semantic variation only because they are the dominant causes of ambiguity in text analysis with ontology [28, 29]. Lexical variation deals with multiple words having the equivalent meaning while semantic variation relates to a single word having multiple distinct meanings (homonym) based on different context. As an example, "*car*", "*motorcar*" and "*locomotive*" could refer to similar concept "*automobile*" while the word "bat" could have quite different meanings in a baseball game broadcast and in a zoology magazine. Because of this, it is dangerous to rely on the match of the values on properties and attributes to do semantic mapping. This logically rules out the simple use of keyword search and associative text mining to find all possible semantically relevant concepts.

(3) Language divergence

It would be less problematic if there was only one ontology language to express and construct a domain ontology. In reality, there are a number of options in terms of languages and tools to build ontology. DAML+OIL (DARPA Agent Markup Language + Ontology Interchange Language) is a combination of two languages to enable the specification of facts and operation of logical inference. DAML is an extension of Resource Description Framework (RDF). In fact, it is using the RDF triples to define concepts and their associated semantic relationships. Similarly, OIL can be expressed as an extension of RDF. OIL combines more widely used

modelling primitives from frame-based languages and formal semantics and logical reasoning capabilities from descriptive logic. It attempts to provide more expressivity to the language to model knowledge in ontology. On the other hand, there is OWL (Web Ontology Language) created by W3C for publishing and sharing data using ontologies on the Internet. It is derived from the DAML+OIL language so it is also an extension of RDF constructs. OWL comes with three versions: Lite, DL and Full. They provide different levels of complexity and functionality for modelling different ontologies [30][31]. It contains additional expressive primitives and vocabulary to define richer semantic relationships. Due to the divergence in ontology languages, it adds another layer of consideration when performing semantic mappings between ontologies built from different languages. One would need to use business logic in a program to identify information where possible mapping could be provided because values could be stored differently with diverse language constructs. Complex mapping such as 1 to n, n to 1 or even n to n could make the mapping task very complicated.

(4) Trade-off on computational complexity and ease of use

We recognize that we can provide robust semantic mapping between specific domain ontologies with moderately high accuracy. However, it would take quite some time to develop and implement the mapping due to the complexity of many mapping algorithms. Besides, very often the systems based on those mapping algorithms require a fair amount of

human (domain expert) intervention to control the mapping process which may not be desirable in some situations [32]. Furthermore, the mapping relations will need to be re-examined and modified when new concepts are added or new relationships are established in the ontology. Therefore, it would be costly for organizations and inefficient in terms of time to develop the mapping tools to provide interoperability between ontologies for some particular application use. In this view, for most applications, they need to weigh the accuracy of the mapping against the cost and time to provide such a mapping. Then again, most researchers in semantic mapping agree that it is still out of our grasp to provide generic mapping in domain independent ontology with current mapping paradigms [25].

3.4 Review of current approaches

This section presents a brief overview of two approaches on semantic mapping. It is by no mean an extensive review of every detail of these approaches. This only serves to give the audience a general idea of these approaches in order to illustrate different strategies in 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.

(i) GLUE

This is a semantic mapping system that employs machine learning techniques to find mappings between two ontologies. With two domain ontologies, for each concept in one ontology GLUE claims to find the most similar concept in the other ontology [23]. There are a number of features distinct GLUE from other similar mapping systems. First, unlike most 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. The similarity measure that is employed by GLUE is the joint probability distribution. More precisely, it is Jaccard coefficient:

$$Jaccard - Sim(A,B) = \frac{P(A \cap B)}{P(A \cup B)}$$

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.

Overall, according to the results of the experiment performed in [23], the accuracy rate of ontology mapping ranges from 66 – 97%. However, based on the observation to achieve this accuracy, a lot of processing in terms of time and effort to implement the strategy is needed to achieve this accuracy. Remember that in order to find similarity between different concepts, GLUE needs to compute a set of similarity functions and determine which one to use on what constraints. Therefore, to map all

concepts from O_1 to O_2 , the total number of calculations is roughly $4|O_1||O_2|$, where $|O_i|$ is the number of concepts in ontology O_i [23]. Moreover, the relaxation labelling technique is sometimes susceptible to converge to local maxima and the converging condition is still not well known yet. Finally, GLUE will also be confused by linguistic ambiguity that prevents it from distinguishing similar concepts like "*networking*" from "*communication devices*".

(ii) MAFRA

MAFRA (Mapping FRAmework) is another ontology mapping methodology that prescribes "all phases of the ontology mapping process, including analysis, specification, representation, execution and evolution" [33]. 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. The internal processes of MAFRA include: *Lift and Normalization*, *Similarity*, *Semantic Bridging*, *Execution* and *Postprocessing*. The details of each stage will not be discussed here and please refer to [33, 34] for complete references.

One of the most innovative aspects of MAFRA is the use of SBO to process ontology mappings. However, MAFRA heavily relies on domain experts to predetermine the mapping relations between two ontologies. This process could be very tedious and error-prone. On the other hand, when the number and variant of concepts grow in the ontology, it would require modification in the SBO to correct the mapping specification. By extension, when the number of concepts in ontology is very large, it becomes impractical to examine all the classes to find out the mapping relations.

CHAPTER 4: BROWSING WITH SEMANTIC SIGNATURE

This chapter first introduces the concept of semantic signature used to represent a class of concepts and describes how it is constructed with WordNet word senses. Then, it further develops the idea of using a semantic signature to facilitate semantic mapping and demonstrate how a semantic signature can be employed in concept browsing in a distributed collaborative learning network.

In a collaborative learning network, institutions commonly organize their metadata according to their fixed viewpoints without taking a global perspective or dispersed users' interests into account. Coupled with that, as previously pointed out in Chapter 3, various technical difficulties in semantic mapping between independent ontologies make it difficult for traditional keyword-based or label-matching-based retrieval in a distributed learning environment to yield satisfactory results that are consistent with users' perceptions which are often based on local ontological concepts.

To overcome the problems with different conceptual views represented in the local ontologies, a unifying global semantic view can be considered as a potential solution. To assist distributed learning repositories to organize and manage their metadata in compliance with a global semantic view, it is worthwhile to explore the use of a semantic-based search of learning object metadata by category across different repositories to enhance browsing experience in a collaborative learning environment. In this work, the aim is not to

invent a new word sense disambiguation algorithm but to extend and combine existing techniques in semantic mapping, information integration and text retrieval with word sense disambiguation. The goal is to create a semantic mapping strategy using WordNet for cross-repository metadata browsing in a distributed learning network. The result can be used to prove the feasibility and merit of applying semantic-based indexing on metadata for providing an interoperable searching platform in repository networks.

4.1 Semantic Signature Definition

A semantic signature in the concept browsing context can be defined as a logical grouping of representative word senses for a class of metadata. In essence, it is a semantic representation of an ontological concept with important WordNet senses with respect to context in which the concept is used. 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 senses BS_{di} in document d_j

T = all keywords in document d_j

Fav = selection function to find best sense

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

To explain briefly, the semantic signature of a class of metadata is built from a set of important document senses from all metadata records belonging to a

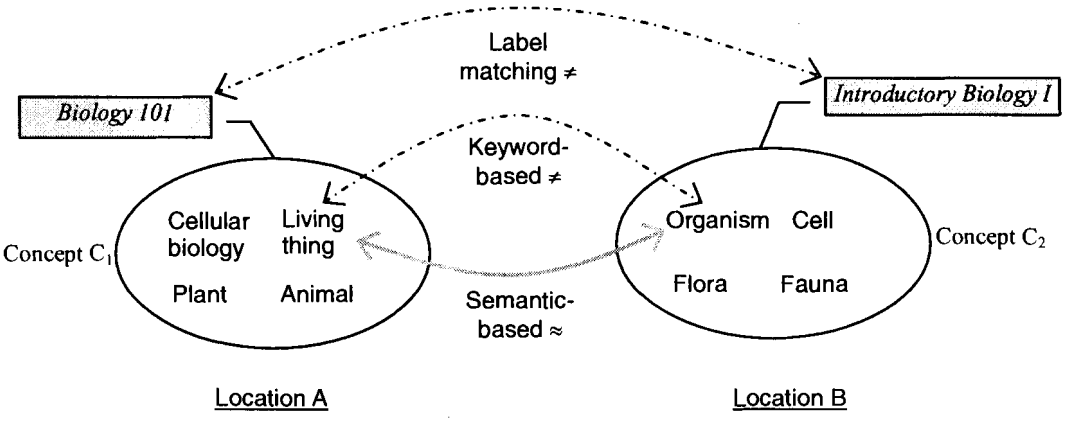
particular class. In turn, document senses are generated from a collection of best WordNet senses for all representative keywords for a particular document.

