Knowledge Representation for Web based Services in a Multi-cultural Environment

Hiranmay Ghosh, N. Rajarathnam
Tata Infotech Research Group
Ground Floor, Library Building
IIT Campus, New Delhi 110016
{hiranmay.ghosh | rajarathnam.n}@tatainfotech.com

Santanu Chaudhury
Department of Electrical Engineering
Indian Institute of Technology
New Delhi 110016
santanuc@ee.iitd.ernet.in

Abstract

Internet being a global resource, web based applications need to break the barriers of language and culture. The core of an intelligent web based application comprises an ontological description of the domain. A domain ontology needs a medium for expression, which usually consists of terminology borrowed from a natural language. Thus, a knowledge based application becomes susceptible to linguistic and cultural context. In this paper, we present a new knowledge representation technique that distinguishes between the abstract concepts in a domain and their expressions. It can associate expressions from different languages with the concepts in an ontology network. Non-textual symbols and media property specifications can also be used to express the concepts using this technique. The resulting ontology can thus be used in a multi-lingual and multi-cultural environment. An RDF based language is used as a vehicle for the knowledge representation scheme.

1 Introduction

In recent times, the Internet is fast evolving as a platform for active web based services. Examples of such applications include travel services, automated call centers and e-business systems. These applications need to process information stored in the network and being exchanged across the system. The processing is not limited to the interpretation and manipulation of the syntactic structure of the information, but also requires semantic understanding and decision making based on the semantics. Internet being a global resource, these web-based applications operate across national boundaries and need to address heterogeneity of languages and cultural contexts. The ontological description of a domain needed for semantic interpretation of data, should ideally capture the abstract concepts that are independent of language or culture. Unfortunately, the vehicle for expressing ontology is symbols (words or phrases) derived from a natural language, making ontology sensitive to linguistic contexts. Besides, the ontology often falls prey of cultural context because of the cultural affiliation of the domain expert. For example, the perception of "loud music" or "gaudy color" varies from culture to culture. Another significant shortcoming of existing ontologies is that they are totally dependent on textual descriptors and ignore the non-textual modes of expression. The non-textual modes of communication, such as images and natural sounds, often have a much richer expressive power than textual or verbal expressions and have a potential to cut across linguistic and cultural barriers.

In this paper, we propose a method for knowledge representation that integrates textual and non-textual modes of communication. We clearly distinguish between the concepts and the expressions. The concepts are abstract and universal. Expressions are means for communicating concepts. They are generally based on textual and non-textual symbols (e.g., words and icons) which varies in the context of language and culture. While the communication within a web based service needs to use expressions, the computational logic for an intelligent web service should use the underlying concepts. We also recognize that data in multimedia form is a powerful medium of expression in a cross-cultural environment. Our knowledge representation technique seamlessly integrate multimedia pattern specifica-
tion with textual expressions.

We have used a standard mark-up language, namely the Resource Description Framework (RDF) [18] for expressing the enhanced ontology in a machine readable form. The underlying principles are quite general and are applicable to other knowledge representation languages, such as DAML, as well.

The rest of the paper is organized as follows. Section 2 reviews the state of the art in knowledge representation on the semantic web. Section 3 explains the role of ontology in knowledge based applications. Section 4 introduces Semantic Network a knowledge representation technique. Section 5 provides an overview of knowledge representation languages on the web. The extensions to a domain ontology proposed by us to cater to a multi-lingual and multi-cultural environment is presented in section 6. An example implementation is discussed in section 7. Finally, we conclude in section 8.

2 Related work

The words and the phrases in a natural language are symbolic representations of real world concepts. Information systems have traditionally associated semantics with keywords to index and retrieve information. However, ambiguity of word meanings and variation of user vocabulary result in unsatisfactory performance of these information systems. An online lexical database, such as the WordNet [17], distinguishes the "word meanings" (the intended concepts) from the "word forms" (the utterances) in English and establishes several lexical and semantic relations between the word meanings. The database has been used in several knowledge-based applications that attempt to "interpret" a message containing some user request or other forms of information [2, 7]. Similar lexical databases have been developed in other languages also[1, 14]. The major drawback of such lexical databases is that they are confined to a single language.

