# ONTOLOGY BASED EXTRACTION AND INTEGRATION OF INFORMATION FROM UNSTRUCTURED DOCUMENTS

Naychi Lai Lai Thein, Khin Haymar Saw Hla, Ni Lar Thein
*University of Computer Studies (Yangon)*

Keywords: Semantic Web, Shared Ontology, Information Extraction, Shared Terminology, Web Ontology Language, Source Ontology, Local Ontology, Semantic Mapping.

Abstract: The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. One of the basic problems in the development of Semantic Web is information integration. Indeed, the web is composed of a variety of information sources, and in order to integrate information from such sources, their semantic integration and reconciliation is required. Also, web pages are formatted with HTML which is only a human readable format and the agents cannot understand their meaning. In this paper, we present an approach to extract information from unstructured documents (e.g. HTML) and are converted to standard format (XML) by using source ontology. Then, we translate XML output to local ontology. This paper also describes a key technology for mapping between ontologies to compute similarity measures to express complex relationships among concepts. In order to address this problem, we apply machine learning approach for semantic interoperability in the real, commercial and governmental world.

## 1 INTRODUCTION

Most of today's web content is easily understood by humans but difficult to understand by computers. A significant portion of the data on the web is in the form of HTML pages. Since content, navigational information and formatting have no clear separation in HTML, the conventional information retrieval systems have the additional task of dealing with noisy data when providing full text search.

However, the number of different information sources is growing significantly and therefore; the problem of managing heterogeneity is increasing. Heterogeneity can be classified into four categories: *structure, syntax, system and semantic*. These problems that have to be faced are due to the lack of common ontology, causing semantic differences between information sources.

## 2 RELATED WORK AND PROBLEM ISSUES

Integrating information from web sources starts by extracting the data from the web pages exported by the data sources. Although XML is supposed to reduce the need for this extraction, relatively few sources are currently available in XML, and legacy HTML sources will be around for years to come.

Because of the semantic heterogeneity among sources, merely extracting the data from web pages is often insufficient to support integration. The problem is that information might be organized in different ways with different vocabularies. A solution to the problem of semantic heterogeneity is to formally specify the meaning of the terminology of each system and define a translation between each system terminology and the shared terminology.

Once a system can extract information from the various sources and has a semantic description of these sources, the next challenge is to relate the data in the sources. So, to integrate data across sources, an integration system must be able to accurately determine which data in two different sources refer to the same entities. This method uses the mapping between the relations and attribute names among the schemas of the individual data sources to help determine the object mappings. In order to address the problem of semantic information integration, we apply the idea of ontologies as a tool for data integration.



Lai Lai Thein N., Haymar Saw Hla K. and Lar Thein N. (2005).
ONTOLOGY BASED EXTRACTION AND INTEGRATION OF INFORMATION FROM UNSTRUCTURED DOCUMENTS.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 457-460
DOI: 10.5220/0002555504570460

# 3 FRAMEWORK FOR SEMANTIC INFORMATION INTEGRATION

The integration method is based on a three-step approach. As described in Figure 1, the first aspect is tackled by wrappers that lift selected content of individual information sources to a common data model. The latter part is done by mediators that provide the glue. Ontologies simplify the job of mediators by defining an integrated semantic model that gives an explicit representation of the semantics of information components. So, ontology can be used as a common model for our purpose. Our intended use of ontology is to describe a data model and to give semantics to data stored in web pages, rather than knowledge (Staab, 2002).

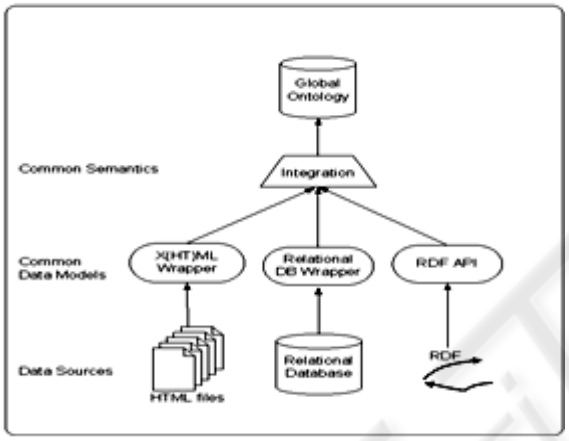


Figure 1: Conceptual Flow of Information Integration

Our hypothesis is that many web sites do not often change their organization of information. New information is published with a rather steady structure. Then, if we can recognize how information is organized, a precise data extraction can take place (Maedche, 2002).