4.2 Why use Semantic Signature?

Before delving into the details of creating a semantic signature, it is worthwhile to clarify the rational proposition for use of this semantic rich representation. The use of a semantic signature is mainly motivated by three observations. First, metadata for learning resources are generally encoded into semi-structured XML documents (e.g. IEEE LOM or DS) with a set of predefined elements. The content of these metadata elements is textual in nature. However, due to the ambiguity problem of the free text, this makes syntactic-based keyword search ineffective to retrieve semantically relevant metadata. This problem cannot be completely resolved even with the vocabulary of the metadata content that may have been defined with concepts in a local ontology. It is because the definition for similar concepts in distributed repositories could vary from ontology to ontology due to conceptual differences. Therefore, it is natural to expect that the set of vocabulary used will also vary morphologically. Thus, it is believed to be better to develop a unified semantic representation scheme to denote a class of metadata independent on a local ontology to facilitate distributed semantic retrieval. This situation is exemplified with the case in Figure 4.1 that shows that the concept *Biology 101* maybe defined with vocabulary <*cellular biology, living thing, animal, plant*> in a local ontology while concept *Introductory Biology I* maybe defined with vocabulary <*cell, organism, fauna, and flora*> in another distant ontology. Both label-matching and keywords-based

mapping would not be able to tell if *Biology 101* is in fact conceptually equivalent to *Introductory Biology I*. However, if a WordNet signature were able to wrap both geographically separated concepts into a semantic representation, it would enable semantic mapping to understand *Biology 101* is indeed conceptually like *Introductory Biology I* because the property *Living thing* is a child of *Organism* and *Flora* is synonymous with *Plant*, and there are other semantic relationships connecting

Figure 4.1 Semantic vs. syntactic matching in different ontologies



these two concepts. In this regard, when looking at theses two concepts at a semantic level, they are very similar. Hence, semantic representation can be used as a key to discover ostensibly unrelated concepts in distant repositories.

Second, the use of WordNet to derive the semantics of word term originate from another important observation that given the topic of a text, there is a high probability that most of the words are closely related semantically to other words used to describe the topic. For example, in a metadata document about *organic chemistry*, it can be expected that many words related to

compound, *molecule*, *bonding* etc will be found. According to [35, 36], when mapping a set of closely related words to WordNet, the returned word senses will be concentrated in an area of high conceptual density with minimum conceptual distance. Therefore, if this hypothesis is correct, then the use of WordNet sense to serve as classifying feature may generate good results compared to the use of keywords because the semantic signatures for similar conceptual topics will be expected to share many common word senses. Then, a distance function can be employed to measure the closeness between semantic signatures to distinguish one class from another.

Third, recent advancements in WordNet make it a popular tool for word sense disambiguation or for semantic rendering in Natural Language Processing research community [36-40][41]. In this case, WordNet can be easily utilized as a mediatory source for providing lexical information to replace keywords representation in most text retrieval approaches. Therefore, to combine techniques from text retrieval with semantic mapping, it is plausible to produce a semantically rich signature to characterize a class of metadata. As a result, semantic-based categorical searching can be realized by matching signatures rather than relying on matching vocabularies in potentially different ontologies. The mediatory approach to provide semantic mapping is believed to be most cost effective since it is not algorithms dependent like the machine learning approach and human expert dependent like some ontology mapping methods.

To summarize, the major thrust of using the semantic representation of a category of metadata is to avoid the drawbacks of keywords-based retrieval as

mentioned before, and more important to enable the retrieval of semantically related metadata to enhance the relevance of the result without resorting to complicated semantic mapping algorithms.

4.3 Building Semantic Signatures with WordNet senses

The generation of a semantic signature for a class of metadata is divided into three distinct phases. In what follows, the general architecture of the methodology will be illustrated and then each phase will be discussed in detail together with illustrating examples.

4.3.1 General Methodology

In devising a methodology for creating a semantic signature for better browsing of distant metadata semantically, the methodology relies heavily on the following assumptions:

- The aggregates of all semantic information from all metadata records annotated by a concept are a good semantic representation of that concept. In fact, metadata is a semantic description of an instance of a concept in ontological framework.

- It is assumed that semantic information of a class can be approximated by the set of important word senses from all metadata for the class.

- Besides, semantic word senses specific to the context can be found through WordNet for terms extracted from metadata.

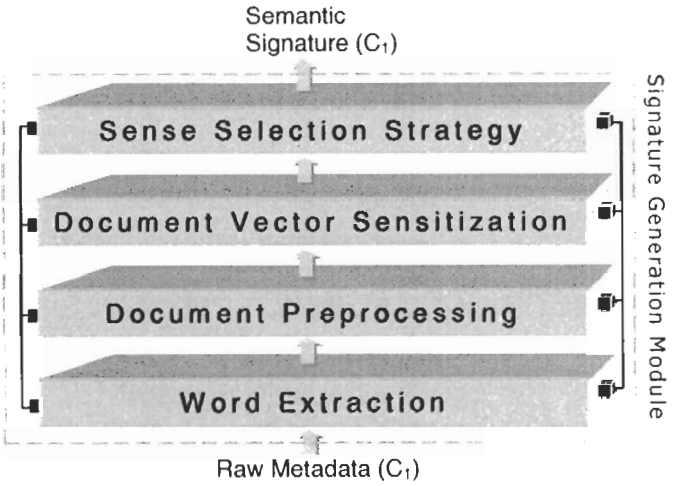
- Finally yet importantly, it assumes that the local semantic signature for a class of metadata is similar to signatures for metadata of semantically equivalent concepts in distant repositories.

The methodology uses a k-Nearest Neighbour (kNN) search algorithm [8] to classify semantically relevant concepts in distant repositories based on local semantic signatures. The instances (metadata) of concepts in a local repository serve as the training dataset. Based on semantic features of local metadata, semantic signatures for each class of concepts are formed. Assuming remote repositories create signatures for their concepts in a similar way, to find semantically relevant concepts in distant repositories, a distance function is defined and used to measure closeness between query signature and semantic signatures for concepts in distant repositories. Eventually, the metadata annotated with the k most similar classes of concepts related to the query signature will be retrieved from remote repositories.

The core of this methodology depends on a good semantic representation of underlying concepts in WordNet word senses. To discover a semantic signature of metadata for concepts, a signature generation module is developed. As shown in Figure 4.2, the module contains four phases. Phase I is called *Word Extraction*. In this phase, representative features will be extracted from the metadata document. Phase II is called *Document Preprocessing*. In this phase, irrelevant information will be eliminated and all non-noun words will be removed. Phase III is called *Document Vector Sensitization*. In this phase, all the

representative keywords will be used as a seed to find the corresponding word senses from WordNet.

Figure 4.2 Semantic Signature Generation Framework



Phase IV is called *Sense Selection Strategy* (S^3). In this phase, the best word sense to represent each word term will be selected among all senses.

4.3.2 Signature Generation in Action

Phase I: Word Extraction

At first, the input of metadata will presumably be in IEEE LOM format. Otherwise, all metadata will be transformed to comply with the standard using the XSLT transformer. Then, adapted from the Edmundsonian paradigm [42], content from *<Title>* and *<Description>* elements will be extracted to represent the whole metadata document (and indirectly, the learning object itself). It is believed that the content from these two elements carry important weight as a cue phrase to be able to represent the whole document [43]. 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

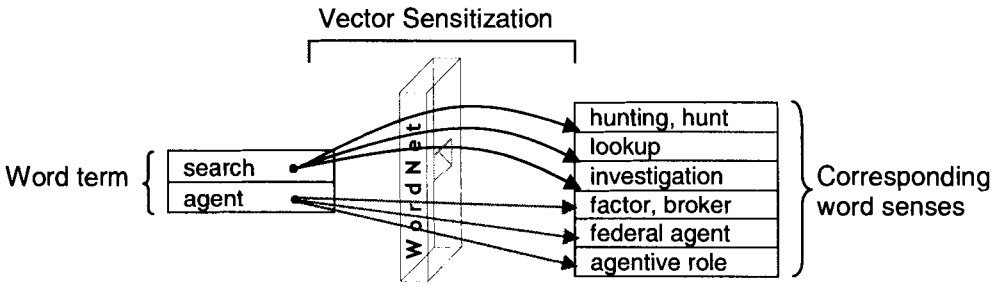
The condensed metadata with only the *<Title>* and *<Description>* elements will be subjected to cleaning in this phase to remove all stopwords, punctuation information, numerical values and irregular symbols. Next, all non-noun words will be removed using a part-of-speech tagger except some commonly used phrasal words which carry a special sequence for specific intended 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 that this approach only uses nouns as the base keywords is according to [44, 45], long phrases are not easily disambiguated compared to single noun terms and binary noun terms. Through previous experiments in [45], it has been shown that in some situations the accuracy of using phrases as distinguishing features for document classification in fact will not necessarily be higher. On the other hand, it is believed that the use of noun carries good salient expression to serve as distinguishing feature for doing text classification [46].