There have been a few attempts to develop bilingual and multi-lingual ontologies [3, 12, 21]. These ontologies set equivalence across the terminology in a few European languages. These "general purpose" ontologies encode some semantic relations, namely is-a and is-part-of. However, they do not encode domain specific relations, e.g. "a monument is built with marble", that may exist between the concepts in a specific domain. Domain-specific ontologies [8, 19, 20] extend the concept of WordNet to include domain-specific relations. Again, these ontology descriptions are language specific and can cater to applications in that language only. The international community recommends development of equivalent ontologies in different languages for multi-lingual applications [13]. The cultural differences make development of "equivalent" ontologies difficult in many application domains. Further, this approach cannot be used when a message contains vocabulary from multiple languages, which happens to be a contemporary practice.

A significant shortcoming in all existing ontology definitions is that they are text-centric. They totally ignore potentially rich non-textual modes of expression. A more serious issue is that they use the same words, that are used in the expressions, to label the concepts. The practice causes potential confusion between the concepts and the expressions.

The intelligent applications on the web, popularly known as the "semantic web", requires a language to express domain ontology in a machine understandable format. The mark-up languages XML, RDF and DAML, have successively evolved to cater to this need [6, 11, 16]. The expressive powers of these languages have been progressively improved to express the complex relationships that exist in the domain. Most of the ontologies designed for web based applications are expressed either in RDF or DAML. We shall review the evolution of these languages in a later section.

3 The role of ontology in web based services

The domain ontology is a critical component of an intelligent internet application. It establishes the semantics of the information stored in the network or any message that flows between the users and the different components of an information system. Ontology can be formally defined as "the specification of conceptualization of a domain" [15]. It describes the entities and the relations that exists in an abstraction of the domain. It is interesting to note that an ontology is not the description of the domain, but that of an abstraction of the domain. The level of abstraction depends on the purpose of the description. The entities and the relations that are defined in an ontology is determined by the judgment of the knowledge engineer in the context of the application. For example, while capturing the ontology for the domain of automobiles for use of automobile engineers, a knowledge engineer might focus on the engineering parameters, such as the engine torque and the rpm; for an e-business applica-
tion, the focus may rather be on available colors, looks and comfort, fancy features and price. The use of ontology results in a standard vocabulary and its consistent use. However, an ontology is much more than a thesaurus. While a thesaurus captures a few linguistic relations with predefined semantics, an ontology defines many more domain specific relations.

It is interesting to note that some ontology is either implicitly or explicitly used in any knowledge based application. The use of an explicit ontology has several advantages. An explicit ontology can be encoded by a domain expert and can be reused in several application contexts. A declarative ontology description is easier to maintain and enhance than if it were embedded in an application. An explicit ontology can be more easily understood by the system designers and used by several system components (typically human or software agents). An ontological commitment is defined as an "agreement to use a common ontology" by several interacting agents.

| class-def | monument |
| class-def | tomb |
| subclass-of | monument |
| class-def | India |
| class-def | Delhi |
| property | capital_of, value | India |
| property | near, value | Delhi |
| class-def | Mughal_Dynasty |
| class-def | Shahjahan |
| property | belongs_to, value | Mughal_Dynasty |
| class-def | Tajmahal |
| instance-of | tomb |
| property | located_at, value | Agra |
| property | built_by, value | Shahjahan |

Figure 1. A tiny ontology segment

In order to appreciate the role of ontology in an intelligent web service, consider the tiny ontology segment depicted in figure 1. The notation used in defining the ontology is self-explanatory. A suitable reasoning algorithm can utilize the ontology to map a descriptive specification like "the famous tomb built by a Mughal king" to the Tajmahal and ascertain its location to be near the capital of India. This conclusion can be useful for an intelligent travel service to make necessary arrangements for a traveler willing to visit the monument.