# 4 UNSTRUCTURED DOCUMENTS TO STANDARD USABLE FORMAT

There is a tremendous amount of information available on the web but much of this information is not in a form that can be easily used by other application. So, we have developed the technology for rapidly building wrappers for accurately and reliably extracting data from unstructured sources.

Documents in HTML do not allow for direct querying. Therefore, we first convert the HTML document into XML by following their hierarchical structure. Possible errors are corrected using *HTML* TIDY. Here, we present an approach to extract information from unstructured documents based on a source ontology that describes a domain of interest. Starting with such ontology, we formulate rules to extract constants and context keywords from unstructured documents. Once a web page is transformed into XML using source ontology, portions of data can be easily used to create local ontology. We use OWL (Web Ontology Language) to build ontology (Embley, 1998).

As case studies to test these ideas for this paper, we consider the history of Myanmar Pagodas to extract information. This case is data rich and narrow in ontological breadth. This also includes information about name, year, features, enshrined things and location.

## 4.1 Data Extraction and Structuring from Unstructured Documents

In this section, we present the framework to extract and structure the data in an unstructured document. As shown in Figure 2, there are three processes in this framework: an ontology parser, a constant/keyword recognizer and a structured text generator. The input is source ontology and a set of unstructured documents. The output is XML structured format filtered with respect to the source ontology and local ontology which is converted from XML output. In this paper, the only step that requires significant human intervention is the initial creation of source ontology for pagoda. However, once such an ontology is written, it can be applied to unstructured documents from a wide variety of sources, as long as these documents correspond to the given source domain (Embley, 1998).

In order to extract information from the HTML code, we need to understand how the data is presented within the HTML code and analyzes the page's tag structure. Given the inputs, parser module (PM) starts by fetching HTML code of the web page to be parsed. Then, it analyzes the tag structure and tries to determine how the data is presented in this code.

After determining the tag structure of the page, PM finds the starting point of data and proceeds to divide the data into fields and records using the pre-defined record separator tags. After invoking the parser, the constant/keyword recognizer uses the source ontology to recognize each regular expression. The structured text generator uses the object/relationship/constraints lists to generate XML

format. The ontology generator uses this XML output to transform local ontology using each regular expression.

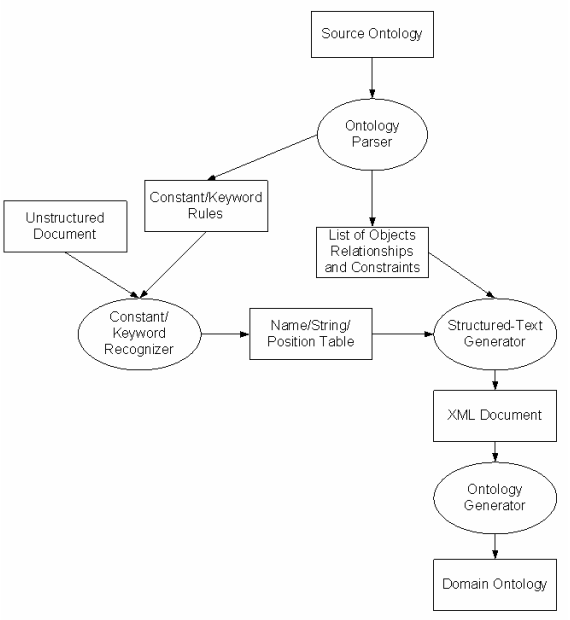


Figure 2: Framework for Ontology Based Information Extraction and Structuring

```
  <owl:Class rdf:ID="Pagoda">
 <owl:Restriction>
       <owl: onProperty
       rdf:resource=”#hasName”/>
       <owl:allValuesFrom
       rdf:resource=”#Name”/>
       </owl:Restriction>
       <owl:Restriction>
        <owl: onProperty
        rdf:resource="#hasKing"/>
        <owl:allValuesFrom
        rdf:resource="#King"/>
       .
       .
     </rdfs:subClassOf>
</owl:Class>
```

Figure 3: OWL Source Ontology for Pagoda

While this approach provides a solution for the problem of extracting information from weakly structured resources, the problem of integrating information from different sources remains largely unsolved (Cui, 2001).

```
<Pagoda>
<Name>Shwezigon</Name>
<StartKing>Anawrahta </StartKing>
<FinishKing>Kyansittha</FinishKing>
<City>Bagan</City>
<Division>Mandalay</Division>
</Pagoda>
```

Figure 5: Resulting Sample XML File.

```
<owl:Class rdf:ID="Shwezigon">
<rdfs:subClassOf
rdf:resource="#Pagoda"/>
<hasStartKing rdf:
resource="Anawrahta"/>
<hasFinishKing rdf:
resource="Kyansittha"/>
<hasCity rdf: resource="Bagan"/>
<hasState rdf: resource="Mandalay"/>
</owl:Class>
```

Figure 6: Sample Output Local Ontology

# 5 ONTOLOGY BASED SEMANTIC INFORMATION INTEGRATION

Information integration is concerned with unifying data sharing some common semantics but is originated from unrelated sources. There are a lot of advantages in the use of ontologies for information integration. In general, three different directions can be identified: *single ontology approach*, *multiple ontologies approach* and *hybrid approach*.
As our approach is based on the hybrid ontology, it has two main advantages:
(1) New information sources can be added without the need of modification.
(2) Shared vocabulary and the mappings among the local ontologies make them be comparables.