Phase III: Document Vector Sensitization

After all irrelevant information has been eliminated, the physical metadata document will be projected into the vector space model. The document vector becomes the logical representation of the physical metadata. Next, most significant terms across all document vectors are selected using TFIDF weighting

scheme (Chapter 2) to represent a category of metadata. After that, each word term with a TFIDF score higher than the threshold is sent to WordNet to retrieve the corresponding word senses and its definition. The threshold is determined by a trial and error approach since there is no standard way to determine the best threshold in the TFIDF approach. It is a well-known disadvantage for this method [10, 47]. The rule of thumb is to find a threshold that can cleanly separate relevant and irrelevant data. A single word term could have multiple word senses retrieved as in Figure 4.3. The word "search" can be mapped to WordNet senses as <*hunting, hunt*>, <*lookup*> and <*investigation*>. With such mappings, a single word term can be denoted by a triple construct in the form $\langle T, S, D \rangle$ where T is the original word term, S is the synset of T and D is the definition of T. Take the word term "*search*"

Figure 4.3 Word term to WordNet sense mapping



as an example; after the sensitization, it becomes $\langle search - \{hunting.\ hunt\} =$ "*the activity of looking thoroughly in order to find something or someone*" (TFIDF 0.623101)$\rangle$ in triple construct. The triple construct format will be used to substitute the original word term in the master document vector. However, since a single word term

could be mapped to different word senses through WordNet, and each word sense is represented in a synset that may have multiple synonymous terms, the length of the vector will grow considerably. This problem will be addressed in the next phase.

Phase IV: Sense Selection Strategy (S^3)

This is the last, and the most crucial phase in this method. It is to choose 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 cases, after the sensitization procedure the dimensionality of the vector will grow significantly. 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. With such a high dimension, it will not only negatively affect the efficiency of the similarity computation but more seriously many of the senses are actually noise that does not carry actual meaning of the word in the context of a document. Including irrelevant senses will distort the semantic representation of the signature and lower the accuracy in a 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 meanings are clarified in a specified context [46]. Without that, the semantic signature is just a bigger collection of keywords and would have little use in identifying the class of metadata based on a semantic relevance in the signature. Therefore, it is necessary to find a way to

reduce the dimension and select only the sense that conveys the main idea of the word from the author's perspective.

To select the best sense to represent a word term, a contextual-based Senses Selection Strategy called S^3 is applied on retrieved word senses. The strategy is based on the assumption that the local contextual information of a document serves as a good hint to choose the best sense to represents the actual meaning of the word term. The S^3 approach can be summarized in the following algorithm:

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

For each metadata document $D \in C_l$
  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 S_l 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_i

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 synsets of other word terms in the same document. As shown in Figure 4.4, a document vector D_1 consists of three words say "Windows", "OS" and "Computer". After retrieving all word senses from WordNet for each word, each word may contain one or more than one synsets. In this example, the word term "*Windows*" has three senses represented by three synsets. They are "*<windowpane, window>*", "*<operating system, computer screen>*" and "*<framework, opening>*". To find the best sense for word term "*Windows*" using strategy 1, it computes the associative frequency of each sense

Figure 4.4 Associative frequency calculation between word senses

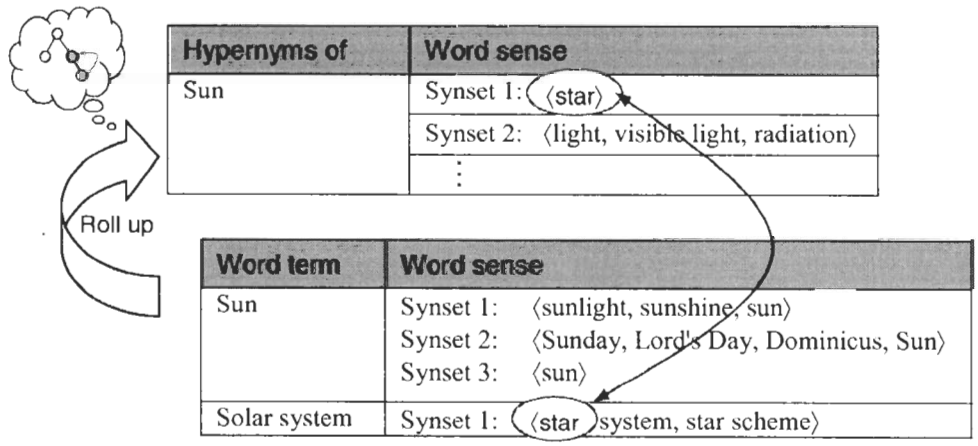
Document vector D_1

Word term	Word Sense
Windows	Synset 1: ⟨windowpane, window⟩ Synset 2: ⟨operating system⟩ ⟨computer screen⟩ Synset 3: ⟨framework, opening⟩
OS	Synset 1: ⟨os⟩ Synset 2: ⟨osmium, Os, atomic number 76⟩ Synset 3: ⟨operating system, OS⟩ Synset 4: ⟨oculus sinister, OS⟩
Computer	Synset 1: ⟨computer device, computing machine, data processor⟩ Synset 2: ⟨calculator, reckoner, figurer, estimator, computer⟩

in all synsets for "*Windows*" with other word terms' synsets. Hence, the sense "*operating system*" for word term "*Windows*" has a high associative frequency with senses like "*computer device*" for word term "*computer*" and with sense "*operating system*" for word term "*OS*", compared to senses like "*framework*" or "*opening*" to other word terms in the document vector with low af. Associative frequency is the metric used to measure the occurrence frequency of a particular word sense of a word term in the document. In this case, the sense "*operating system*" will be marked as most frequently occurring sense for the word term "*windows*" in strategy 1. From this, the most frequently occurring word sense will be used to substitute as semantic representation of the word term.

Next, if the word sense of a word term cannot be discriminated using strategy 1, the algorithm generalizes the word term to the k-order parent senses. In this approach, the value of k is 1. In other words, it will generalize to the

Figure 4.5 Word sense generalization to immediate (1-k) parent



immediate parent sense. Referring to Figure 4.5, 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 "*Sun*" will be rolled up to its immediate parent through hypernym (is-a) relation in WordNet hierarchy. Then, the parent's synset will be used to calculate the associative frequency with respect to other word senses from other word terms. The reason that it uses immediate parent senses (k=1) to compute the associative frequency is that according to [23, 48], 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. Then, it would not be a good feature to be used for the classification purpose.

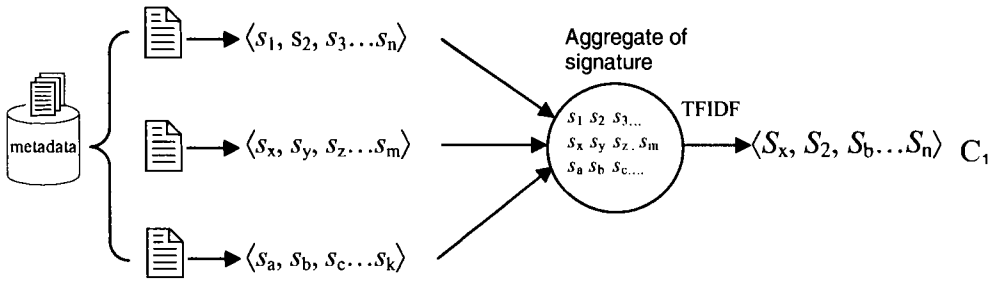
Finally, as arranged by WordNet, the word senses retrieved from WordNet for a particular word are a partial order set ranked by popularity in English usage. If the previous two strategies can not find the best sense to represent the word term, then the most popular sense offered by WordNet will be adopted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

At last, the best word sense will be selected based on the preferential order of strategy 1 > strategy 2 > strategy 3. In other words, the sense selected by strategy 1 will be used as the best sense over the other two strategies. The principle behind this preference ranking is derived from observations and the hypothesis that the local context is the most specific and relevant to provide contextual meaning of a sense for word term. 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 this strategy, then it will compare the immediate parent sense to the other word senses to check if the parent sense is the most frequently occurring sense for the underlying word term. Eventually, it resorts to the most popular sense 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 will become a semantic signature of a document. In order to generate the final semantic signature for a class of documents referring to a particular concept, the TFIDF scheme will be applied again to each word sense in all document signatures. Based on the score, the most relevant senses to characterize the class of metadata will be aggregated to form the final signature for the class as in Figure 4.6.

