# OntoGenie: Extracting Ontology Instances from WWW

Chintan Patel, Kaustubh Supekar, and Yugyung Lee

School of Computing and Engineering
University of Missouri-Kansas City
{copdk4, kss2r6, leeyu}@umkc.edu

**Abstract.** Web has become a tremendously huge information source on planet. However, the information is not machine perishable. Standardized Ontological representation of knowledge solves the problem as proposed by Semantic Web. One of the major challenges is to convert the information present on current Web into Ontologies for Semantic Web. We have developed a solution, OntoGenie, that parses the Web pages to create knowledge instances for a given Ontology using WordNet as a bridge, mapping between the Ontologies and the Web page terms. OntoGenie was tested over Ontologies available on the Semantic Web and some motivating results were obtained.

## 1 Introduction

Semantically enriched Web would allow leveraging intelligent applications such as semantic search, Semantic Web services and Semantic Grid. The knowledge in Semantic Web is encoded in webized way, as simple directed graphs [3]. Thus, Ontologies, representation of domain knowledge in Semantic Web, provide the explicit formalization and specification of the concepts and their corresponding relationships [2]. It should be noted that Ontologies have associated specific instances for the corresponding concepts. These instances contain the actual data that are being queried in knowledge based applications. Ontologies are largely developed manually by domain expert, filling in the instance data manually is an arduous task. It is infeasible to manually construct all instances corresponding to a concept defined in an Ontology.

In this paper, we focus on creating Ontology instances that can be automatically extracted from unstructured data on Web including plain text and HTML. Also, to accelerate the nurturing and growth of Semantic Web, there is a pressing need to develop tools that would provide smooth transition from current Web to Semantic Web. We have developed a tool, OntoGenie, that uses WordNet[1] to convert *unstructured data* from Web to *structured knowledge* for Semantic Web. The tool was developed as a part of ongoing BEE-SMART (A Natural Language Interface to Semantic Web) project at University of Missouri[2]. The architecture of the tool and the results obtained are discussed.

[1] http://www.cogsci.princeton.edu/ wn/
[2] http://sice527.ddns.umkc.edu/BeeSmart/

## 2 OntoGenie Functionality: What's your wish master?

The OntoGenie is a semi-automatic tool that takes as input domain ontologies and unstructured data from Web (plain text or HTML), and generates Ontology Instances (OI) for the given data. The tool uses the linguistic ontology, WordNet, as a bridge between domain ontologies and Web data.

**Step 1: Map the concepts in a domain ontology into WordNet**- Retrieve a concept $C_d$ from domain ontology $O_d$ and map it into a concept $C_w$ in the WordNet ontology $O_w$. The mapping is performed by canonizing the English terms defining the Concepts ($C_d$ and $C_w$). One important issue in this regard is that many terms in WordNet may map into a same concept from $O_d$. For example, the concept *University* in WordNet has more than one senses such as an 'educational institution' or a 'group of persons associated by some common tie'.

**Step 2: Capture the terms occurring in Web pages**- OntoGenie utilized the search service (Google Web service[3]) and the directory service (dmoz directory[4]) to retrieve Web pages for a particular domain. The web pages are parsed word by word, each word $W_i$ is canonized and compared with the $C_w$ present in the WordNet. Interestingly, we can visualize a connection among the Ontology Concepts ($C_d$), the WordNet Concepts($C_w$) and the Web page terms ($W_i$). Consider the Ontology in Figure 1: the concept *Country* ($C_d$) in the domain ontology $O_d$ is mapped to similar concept *Country* ($C_w$) in the WordNet ontology $O_w$ in Step 1 and then during Step 2, the Web term *USA* ($W_i$) is mapped into a hyponym of *Country* $C_w$ which has already being mapped to *Country* $C_d$.

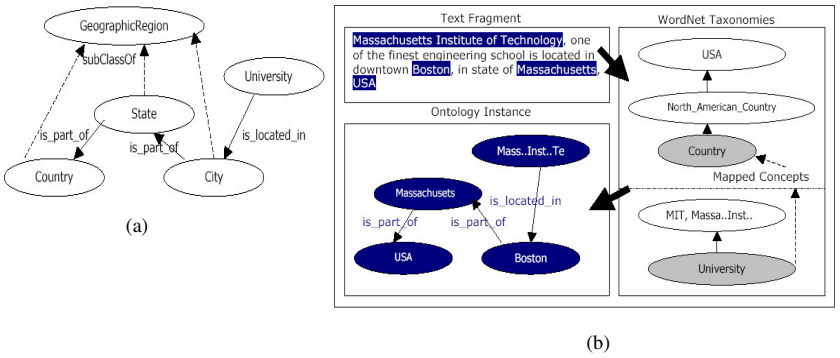


**Fig. 1.** (a) University Ontology Excerpt (b) Flow of OntoGenie Algorithm

**Step 3: Discover relationships**- Once the mappings are accomplished for a Web page, we discover the relationship that holds between the instances of the concepts extracted. Conventionally, the task of discovering relationships was