In this system, shared ontologies provide a vocabulary in order to specify the semantics of information in different sources. Formally, we define a shared terminology as a set of words and a partial function over pairs of words [Stuckenschmidt, 2002].

## 5.1 Ontologies Alignment Using Shared Terminology

It has been argued that semantic heterogeneity can be resolved by transforming information from one context into another.

A conceptual model of the context of each information source builds a basis for integration on the semantic level. In this process, we take the information about the context of the source providing a new context description for that entity within the new information source. Here, we focus on context transformation by classification [8].

## 5.2 Mapping between Ontologies

In this system, we apply machine learning techniques to semi-automatically create semantic mappings. Since taxonomies are central components of ontologies, we focus on finding correspondences among the taxonomies of two given ontologies: for each concept node in one taxonomy, find the most similar concept node in the other taxonomy. The first issue we address is the meaning of similarity between two concepts. In our approach similarity measure is based on the joint probability distribution.

To match concepts between two taxonomies, we need a measure of similarity. First, we would like the similarity measures to be well-defined. Second, we want the similarity measures to correspond to our intuitive notions of similarity.

Many practical similarity measures can be defined based on the joint distribution of the concepts involved. A possible definition for the exact similarity measure is

$$Jaccard - sim(A,B) = \frac{P(A \cap B) / P(A \cup B)}{P(A,B) + P(A,\overline{B}) + P(\overline{A},B)}$$

This similarity measure is known as the Jaccard coefficient. It takes the lowest value 0 when A and B are disjoint, and the highest value 1 when A and B are the same concept [Doan, 2002].

# 6 CONCLUSIONS

In this paper, we dealt with the problem of information integration from different sources. Integrating information from web sources starts by extracting the data from the Web pages exported by the data sources. So, we have proposed a framework for extracting reliable data and to convert standard form.

By using shared ontology, we also address terminological semantic heterogeneity in semantic integration. With the proliferation of data sharing applications that involve multiple ontologies, the development of automated techniques for ontology matching will be crucial to their success.

# REFERENCES


Bayrak C., Kolukısaoğlu H., *2003*. Data Extraction from Repositories on the Web: A Semi-automatic Approach, Computer Science Department, University of Arkansas at Little Rock, Little Rock, AR, U.S.A. , *SEPTEMBER, Vol. 7, No. 4, pp. 13-23.*

Cui Z., Jones D., Brien P. O, 2001. Issues in Ontology-based Information Integration, Intelligent Business Systems Research Group Intelligent Systems Lab.

Doan A., Madhavan J., Domingos P., Halevy A., 2002. Learning to Map between Ontologies on the Semantic Web. *In Proceedings of the World-Wide Web Conference (WWW-2002), pages 662-673, ACM Press.*

Embley D. W., Campbell D. M., Smith R. D., and Liddle S. W., 1998. Ontology-based extraction and structuring of information from data-rich unstructured documents. I*n International Conference on Information and Knowledge Management (CIKM).*

Gruber T.R., 2003. A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, 199–220.

Hendler J., Lee T.B., Miller E., 2002. Integrating Application on the Semantic Web, Journal of the Institute of Electrical Engineers of Japan, Vol. 122 (10).

Maedche A., 2002. Tying Up Information Integration and Web Site Management by Ontologies, IEEE Data Engineering Bulletin.

Stuckenschmidt H., 2002. Information Sharing on the Semantic Web, AI Department, Vrije University, Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands.

Soderland S., 1998. Learning information extraction rules for semi-structured and free text, www.cs.washington.edu/homes/soderlan/ WHISK.

Staab S., Maedche A., 2001. Comparing Ontologies Similarity Measures and a Comparison Study, Internal Report No. 408.

Staab S., 2002. The Semantic Web-New Ways to Present and Integrate Information, Institute of Applied Informatics and Formal Description Methods (AIFB), University of Darlsruhe.

Wache H., Vögele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H., Hübner S., 2001. Ontology-Based Integration of Information: A Survey of Existing Approaches.