RIT Department Of Computer Science

MSc Project Pre-Proposal:

Prototype - Automation of Ontology Creation

Varun Goyal
April 26, 2012
Chair: Dr. Xumin Liu

A. Terminologies:
1. Interfaces or data interfaces: It is a source of attributes required to query the web-service at the interface of source web sites.
2. Data pages: search results of such web sites.
3. Ontology
4. Semantics
5. Visible Web / Surface Web: The data which is available at the surface of internet, i.e. which can be viewed or extracted using web UIs.
6. Invisible Web / Deep Web: All the data available on the website server, i.e. the data which is not accessible via the surface web.

B. Problem Description & Motivation
   Semantic search is a search of the content present on the world wide web based upon the meaning of the search query and the meaning of the content being searched into. By doing this, Semantic search seeks to improve search accuracy over the traditional keyword based search implemented by most major search engines.
   However, semantic search is greatly dependable upon the whole WWW to implement ontology. An ontology formally represents knowledge as a set of concepts within a domain and the relationship between those concepts. So until we know about all the concepts defined on a website and their relationships we cannot implement a semantic search on the website.
   Thus, we realize that creating ontology is a problem here. However, this problem can be addressed with the help of domain expert and few developers. But then again this might be expensive with no increase in traffic to the website in near future (till the time every-other website creates and ontology or builds upon existing ones and the search giants implement semantic search). Thus, we come to the conclusion, that we need a tool for automatic acquisition of domain ontologies.

C. Goal
   Goal of this project is to create a prototype tool for automation in creation of ontology for a website.

D. Implementation
   This project will use the concept of DeepMiner[1]. The DeepMiner system learns domain ontologies from the source web sites by creating and training the label classifiers (C^l) and the instance classifiers (C^i), which set a relationship between the concepts and their instances. Once this relationship is understood DeepMiner constructs ontology. DeepMiner would reiterate on these steps to on next set of C^l and C^i to improve upon the developed ontology.
   However, the problem with DeepMiner is that it need a lot of data input i.e. Interfaces and data pages. DeepMiner only uses only visible web to acquire this

data and give good results. We propose to include the data from invisible web in the deep miner to improve its efficiency and accuracy.

E. References

1. W. Wu, A. Doan, C. Yu, and W. Meng. Bootstrapping domain ontology for Semantic Web services from source web sites. In Proceedings of the VLDB-05 Workshop on Technologies for E-Services, 2005.
2. A. Maedche and S. Staab, Ontology learning for the Semantic Web, IEEE Intelligent Systems: Special Issue on the Semantic Web 16 (2001) 72–79.
3. Zhou, L. (2007), Ontology Learning: State of the Art and Open Issues, Information Technology and Management, 8(3), 241-252.
4. G. Bisson, C. Nedellec and L. Canamero, Designing clustering methods for ontology building – The Mo'K workbench, presented at ECAI Ontology Learning Workshop, Seattle, WA, 2000.
5. Decker, S.; Melnik, S.; van Harmelen, F.; Fensel, D.; Klein, M.; Broekstra, J.; Erdmann, M.; Horrocks, I.; , "The Semantic Web: the roles of XML and RDF," Internet Computing, IEEE, vol.4, no.5, pp.63-73, Sep/Oct 2000 doi: 10.1109/4236.877487URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=877487&isnumber=18994