[3] http://www.google.com/apis/
[4] http://dmoz.org/

done via morphologically determining the verbs and the relationships to noun [1]. The approach works for simple *toy* cases, but fails practically in real world cases, dealing with large amount of ontological instances. We propose to use a simple approach using *principle of locality* ($\delta$), the idea is to flexibly assume a set of concepts discovered in predetermined locus around the concepts to be related. To better understand the idea, consider the Ontology as a graph, with Concepts represented as nodes and the Relationships as links. The distance between Concepts in the set can be defined as number of links encountered traversing the links between the Concepts (we assume the shortest path).

Consider an example, as described in Figure 1b, an instance of University $MIT$ and an instance of $Country$, we can assume a relationship to hold between them. It should be noted however that if the instances of intermediate nodes are unknown (e.g., *State* in this case), we still consider them as blank nodes. Such blank nodes can be filled on while scanning other Web pages for the given domain. The purpose of incorporating the principle of locality is to increase the recall by discovering largely disconnected knowledge instances and then linking them by information discovered from other pages.

## 3 OntoGenie Implementation and Results

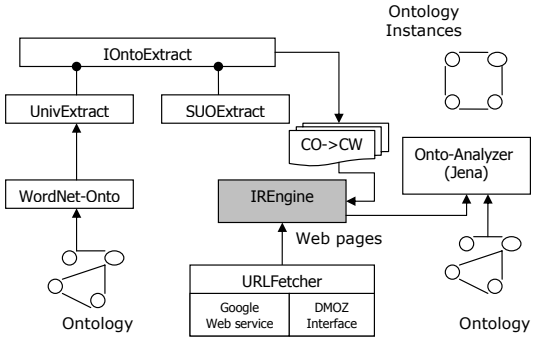


**Fig. 2.** OntoGenie Implementation Framework

The architecture of OntoGenie has been designed to exploit the functionality provided by the existing available tools. Figure 2 shows the implementation details for the OntoGenie framework. The OntoGenie implementation interacts with the Java WordNet and Jena APIs for Ontological and Web data parsing, computing locality-based distance between concepts, and creating Ontology instances. To disambiguate the Concept mappings to WordNet, as mentioned in Step 1, we have developed a graphical user interface for a domain expert to select the right sense for the automatically discovered mappings. We used KAON as our backend data store for crawled Ontologies. To interface Google Web service,

we used Java Web Services Developer Pack[5] (JWSDP). One of the noteworthy idea being incorporated is providing an abstract Interface *IOntoExtract*, wherein we can develop different plugins to test variety of mappings. For example, SUO[6] was mapped to WordNet within the OntoGenie framework. Similarly, with the component URLFetcher, one can add variety of interfaces to retrieve web pages (Google web services and DMOZ URL extractor were used in OntoGenie)

We tested the OntoGenie framework with University Ontology[7] and extracted the university related Web pages[8]. The OntoGenie has successfully discovered Knowledge Instances from the Web. Table 1 shows one of the RDF instance being discovered for the University Ontology. The excerpts says that **Librarian** is an instance of the concept *Person* and is the member of an *Organization* whose instance is **Library**.

| |
|---|
| <rdf:Description about= "http://tempuri.com/15univs.daml#Librarian"> |
| <rdf:type resource= "UNIVURI#Person"/> |
| <km:member rdf:resource= "http://tempuri.com/km#Library"/> |
| </rdf:Description> |
| <rdf:Description about= "http://tempuri.com/km#Library"> |
| <rdf:type resource= "UNIVURI#Organization"/> |
| </rdf:Description> |
| UNIVURI = http://www.cs.umd.edu/projects/plus/DAML/onts/cs1.0.daml |

**Table 1.** Experimental Results

## 4 OntoGenie Conclusions: Getting back into Lamp!

We presented a simple, practical and implemented framework, OntoGenie that solves the highly critical and important problem of discovering Knowledge instances from Web. OntoGenie is based on creating mappings from Ontology to Web page terms using WordNet as a effective bridge. We showed implementation details and a glimpse of the results obtained.

## References

1. M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, Learning to Extract Symbolic Knowledge from the World Wide Web, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98).
2. T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, Proceedings of International Workshop on Formal Ontology, Padova, Italy, 1993.
3. Tim Berners Lee, Semantic Web Roadmap, http://www.w3.org/DesignIssues/Semantic.html

[5] http://java.sun.com/webservices/webservicespack.html
[6] http://suo.ieee.org/
[7] http://www.cs.umd.edu/projects/plus/DAML/onts/cs1.0.daml
[8] http://www.mit.edu:8001/people/cdemello/univ-full.html