Figure 4.6 Aggregation of document signature to generate class signature

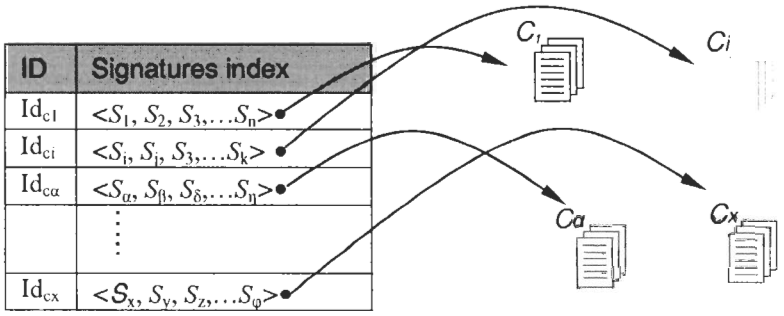


4.4 Federated Concept Browsing in a Repository Network

After the semantic signatures were generated, they can be used to index the actual class of metadata for fast distributed browsing. A common technique in database indexing, the inverted file system, can be applied here. As shown in Figure 4.7, a collection of semantic signatures as unique identifiers representing

concepts in a local ontology can refer to a set of metadata documents. Unlike normal inverted indexing, for the sake of simplicity, in the current model each

Figure 4.7 Inverted index by Semantic Signature



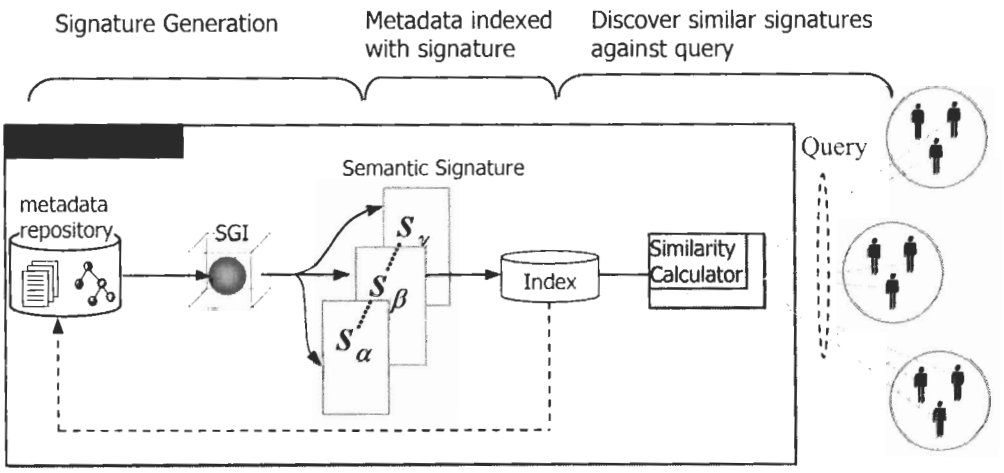
signature can only represent one class of metadata even if there may be shared elements among the signatures. In order to realize the semantic indexing of metadata of learning resources, a toolkit called Signature Generation Indexer (SGI) is developed to generate semantic signatures for metadata of learning resources. The generated semantic signatures will be used for metadata indexing in order to facilitate searching and retrieval of metadata. Focusing on the efficiency, the design of SGI is to allow users to produce semantic signatures for classes of learning resources metadata easily without tedious human interaction, or complicated implementation (see Appendix A).

4.4.1 Browsing distant metadata with semantic signature

In the end, the ultimate goal is to achieve semantic search on E-learning topics 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 fulfil this anticipation, it is important to assume that all participant repositories

in the collaborative network must employ the same strategy to index learning resources metadata with WordNet semantic signatures. The overall operation of semantic-based browsing of learning resources metadata is shown with Figure 4.8.

Figure 4.8 Integrated process of semantic-based browsing of metadata



When users initiate a query by selecting the view of a specific topic which is similar to a class of metadata from a local user interface, the corresponding semantic signature representing the topic is retrieved from the local database. Then, it is sent across the network to participating learning repositories. The query, in the form of a semantic signature, is entered into the Similarity Calculator in distant repositories. The Similarity Calculator is used to compute the similarity to topic signatures in each of the learning repositories. The cosine similarity [7] is adopted as a distance function, so that the more matched elements in 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 title sense makes a higher contribution to the overall score than the ones from the description tag.

After all, in order to ensure the global accuracy of the result, results from participating remote repositories are merged and sorted in descending order based on the cosine similarity score. Then, the top k (k=5) topic of metadata are offered as an answer to local query.

CHAPTER 5: EXPERIMENT AND EVALUATION

The efficacy of the proposed semantic mapping strategy is tested and evaluated in two different settings. The primary goal of the evaluation is to validate the use of WordNet to provide semantic knowledge to represent categorical data for semantic browsing in a federated network. Secondarily, it evaluates the usefulness of using immediate parent concept as a substitute for word terms in selecting the best sense. The design of two experimental settings is to fulfil these objectives.

5.1 Evaluation settings

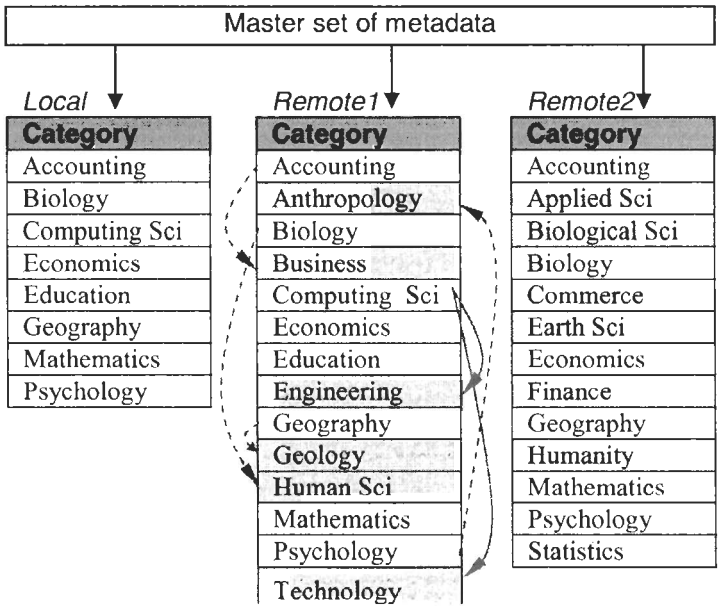
First, in order to test the hypothesis of using semantic signatures to enable semantic browsing and improve relevance, simulated distributed concept retrieval must be run to measure the relevance rate compared to the traditional keyword-based method. To replicate the distributed repositories in a collaborative E-learning network, three independent databases are set up. They are referred to as "*local*", "*remote1*" and "*remote2*" where the *local* of course denotes a local data source and both *remote1* and *remote2* denote distant data sources. A single master set of 2235 metadata in 8 different categories is distributed evenly in number and randomly in nature into the three simulated databases. The gathered learning resources metadata have been transformed to conform to IEEE LOM format. The dataset characteristics will be discussed in more detail in Section 5.3. After the distribution, the *local* database contains metadata that represent

the set of training data for the classifier. During the training phase, a kNN classifier will use the metadata records from the *local* database to learn the features that identify the class of metadata. It starts by extracting important word terms from each class of metadata and projecting them into a vector space model. Next, after running through the signature generation module, a semantic signature for each class of metadata will be produced and used to index the class of metadata in the database.

The datasets in both *remote1* and *remote2* will be controlled to model the situation of potentially different ontological classification in a distributed environment. To simulate the effect of varied labelling of classes in different ontologies, the original 8 categories of metadata will be expanded to 14 categories in *remote1.* The reason to have 14 categories is to allow some mislabelling in some classes but not all, due to the limited dataset. In an ideal situation, it would be better to have two large datasets that are annotated with two different ontologies with the known mapping. In the case of *remote1*, the 6 derived categories are labelled with different class names from their respective sources, and metadata are reallocated to these derived categories from their original categories. Each newly derived category contains metadata belonging to the same class. To illustrate, part of the metadata from the category "*computing science*" will be distributed to derived categories "*technology*" and "*engineering*" respectively in *remote1*. Thereby, the metadata for the concept "*computing science*" is now grouped into "*computing science*", "*technology*" and "*engineering*". Essentially, this simulates the situation that a class "*computing*

science" could be categorized differently into classes like "*technology*" and "*engineering*" in another repository. The same distribution principle applies on the *remote2* database that includes 13 categories with 7 derived categories. Figure 5.1 shows the metadata distribution in 3 separate databases diagrammatically. Derived classes are shaded.

Figure 5.1 Metadata distribution in simulated distributed data sources



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

Second, it is relatively trivial to set up the experiment to test the effectiveness of using "*immediate-parent*" in the hypernym relationship to replace word sense in generating a semantic signature. By modifying the sense substitution process, a semantic signature can be generated without the sense generalization effect. The metric that we used to test the value of "*immediate-parent*" is different from the previous one. Instead of comparing the derived relevance, here we compare the raw similarity score to evaluate the effect of "*immediate-parent*" in finding the matched signature.

