An Architecture to Support Web-Based Information Extraction Using Domain Ontologies

Vijayan Sugumaran[1,2] and Farid Meziane[3]

[1] Department of Decision and Information Sciences
School of Business Administration
Oakland University, Rochester, MI 48309, USA
[2] Department of Service Systems Management and Engineering
Sogang University
1 Shinsoo-Dong, Mapo-Gu, Seoul 121-742, South Korea
[3] School of Computing Science and Engineering, Newton Building
The University of Salford, Salford M5 4WT, UK
sugumara@oakland.edu, f.meziane@salford.ac.uk

Abstract. The web is the world's most valuable information resource. However, a wide gap has emerged between the information available for software applications vis-à-vis human consumption. In response to this problem, new research initiatives have focused on extracting information available on the web with machine-processable semantics. Ontologies play a large role in information extraction, particularly in the context of semantic web, and applications should be able to find appropriate ontologies on the fly. However, existing tools do not adequately support information extraction and ontology selection. This research-in-progress paper presents the architecture for an information extraction system which relies on domain ontologies and lexical resources. We also provide an approach for easy identification of appropriate ontologies for a particular task.

Keywords: Information Extraction, Semantic Web, Ontology, Ontology Selection.

1 Introduction and Motivation

1.1 Introduction

The development of the web has seen an exponential expansion of information available to users. It contains huge collections of documents that span different domains, languages and levels. Hence, turning the web into the primary source of information for many users and its associated information services will undeniably be the richest source of information. The web serves a huge, widely distributed, and diverse community of users.

The diversity and density of the web has created a significant data extraction problem as its present structure makes it difficult to make use of that information in a systematic way [4]. The content of the web is created and managed by many individuals and organizations and adhere to various standards and formats, which provide

R. Sharman, T.S. Raghu, and H.R. Rao (Eds.): WEB 2009, LNBIP 52, pp. 1–9, 2010.

undeniably convenient and intuitive ways of representing and using information to humans. However, this imposes a real challenge to machines and the automatic extraction of relevant knowledge and information. Hence, this limits the benefits the web could bring to some communities and is frustrating when searching for incomplete, imprecise and ambiguous information. Hence, knowledge discovery from web resources is becoming a priority for many researchers and industries. The most common way currently in use for searching and retrieving information from the web is based on keywords search or similarity based search using one or more search engines, and then in order to extract relevant information, the user has to browse the large number of returned URLs. Moreover, these approaches can encounter many major difficulties including synonymy and polysemy problems.

On the other hand, the development of the semantic web aims at reducing these difficulties by stressing more on the semantics associated with the web content than its keyword based search. There are various definitions and developments of the semantic web which is becoming the main research interest of many communities that include artificial intelligence, and information retrieval and extraction communities. Ontologies are seen as the backbone of the semantic web research. In this paper, we present an architecture to support efficient information extraction from web resources. This architecture is based on domain specific ontologies and is aimed to be generic enough to be applied to various domains. In addition, we present an approach for the selection of the best ontology for specific domain searches.

1.2 Research Context and Limitations

There are two approaches to information extraction, the knowledge engineering approach and the automated training approach [7]. In the knowledge engineering approach, a knowledge engineer will develop some hand crafted rules to use on a specific corpus. Being an interactive process, the rules will be improved until they yield a satisfactory result. Although intuitive, relatively easy to develop and not requiring high computational resources, it is hard to generate rules that are general enough to be used on unseen documents and applied to different domains [15]. On the other hand, extensive corpora annotation is used in the automated training approaches. The Information Extraction (IE) system has to learn extraction patterns from the annotated corpora. This approach requires human intervention for the annotation of large corpora and is known to require extensive computational resources. In our research, we adopt a knowledge engineering approach for two main reasons.

1. The model we are developing is not a single architecture that will fit all domains, but a generic one whose instances can be applied for specific domains. Hence the rules will be specific to a single domain and different domains will require different sets of rules.
2. We attempt to minimize human intervention and particularly the manual annotation of large corpora. It is just not feasible to annotate large corpora for systems that attempt to extract knowledge from the web.

The proposed architecture does not support the development of ontologies but attempts to use existing ones. Indeed, experts suggest that in the semantic web, it is unlikely that we will have a large number of comprehensive, high-quality ontologies but rather many smaller, domain-specific ontologies [6,13]. Based on these assumptions, our system attempts to make use of this plethora of ontologies and incorporate in the architecture a component that selects the most suitable ontology for the application to be considered.

The rest of the paper is organized as follows. Section 2 summarizes the benefits and vision of the semantic web and in section 3 we introduce the concept of ontologies and their use in information retrieval. We describe our proposed system in section 4 and the ontology selection process in section 5. Section 6 concludes the paper.