4 Semantic net as a knowledge representation technique

The unstructured nature of a body of knowledge makes it difficult to map it to a well-defined schema in a relational database system. This has motivated research in several alternative representations for domain knowledge. A frame based representation allows the concepts to be organized in "slots" with "properties" interconnecting them. The properties are essentially binary relations linking two concepts in a domain. Thus, a frame based system results in a graph, where the nodes represent the concepts and the edges the binary relations between them. Several knowledge representation schemas [4, 17] use a underlying frame structure. A semantic network is a generalization of the frame structure, where the properties are treated as first-class objects. Thus, it is possible to express relations between the properties in a semantic network. The flexibility allows definition of new domain-specific properties and establishment of their semantics. The semantic network has been used to represent the domain knowledge in many application domains, e.g. [8, 19, 20]. Figure 2 represents the ontology in figure 1 as a semantic network. Each class defined in an ontology maps to a node in the knowledge graph. Each directed edge of the graph represents a relation between a pair of concepts. A property-value specification in the ontology is represented by a link. The edge is labeled with the property name and the destination leads to the value class. Note that the nodes in a knowledge graph represent different types of entities, e.g. classes, individuals and property values. Thus, it may be necessary to add additional labels with the nodes to distinguish them. The "subclass" relation is indeed a special case of property-value specification in a knowledge graph. The properties of the properties are not shown in the graph, but will be discussed in a later section.

5 Knowledge representation on the web

The ontology expressed as a knowledge graph needs a language as a vehicle for textual representation. HTML has been regarded as the lingua franca of the internet since its inception. Its mark-up mechanism proves to be a simple intuitive mechanism to establish semantics to the contents of a document. However, there are a limited number of tags in HTML, each conveying a specific structural
meaning. There has been several attempts to extend HTML with additional semantic mark-up tags for use in web based applications. The effort has culminated in development of XML, which allows definition of arbitrary tags in a document, whose meanings are defined by the application. Many authors [6] view XML to be an option to encode the web contents for semantic modeling. A major limitation of XML is that it does not allow representation of the relation between the different semantic entities. This results in difficulty of a common interpretation of the document contents. Though a DTD or a schema definition can establish some private ontology, the focus is primarily on the document structure.

RDF [18] uses XML as a metalanguage and provides a mechanism to build complex data models. The base element of RDF is a triple, comprising a subject, an object and a predicate as shown in figure 3(a). Each component of a triplet is viewed as a resource in RDF and can be identified by an URI. A resource can also be a text segment as a special case. Several RDF triplets make a directed resource graph (see figure 3(b)).

The structure of the RDF graph can be directly mapped to a knowledge graph shown in figure 2. The entities in a knowledge graph maps to the nodes of the RDF and the relations to its edges. A powerful feature of RDF is its uniformity of treatment for all elements, namely the concepts and the properties, using URIs. The properties are therefore treated as first class objects. Thus, the RDF have enough expressive power to represent a semantic network. Moreover, a triplet, or even a complete RDF description, can also be treated as a resource in RDF. Thus, it is possible to develop arbitrary complex data models using several levels of nesting. The expressive power of RDF in data modeling has prompted its use as a language for ontology representation [5, 11] for web based applications. A complex web based application that requires automatic discovery, selection, invocation, monitoring and integration of several primitive services require an ontology for utilization of these services. The DAML family of languages [9, 10] have been proposed to standardize the way to characterize the services. Both the languages are extensions of RDF and inherits its expressive power. All these languages are, however, merely tools to express ontology and should not be confused with the ontology representation itself. The power of an ontology is determined by the underlying schema (e.g. a semantic network) and not by the language used for expressing the schema.

6 Extension to knowledge graph - descriptors and media patterns

The description in the knowledge graph is sensitive to language and cultural context since the classes and the relations are labelled by textual symbols derived from a natural language. Thus, it may be difficult to map an expression in another natural
language to a set of entities on the ontology map. In order to make an ontology usable in multi-lingual context, we distinguish between the concepts and their descriptors. The concepts are abstract entities and are local to the knowledge base. They are denoted as nodes in the knowledge graph. Descriptors are used to communicate concepts between the users and the different modules of a knowledge based system. The descriptors can be symbolic entities, such as the words from a natural language or commonly used icons. We associate a set of descriptors with every concept node. These descriptors may be borrowed from different languages to facilitate multi-lingual applications. A concept is identified when any of the descriptors are found in an expression in context of a web application. For example the class definition for "tomb" in figure 1 is modified as in figure 4 to contain a few descriptors in English and some Indian languages. Any of the expressions "tomb", "grave" "samadhi" or "maqbara" will all lead to the same concept, labelled as tomb, on the ontology map. Note that the label is just a unique identifier for the concept and has no relation with the natural language. Though we have used the word "tomb" for easier human understanding, it could be anything like "xyz" or "concept0124".

```

class-def tomb
    subclass-of monument
    property descriptor, value {tomb, grave . . . }
    property descriptor, value {maqbara, samadhi . . . }

... class-def Tajmahal

... property sample_image, value http:// . . /Taj.jpg
```