5.2 Assumptions

The experiment is carried out based on a limited set of assumptions. First, it is built on the belief that in a federated E-learning environment, certain cooperative agreements exist to govern how to provide an interoperable platform for participants to share data. In our case, this implies the agreement to index classes of metadata with WordNet semantic signatures for federated concept browsing. Second, it assumes that a large number of conceptually related metadata reside in separate repositories despite the fact that they may be labelled differently. These assumptions appear to be pragmatic in the context of a collaborative research network, at least in the scope of LORNET[5] for which an interoperable platform for cross-repository information integration is crucial for its success.

[5] http://www.lornet.org/

5.3 Dataset Description

Since there is no publicly available dataset on learning resources metadata that is consistent and large enough for our purpose, to conduct the experiment metadata are acquired through a number of different sources. Table 5.1 shows the categories of metadata acquired and their respective sources. In total, 8 different categories of 2235 metadata are acquired. They are *Accounting*, *Biology*, *Computing Science*, *Economics*, *Education*, *Geography*, *Mathematics* and *Psychology*. The choice of the category is arbitrary and is solely dependent on the abundance. The dataset is partitioned into training and testing groups. As mentioned, the *local* database stores the training dataset while *remote1* and

Table 5.1 Source and Category of Metadata

Category	Source	Number 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

remote2 store the testing dataset. The class labels for all metadata are known in advance. Metadata are distributed randomly to training and testing groups using the Microsoft Excel random generator. The training group contains 723 metadata records (*local*) while the testing group contains 1512 records (*remote1* and *remote2*).

5.4 Metric

In order to gauge the effectiveness of the system, three standard metrics in information retrieval are used in the evaluation of the system performance: they are *Recall*, *Precision* and *F-measure* [7]. *Recall* (R) is defined as the number of relevant documents retrieved over the total number of relevant documents found in the collection. The *Recall* can measure the coverage of the system and its identification capability. *Precision* (P) is defined as the number of relevant documents retrieved over the total number of documents retrieved. The *Precision* is to measure the reliability of the returned result. Mathematically, they can be written as follows:

$$P = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{retrieved\}|}$$

$$R = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{relevant\}|}$$

Both precision and recall have value lying between 0 and 1. In general, the closer these values are to 1, the better the system is. On the other hand, the *F-measure* is a weighted harmonic mean[6] of P and R which combines both the precision and recall into a single formula:

$$F_{measure} = \frac{(\beta^2 + 2.0) \times P \times R}{(\beta^2 \times P) + R}$$

[6] http://en.wikipedia.org/wiki/Harmonic_mean - Harmonic mean is defined as $H = n/ (1/a_1 + 1/a_2 + .. + 1/a_n)$ where $a_1 .. a_n$ are positive real number. The harmonic mean provides the correct notion of average.

where β is the relative importance given to recall over precision. In this case, both precision and recall are of equal importance, and therefore the factor β is 1. The *F-measure* function assumes values in the interval [0, 1] [7]. Similar to precision and recall, a high value would indicate an effective system when both precision and recall are high.

To measure the improvement of using sense generalization in the sense selection strategy, a raw similarity score will be used. The higher the score, the better the chance that the class will be classified correctly.

5.5 Limitations

A set of issues regarding the usability and performance of the methodology is worth mentioning here. First, the major issue would be the size of the training and testing dataset. The small data corpus does not satisfy the need of the learning algorithm to correctly form the base signature of each class and may influence the predictive ability of the base signature as the matching template. On the other hand, when more and more metadata are added to each class, an incremental update on the base signature of each class is needed to reflect new elements found in the recent metadata. However, as this is not within the scope of this research, it is not supported by the current implementation.

5.6 Results

Results from the semantic-based concept browsing are compared with the traditional keywords-based browsing. The keywords-based browsing is to search for relevant concepts based on the match of user supplied keywords with

keywords extracted from <*Title*> and <*Description*> elements. First, representative keywords from elements <*Title*> and <*Description*> are extracted from all metadata records for each concept. Then, the most representative keywords to characterize the concept are selected based on the TFIDF score. Next, the selected keywords are used to index the respective concepts. When finding the relevant concepts, the keywords provided by the users are used to calculate the cosine similarity score against the index keywords. The top 5 most relevant results will be returned as the answer. These results will be compared against the results from semantic-based browsing.

The precision and recall are calculated based on the top 5 results returned from the two remote repositories. In Table 5.2, the rows represent the concept categories while the columns list the results of precision, recall and F-measure for both semantic-based (columns 'S') and keywords-based (columns 'K') browsing. The average scores of semantic-based approach on precision, recall and F-measure are all 0.86. The average scores of keywords-based approach on precision, recall and F-measure are 0.54, 0.65 and 0.58 respectively. This shows that using a semantic signature can improve retrieval relevance in terms of recall and precision on E-learning topics. In most categories, the semantic based retrieval out performs the keywords-based retrieval.

Table 5.2 Comparison on precision, recall and F-measure on concept retrieval

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

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

Table 5.3 Comparison of similarity score using *1-k* parent generalization on remote1

Category	Cosine score with sense generalization	Cosine score w/o sense generalization	Percentage change (%)
Accounting (132)	0.5037	0.4987	1
Biology (92)	0.3448	0.3516	-1
Computing Science (102)	0.3722	0.3139	19
Economic (147)	0.6086	0.5957	2
Education (59)	0.5344	0.2835	88
Geography (88)	0.5273	0.3625	45
Mathematics (58)	0.6436	0.3452	86
Psychology (65)	0.4513	0.3219	41

Table 5.3 shows the results of the experiment evaluating the contribution of the process of sense generalization. In Table 5.3, the rows represent the results for each category while the columns represent the cosine score with sense

generalization and the cosine score without sense generalization, as well as the percentage change between these two scores. The cosine score with sense generalization ranges from 0.3448 to 0.6436 while the cosine score without sense generalization ranges from 0.2835 to 0.5957. The percentage change ranges from 1% to 88%. From Table 5.3, it has been observed that the margin of improvement in cosine score is larger in categories with less number of metadata records. On the other hand, there is negligible improvement in the cosine similarity score in the categories of "*Accounting*" and "*Economics*" when using hypernym generalization compared to cases without using generalization. The cosine similarity score is in fact decreased in the category of "*Biology*" when using hypernym generalization compared to cases without using generalization.

5.7 Interpretation

As opposed to the classical or traditional keywords-based representation, semantic-based indexing with WordNet senses can include more lexicon information than a simple syntactic approach. This implies more features will be added to the class signature representation. Since more features are added, this 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 might be expected the precision is 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. In the category of "*Biology*", there is no difference in terms of retrieval relevance using keywords-based or semantic-based

representation. 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 improves the cosine similarity score in some of the categories while there is no significant increase on categories of "*Accounting*" and "*Economics*". In the category of "*Biology*", the cosine similarity score is actually decreased when using the hypernym generalization. This may be due to the highly specialized words used in the domain of "*Biology*". Thus, in this case using sense generalization may in fact reduce the matching possibility in similarity calculations. With this result, further experimentation and analysis are needed to fully understand the impact of sense generalization in classification of metadata.

Therefore, combined with a good contextually based sense selection strategy, WordNet as a mediator can provide a source for ambiguity resolution and semantic information for the process of semantic browsing. Coupled with that, the selection of a kNN algorithm as the classifier also contributes to the better performance of the system.

The kNN classifier is an instance-based classifier. The performance of instance-based classifiers is more dependent on sufficiency of a training set than it is the case with 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 that more terms or term combinations important for content identification may be missing from the training sample documents. This will

negatively affect the performance of a classifier. Nevertheless, an ontology (e.g. WordNet) guided approach seems to somewhat reduce the negative influence of this problem. The replacement of child concepts with the parent concept through hypernym relationships appears to be able to discover an optimum concept set without adversely affecting performance. This is particularly evident in the classes with a small set of data. In that situation, signature-based retrieval is superior to the keywords-based method to a larger margin compared to the classes with more data. Therefore, by using hypernym generalization an important term that resides low in concept hierarchy may be mapped to a parent concept and included in the class signature for comparison, even if this term is not included in the training set.

CHAPTER 6: CONCLUSION AND FUTURE DIRECTIONS

Semantic-based concept mapping is a critical step in many data management systems particularly in a distributed environment. For an E-learning repository network to be effective, it is important to provide an interoperable platform for learners to access learning objects, and for instructors to discover semantically relevant learning objects for reuse.

Due to the diversity of ontology in a distributed environment, it is difficult to use keywords-based browsing to discover semantic relevant information. To enable semantic browsing, in most situations a complete semantic mapping schema is needed to enable semantic retrieval. To provide such semantic mapping manually is labour intensive, time consuming and error prone. Hence, it is important to develop techniques to automate the mapping process. Given the rapid advancement in WordNet, it is interesting to see if it can be used as a mediator to provide enough semantics for categorical classification in the area of learning object metadata. In essence, it is useful for cross ontology communications by providing a semantic representation of ontological concepts with coherent WordNet senses to create correspondences between concepts.