2 Semantic Web

The web contains a huge collection of documents, which are read, understood, and processed mainly by humans and its current structure is not machine friendly. The amount of electronic information keeps on growing and the internet users are facing the information overload paradox and existing tools and techniques do not provide adequate relief from this problem. Moreover, they are not able to exploit the semantic content of these information sources, so it can be hard at times to find out meaningful relationships between different pieces of information.

These and many other similar problems are the bottlenecks for the future growth and utilization of the web, and in order to overcome them, web contents should be processed by computers if we want to achieve the vision of the semantic web which aims at providing an information enriched with machine processable semantics. This will allow various intelligent services to understand the information and to perform knowledge level information transformation, search, retrieval and extraction [2].

Ontologies are no doubt the most important form of knowledge representation currently in use for the semantic web. In order to overcome the problems caused by present search and retrieval techniques to access information, ontologies are providing ways to retrieve and extract information based on the actual content of a page and help navigate the information space based on semantic concepts. Tools like ontologies facilitate access to and description of the content of documents and are an important step towards offering efficient resource discovery on the web. They can be generic, for example WordNet (http://wordnet.princeton.edu/), or can be domain dependant covering the concepts related to a particular domain. The proposed architecture will be using domain ontologies, which will provide concepts related to a domain of interest, in order to disambiguate word sense, automatic query expansion and for efficient information retrieval and extraction.

3 Role of Ontologies in Information Extraction

The use of ontologies to overcome the limitations of keyword-based search has been put forward as one of the motivations of the semantic web since its emergence in the late 90's [14]. Ontology driven information extraction methods extract relatively shallow information from a very large corpus of documents, instead of performing more exhaustive (and expensive) processing of a small set of documents [9]. Ontologies can be used for both knowledge engineering and automated training approaches [15]. In the knowledge engineering approach, ontologies can be used in the development of rules as well as automated training for the annotation of corpora.

The advantages of using ontology driven information extraction include:

- Driving the entire information extraction process directly from the ontology presents a very natural path for exploiting all kinds of ontological data [9].
- Search-based system using ontologies, allow users to consider a much larger set of documents than could be handled via individual, document-driven processing for example [9].
- Ontology-driven information extraction can be easily focused on the desired results rather than processing all the content from some documents [9].
- Domain knowledge in form of an ontology makes it possible to develop portable Information Extraction system [15].

4 The Proposed Architecture

We propose a generic ontology based information extraction system that constitutes a suitable basis for building an effective solution to extracting unstructured information from the web by providing an extensible architecture and will provide fast and accurate selective access to this information; performing selective dissemination of relevant documents depending on filtering criteria.

The architecture aims at providing a set of integrated software components for accessing heterogeneous data sources and extracts the required information based on domain dependent ontologies. However, ontology selection will be based on the search criteria and the information users are requesting. The architecture of our system is shown in Fig.1. It is composed of 5 modules, which are briefly described below.

Query Processing Module: The query component handles the user's query which is expected to be in a free format probably in natural language or some known keywords. A suitable interface will be developed and this is the main input to the system and if necessary, the query will be pre-processed to remove stop words, stemming etc.

Ontology Selection Module: This component has two tasks. The first is to identify the domain of the ontology if this is not known to the system or provided by the user. This task will rely mainly on the components of the user query and will attempt to link the query to a known domain. The second task is the selection of the best ontology among a set of ontologies. This task is described in details in section 5 of this paper.

Searching Module: This component performs the usual web search based on users' requirements and the ontology. Query expansion and refinement can be performed at this stage. The output of this component would be a set of documents or websites containing the requested information.

Information Extraction Module: Once the documents are retrieved, this module attempts to extract the required information. It uses a variety of natural language processing techniques, text summarization, etc. As mentioned earlier, in its early implementation, our system will adopt a rule based approach for information retrieval.

User Interaction and Presentation Module: This module's task is to gather user input and present the results in a layout suitable to the user needs such as tables, templates etc. The module can be extended to personalize the output taking into account user profiles.