Figure 4. Modified ontology for multi-lingual application

The ontology can be further enriched by encoding perceptual entities like image and audio properties and samples. For example, the last line in figure 4 associates a sample picture "Taj.jpg" with the concept Tajmahal. The automated travel service described in section 3 may validate its deduction with the user using the sample image. The sample picture may also be used to identify relevant non-textual documents using content based retrieval methods. The non-textual expressions for a concept naturally overcome linguistic and cultural barriers.

The concepts and the expressions, in general, form a many-to-many relation as illustrated in figure 5, making identification of a concept difficult from the expression in a message. One way to achieve concept extraction is to model the relation as a cause-effect relationship and use of some form of diagnostic reasoning. The underlying (indented) concepts are believed to cause the appearance of some specific terms or some specific media pattern specifications in an expression. The criterion of "best explanation" can be used to interpret the concepts from the observed keywords or media patterns in a message. It is also possible to associate some causal strength between the concepts and the descriptors in a closed domain. Any algorithm for diagnostic reasoning e.g. the criterion of tightest cover, probabilistic or qualitative belief networks, etc. can be used for deriving the intended concept(s) from an expression. The suitability of a method depends on the size and complexity of the domain and the requirements of the application.

The generality of RDF enables it to support this extended ontology. The schema definition and the RDF representation of a portion of the example is shown in figures 6 and 7. Note that the schema definition includes relations between properties. For example, the properties "subclass" and "instance" are both subproperties of "class" and inherits the later’s semantics.

7 An example implementation

We have implemented a prototype retrieval application in the tourism domain that retrieves relevant documents from the web using an ontology
server. The ontology server encodes the domain knowledge using the representation as described in the previous section. As an example, it is possible to retrieve documents on the Tajmahal with a query like “the maqbara (tomb) built by Shahjahan in memory of his wife”, and/or with sample images. The domain knowledge deduces the concept “Tajmahal” from the expression and uses textual and content based retrieval methods to retrieve relevant documents from the web.

To identify the objects for retrieval, we exploit the descriptors and the relations specified in the message encoding the user query. We have used a reasoning model based on Belief Networks, a detailed discussion of which is beyond the scope of this paper. The concepts identified from a message are abstract entities internal to the ontology server, and cannot be used for any practical communication. The textual and non-textual descriptors associated with the identified concepts (or, a subset of the same) need to be used for that purpose. Thus, the ontology server, in effect, maps the set of descriptors in an incoming message to another set of descriptors in the outgoing message. While the input set of descriptors relate to user vocabulary, the output set relates to the application domain. The "translation" achieved with the ontology server can span over multiple languages and across different textual and non-textual domain (or, vice-versa). For example, a textual description of the Tajmahal (as cited above) is translated into textual descriptors (keywords) as well as sample images in our application. While the former is used to search the Internet using available search engines, the latter is used for Content Based Retrieval.

The reasoning algorithm that is built over a knowledge representation[4] generally exploits the implicit semantics of the properties. Thus, definition of domain-specific relations result in domain specific semantics and hence domain-specific reasoning algorithms. We have exploited the properties of the relations to develop domain-independent reasoning algorithm, still exploiting domain-specific relations. Thus, the application can be easily ported to other knowledge domains.

8 Conclusion

In this paper, we have proposed a new knowledge representation scheme that can be used for the web based applications in a multi-lingual and multi-cultural environment. The ontological description of a domain has been extended to include descriptors from multiple natural languages and media feature specifications. We have argued that RDF(S) possesses enough expressive power to capture the semantics of a domain ontology and that it can be used as a vehicle for representing the complex rela-
tions that exist in a non-trivial domain. Our experiment with a prototype tourism services confirm the claim.

The proposed extensions to domain ontology should not be viewed in this application context alone. The extensions can be used in several web based applications involving users from different countries and cultural backgrounds and dealing with multimedia objects in the domain. Though we have used RDF to encode the domain knowledge, the principles are consistent with and can be used in conjunction with extensions of RDF, such as the DAML family of languages.

References