This work presents important reflections on the exploratory use of WordNet to provide semantic mapping between remote learning repositories in order to enable semantic-based concept browsing.

6.1 Conclusion

This research offers two key contributions. First, it gives a new light weight semantic (ontology) mapping approach to enable cross platform concept browsing in a federated network. Many current practices in semantic mapping require intensive user involvement to provide mapping information in the case of a complex ontology, or resort to a complicated heuristic or rule-based machine learning approaches that could be dataset dependent and require user input as well. This work shows an effective automatic mapping technique that can allow federated concept browsing with semantic signatures. Evident by the experimental results, it establishes the merit of using WordNet to provide semantic knowledge for metadata classification in any domain. The merits include the provision of the semantic representation of categorical data and increased semantic relevance in categorical browsing.

Secondly, by using word sense generalization during the sense selection process, it was shown that it successfully reduces the dimensions in the semantic signature. However, the contribution of sense generalization to increasing the opportunity to find similar signatures by increasing the matching features is not conclusively supported by the experimental results. This creates an incentive to explore the use of other sense generalization techniques to improve the signature matching process.

Although this thesis primarily focuses on the discussion of the E-learning repository network, we believe that the validity of the methodology described in

this thesis can be easily extended to other collaborative networks with minimum modification.

6.2 Future Directions

As demonstrated through the evaluation with a constrained dataset, the use of WordNet to provide semantic referencing between different E-learning repositories can show moderate improvement to enhance the relevance of global concept browsing. However, in order to validate that the same methodology can be applied on other metadata or semi-structured documents, more experimental evidence needs to be collected on different datasets. Regarding the evaluation, it is believed that a larger set of testing corpus with diverse classes of metadata needs to be acquired. This can not only improve the effectiveness of the kNN algorithm but also further establish the validity of the methodology. Taking this mediatory approach to a broader perspective, it is perhaps useful to include multiple thesauruses, which could consider domain knowledge, for rendering semantics to word term instead of relying only on WordNet.

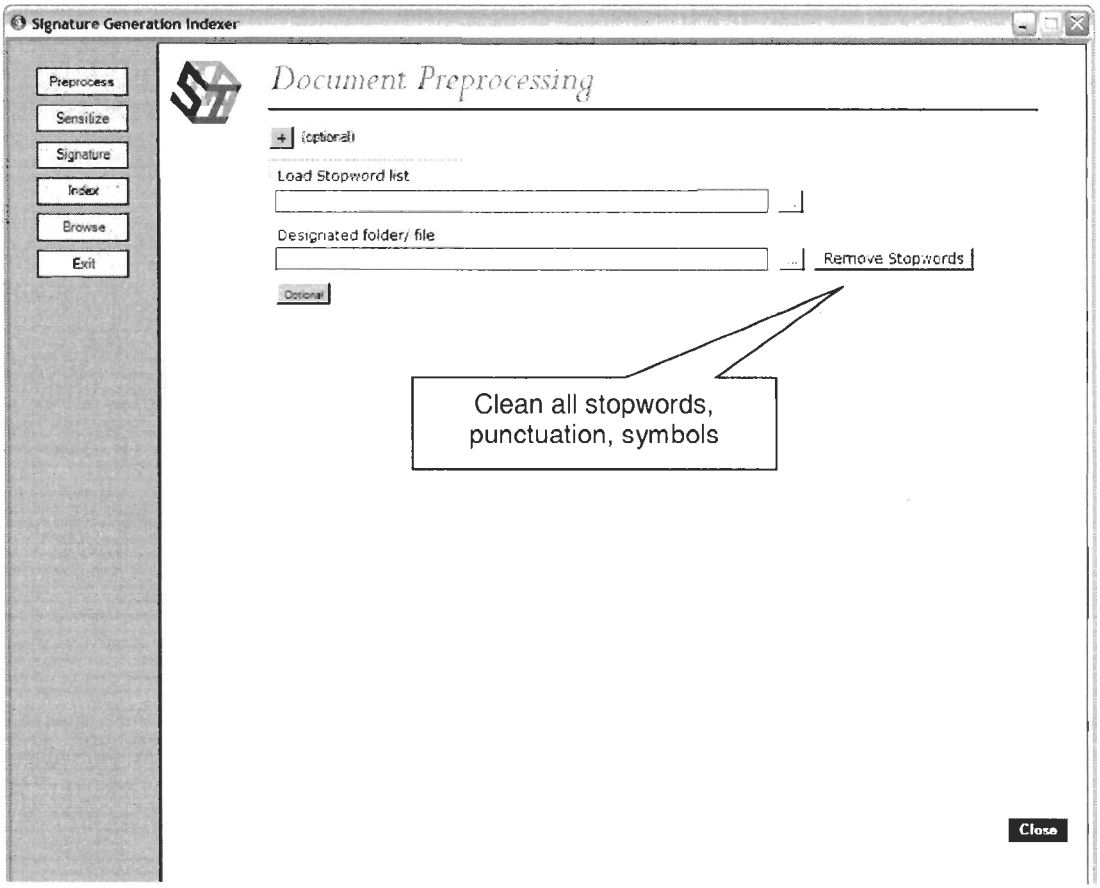
In terms of fine-tuning the suggested method, there are several areas that could be improved. First, more vigorous natural language processing techniques can be utilized to extract meaningful features for sense representation. For example, the inclusion of other part-of-speech word terms (e.g. verbs and noun phrase) and noun phrases may provide sources for identifying key senses for semantic representation. Furthermore, the use of a local domain ontology combined with heuristic-based constraints may also improve the selection of target word terms for semantic characterization. On the other hand, using

semantic distance in WordNet to expand the selection of other word senses as a substitute for the original sense should be examined to test if it can produce a better generalized word sense representation without lowering the precision in the classification. In the classification, other algorithms like Bayesian-based approach or ID3 can be adopted to replace the kNN. Finally, there are some performance improvements that can be achieved by modifying the program.

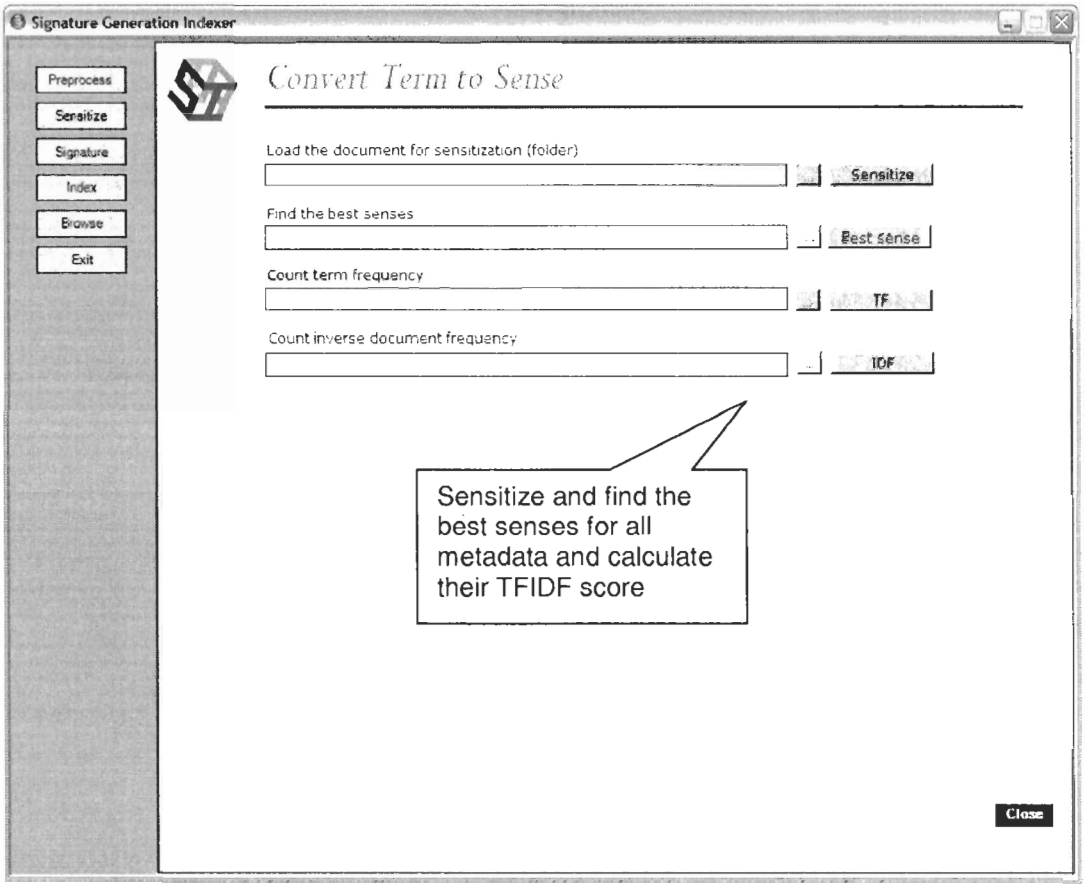
APPENDIX A

The Signature Generator Indexer (SGI) is implemented with 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 separate module implemented in C# as well and connected to the index database.