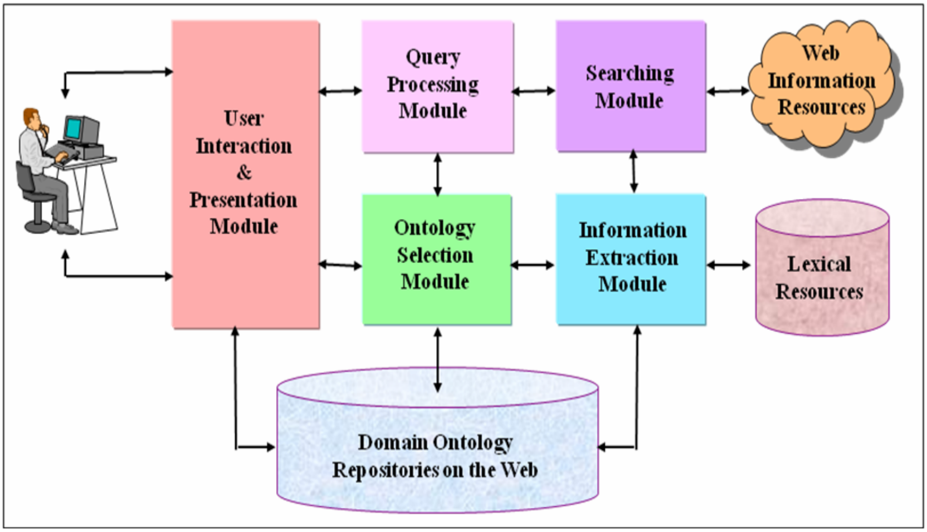
User Interaction & Presentation Module
Query Processing Module
Searching Module
Web Information Resources
Ontology Selection Module
Information Extraction Module
Lexical Resources
Domain Ontology Repositories on the Web

Fig. 1. Proposed Information Extraction System Architecture

5 Ontology Searching and Selection Methodology

Some ontology search engines have been discussed in the literature. For example, SWOOGLE [5] uses traditional Information Retrieval (IR) techniques to retrieve ontologies using keywords. It indexes ontologies primarily designed using OWL language which supplies a set of metadata that could be used for identification of semantic web documents. Since the terms are usually compounded into Uniform Resource Index (URI) reference terms, this approach is not very effective in identifying relevant ontologies. Similarly, link analysis is used to rank ontologies with respect to queries in the OntoSearch system [16] and in the OntoKhoj system [10]. These approaches are limited in focus and do not scale up. Also, they place the burden on the user to determine the appropriateness of the ontologies for the task at hand [1].

Another issue in ontology selection is the difficulty in assessing the quality of ontologies. As the number of available ontologies increases, the evaluation of these ontologies becomes more difficult. A few ontology evaluation approaches have been discussed in the literature. For example, Porzel and Malaka [11] propose an evaluation approach that primarily focuses on the syntactic aspects of ontologies. However, it does not take into account the semantic aspects and user context. The OntoMetric approach [8] establishes a set of processes to evaluate ontologies. It uses predefined characteristics such as ontology language, the methodology used to develop the ontology, and the software environment used. While this is an interesting approach, it greatly depends on the user's familiarity with the domain and is applicable for only small ontologies. The proposed approach attempts to minimize the cognitive burden on the end user who may not be well versed in ontology engineering.

One of the goals of this research is to develop a framework using a novel approach for assessing and selecting the most appropriate ontology for a particular task on the semantic web and implement it in a proof-of-concept prototype system. This system will make use of external knowledge sources available on the web, as well as an internal knowledge repository that contains various ontology evaluation metrics for ranking ontologies and selecting the most appropriate ontology based on task characteristics.

The proposed framework is being implemented using a suitable open source CASE tool and other open source utilities such as inference engines, ontology editors, dictionaries and natural language processors for supporting the various functionalities. The resulting system is expected to enhance the efficiency of systems analysts in developing different types of semantic web applications that require the use of ontologies. The utility and effectiveness of the system will be assessed by comparing it to other ontology search and selection tools such as SWOOGLE. It is anticipated that the users and applications can find their web resources, namely ontologies, on the web more effectively compared to the existing tools.

Our approach for the selection of ontologies on the semantic web consists of the following three steps (Fig. 2): 1) identification of the initial ontology set, 2) ontology evaluation and 3) ontology selection. Users can start the search for ontologies using a natural language query and specifying the domain. The system will initiate the search for the available ontologies in the knowledge-base (KB) repository and identify ontologies within that domain. Then, the ontologies are evaluated and ranked using appropriate metrics. With user feedback and taking into account the task characteristics, the system will recommend the best ontologies to use. The individual steps are briefly discussed below.