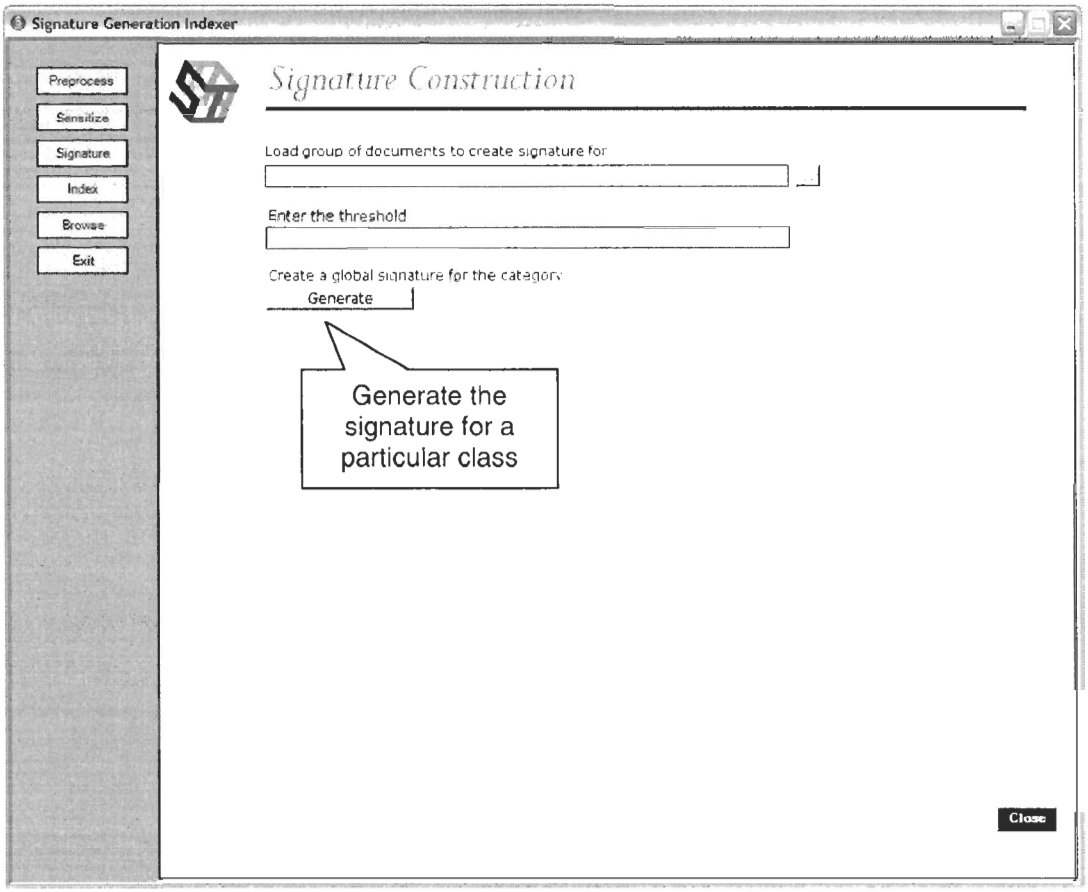
Document Cleaning Module



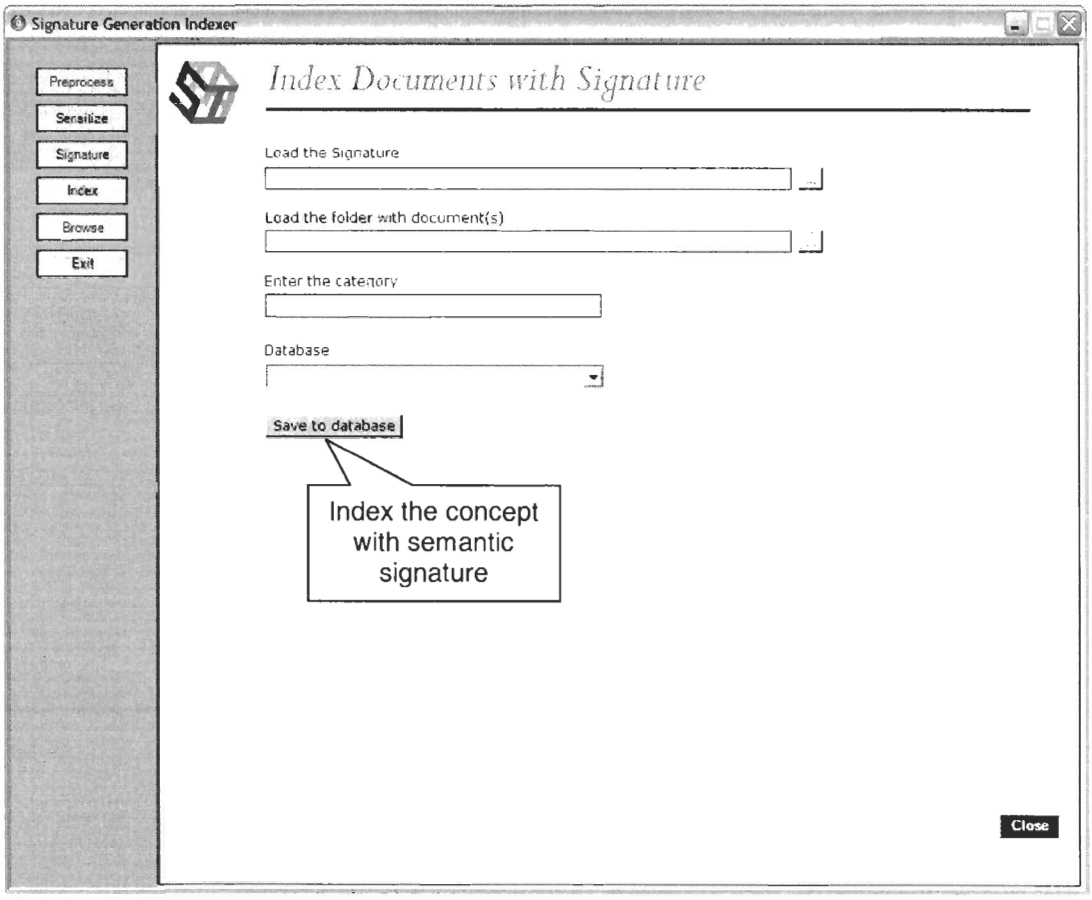
Document Sensitization Module



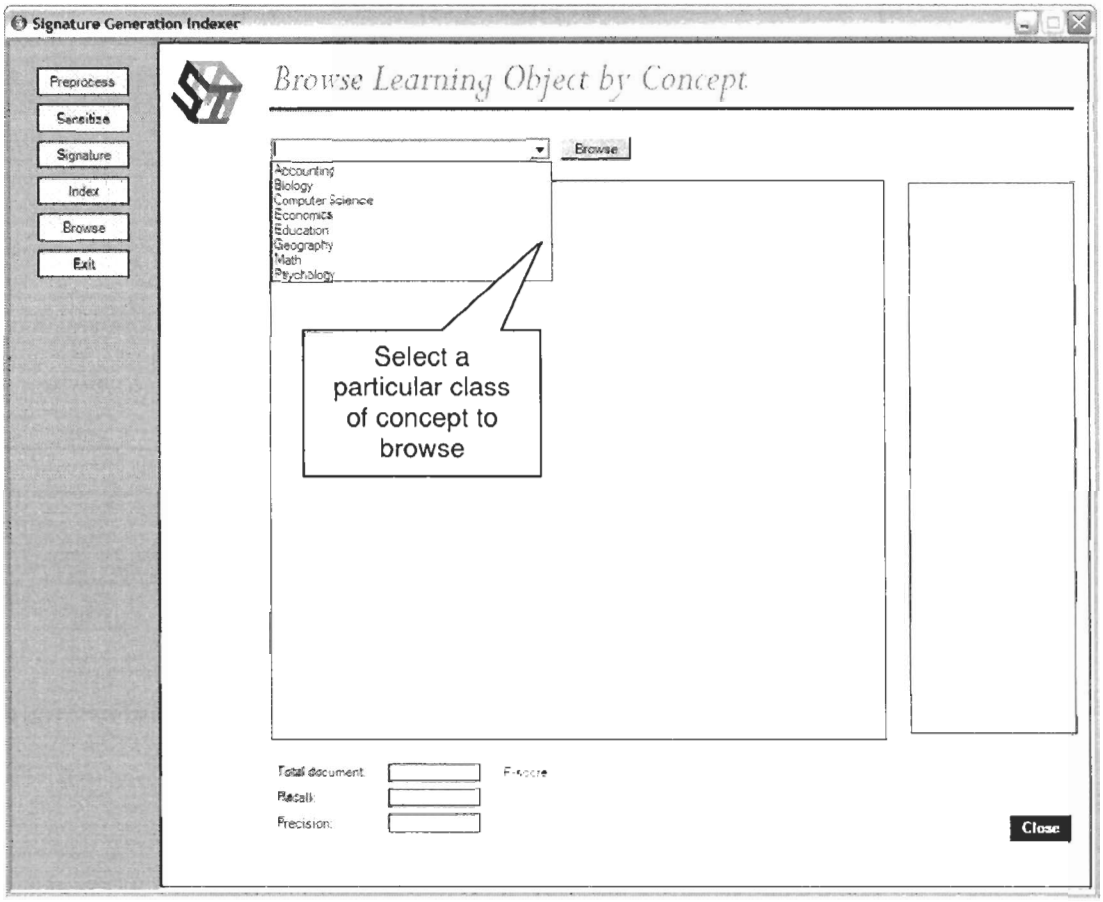
Semantic Signature Generation Module



Concept Category Indexing Module



LO Concept Browsing Module



BIBLIOGRAPHY

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

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

[3] A. Tan, "Text mining: The state of the art and the challenges", http://www.ewastrategist.com/papers/text_mining_kdad99.pdf, Kent Ridge Digital Labs, 1999.

[4] Z. Bellahsene, "Data integration over the Web", *Data & Knowledge Engineering,* vol. 44, pp. 265-266, 2002.

[5] M.T. Özsu and P. Valduriez, *Principles of distributed database systems,* Upper Saddle River, N.J.: Prentice Hall, 1999.

[6] W.B. Croft and Center for Intelligent Information Retrieval, *Advances in information retrieval :recent research from the Center for Intelligent Information Retrieval,* Boston: Kluwer Academic, 2000.

[7] R. Baeza-Yates and Ribeiro,Berthier de Araújo Neto, *Modern information retrieval,* New York: ACM Press, 1999.

[8] J. Han and M. Kamber, *Data mining :concepts and techniques,* San Francisco: Morgan Kaufmann Publishers, 2001.

[9] Y. Jung, H. Park and D. Du, "An Effective Term-Weighting Scheme for Information Retrieval", Department of Computer Science and Engineering, University of Minnesota, Minnesota, U.S, Tech. Rep. TR 00-008, 2000.

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

[11] F. W. Kroon, "Linguistic Variation in Information Retrieval Using Query Reformulation", PhD thesis, Simon Fraser University, 2002.

[12] J. D. Ullman, "Information integration using logical views", *Theoretical Computer Science,* vol. 239, pp. 189-210, 2000.

[13] J.A. Larson, *Database directions :from relational to distributed, multimedia, and object-oriented database systems,* Upper Saddle River, NJ: Prentice Hall PTR, 1995.

[14] A.P. Sheth and J.A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases", *ACM Comput.Survey.,* vol. 22, pp. 183-236, 1990.

[15] Hatala, M. and Richards, G., "Global vs. Community Metadata Standards: Empowering Users for Knowledge Exchange", in International Semantic Web Conference 2002, pp. 292-306, 2002.

[16] T. R. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human-Computer Studies, vol. 43, pp.907-928, 1993.

[17] B. Chandrasekaran, J.R. Josephson and V.R. Benjamins, "What are ontologies, and why do we need them?" *IEEE Intelligent Systems,* vol. 14, pp. 20-26, 1999.

[18] N. Guarino and L. Schneider, "Ontology-Driven Conceptual Modelling", in ER '02: Proceedings of the 21st International Conference on Conceptual Modeling, pp. 10, 2002.

[19] A. Gangemi, N. Guarino, C. Masolo and A. Oltramari, "Sweetening WORDNET with DOLCE", *AI Mag.,* vol. 24, pp. 13-24, 2003.

[20] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. and Miller, "Wordnet: an on-line lexical database", *International Journal of Lexicography,* vol. 3, pp. 235-244, 1990.

[21] A. Gómez-Pérez, M. Fernández-López and O. Corcho, *Ontological engineering :with examples from the areas of knowledge management, e-commerce and the semantic Web / Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho,* London; New York: Springer-Verlag, 2004.

[22] A. Zisman and J. Kramer, "Towards Interoperability in Heterogeneous Database Systems", Department of Computing, Imperial College, Tech. Rep. DOC 95/11, December 1995.

[23] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos and A. Halevy, "Learning to match ontologies on the Semantic Web", *The VLDB Journal,* vol. 12, pp. 303-319, 2003.

[24] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art", *Knowl.Eng.Rev.,* vol. 18, pp. 1-31, 2003.

[25] M. Uschold and M. Gruninger, "Ontologies and semantics for seamless connectivity", *SIGMOD Rec.,* vol. 33, pp. 58-64, 2004.

[26] R. Dhamankar, Y. Lee, A. Doan, A. Halevy and P. Domingos, "iMAP: discovering complex semantic matches between database schemas", in SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 383-394, 2004.

[27] P. Shvaiko and F. Giunchiglia, "Semantic Matching", Department of Information and Communication Technology, University of Trento, Tech. Rep. DIT-03-013, Sep 2003.

[28] G.W. Cottrell, *A connectianist approach to word sense disambiguation,* London: Pitman, 1989.

[29] G. Hirst, *Semantic interpretation and the resolution of ambiguity,* Cambridge Cambridgeshire; New York: Cambridge University Press, 1987.

[30] G. Antoniou and F. Van Harmelen, *A semantic Web primer,* Cambridge, Mass.: MIT Press, 2004.

[31] H. Stuckenschmidt and M. Uschold, "Representation of Semantic Mappings", in Semantic Interoperability and Integration 2005, pp. 53-58, 2005.

[32] N. F. Noy and A. Musen, "Evaluating Ontology-Mapping Tools: Requirements and Experience", Stanford University, 2002.

[33] 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.

[34] N. Silva and J. Rocha, "Complex semantic web ontology mapping", *Web Intelli.and Agent Sys.,* vol. 1, pp. 235-248, 2003.

[35] E. Agirre and G. Rigau, "Word sense disambiguation using conceptual density", in Proceedings of COLING'96, pp. 16-22, 1996.

[36] 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, pp. 171-180, 1993.

[37] E.M. Voorhees, "Query expansion using lexical-semantic relations", in SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 61-69, 1994.

[38] E. Agirre, O. Ansa, E. Hovy and D. Martinez, "Enriching very large ontologies using the WWW", in Procedings of the ECAI 2000 Workshop on Ontology Learning, pp. 27-33, 2000.

[39] G. Ramakrishnan, B.P. Prithviraj, A. Deepa, P. Bhattacharyya and S. Chakrabarti, "Soft Word Sense Disambiguation", in Proceedings of GWC 2004, pp. 291-298, 2004.

[40] C. Stokoe, M.P. Oakes and J. Tait, "Word sense disambiguation in information retrieval revisited", in SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 159-166, 2003.

[41] C. Stokoe and T. John, "Towards a Sense Based Document Representation for Internet Information Retrieval", in The Twelfth Text Retrieval Conference, pp. 791-795, 2003.

[42] H.P. Edmundson, "New Methods in Automatic Extracting", *J.ACM,* vol. 16, pp. 264-285, 1969.

[43] D. Riboni, "Feature Selection for Web Page Classification", In EURASIA-ICT 2002 Proceedings of the Workshop, 2002.

[44] P. Resnik, "Disambiguating Noun Groupings with Respect to Wordnet Senses", in Proceedings of the Third Workshop on Very Large Corpora, 1995, pp. 54-68.

[45] K.M. Hammouda and M.S. Kamel, "Document Similarity Using a Phrase Indexing Graph Model", *Knowl.Inf.Syst.,* vol. 6, pp. 710-727, 2004.

[46] C.K. Cheng, X. Pan and F. Kurfess, "Ontology-based Semantic Classification of Unstructured Documents", in Adaptive Multimedia Retrieval, pp. 120-132, 2003.

[47] M. Ehrig and Y. Sure, "Ontology Mapping - An Integrated Approach", in ESWS 2004: Proceedings of the First European Semantic Web Symposium, pp. 76-91, 2004.

[48] M. Ester, H. Kriegel and M. Schubert, "Web Site Mining : A new way to spot Competitors, Customers And Suppliers in the World Wide Web", in Proc. 8th ACM SIGKDD int. Conf. on Knowledge Discovery and Data Mining, Edmonton, CA (KDD'02), pp. 249-258, 2002.