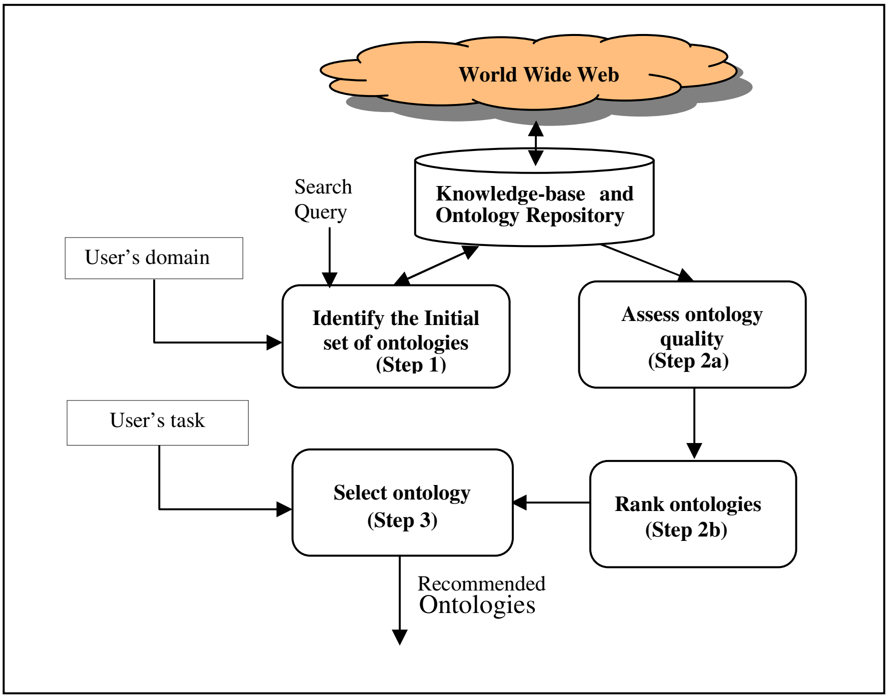
World Wide Web
Search Query
Knowledge-base and Ontology Repository
User's domain
Identify the Initial set of ontologies (Step 1)
Assess ontology quality (Step 2a)
User's task
Select ontology (Step 3)
Rank ontologies (Step 2b)
Recommended Ontologies

Fig. 2. Ontology Selection Approach

Step 1: Initial set of ontologies

There are a large number of ontologies on the web and many ontologies are being created every day. However, finding and accessing these ontologies has been difficult. Hence, we will design a knowledge-base of ontology repositories to identify potential candidate ontologies. Our system will provide an interface for the user to access a particular ontology on the web directly or use our knowledge-based ontology repository as a starting point. For example, ResearchCyc (http://research.cyc.com) is a well-known ontology library with more than 300,000 concepts, 3,000,000 assertions (facts and rules), and 26,000 relations, that interrelate, constrain, and define the concepts. However, it doesn't provide a flexible user interface for searching. Similarly, DARPA Agent Markup Language (DAML) ontology library (http://www.daml.org/ontologies) is the well-known ontology library, which contains 282 publicly available ontologies. Our system will provide an interface to such libraries and the user can search these libraries by providing keywords and domain information. For example, assume that the user submits a query related to air travel. Using the ontology repository, the system identifies travel domain ontologies as a starting point. At the end of this step, the system provides a set of ontologies to consider.

Step 2: Ontology evaluation

The most important step in ontology selection is ontology evaluation. The ontology evaluation step consists of two sub-steps, namely, ontology quality assessment (step 2a) and ontology ranking (step 2b). The quality assessment activity requires a suite of metrics that can be used to assess the quality of ontologies. Burton-Jones et al. [3] have proposed a metric suite for ontology auditing. This metric suite comprises of ten metrics derived from the theory of semiotics that assess the syntactic, semantic, pragmatic and social quality of an ontology. We examine the suitability of the individual metrics in the context of semantic web and adopt only those metrics suitable for SW and operationalize them. We develop an ontology auditing tool to assess the quality of the initial set of ontologies. For each ontology, a numerical score for each metric is computed. The ontology ranking step involves computing the overall quality score for the ontology depending upon the relative importance of the quality dimensions (weights) specified by the user and ranking them based on the overall score. Thus, at the end of this step, the ontologies to consider are further reduced based on the threshold score set by the user.

Step 3: Ontology selection

After evaluating and ranking the ontologies, one can select the highest ranked ontology. However, it may not always be the right ontology to use. For example, assume that the user needs an ontology for a task related to airline reservation. In the previous step, the user may start with travel domain ontologies in general and evaluate and rank them. However, not all of them may be directly applicable for the task at hand. For example, the highest ranked ontology may specialize in other forms of transportation such as train or cruise reservation. Thus, the highest ranked ontology may not always be the best ontology to use. In this step, the ontology selection takes into account the task-artifact fit [12]. In addition to task characteristics, other factors such as reputation of the source, usage of the ontology, and user feedback are used in selecting the ontology. At the end of this step, the most appropriate ontology to use is suggested.

6 Conclusion and Future Work

We have discussed the need for effective information extraction systems for the web and presented an architecture for such a system. These systems rely heavily on domain ontologies and selecting an appropriate ontology for the task at hand is not trivial. We have presented an approach for ontology selection in a systematic way. The work outlined in this paper is research in progress and the proposed system is currently under implementation. There are many areas which can enhance and influence the current architecture. The processing and expansion of natural language based queries will have an impact on the proposed architecture particularly since ontologies are central to the architecture. Another enhancement could be the development of a module for automatic generation of rules as suggested by [15]. An issue that is becoming inherent to the information extracted from the web is its quality. Indeed, the web is replete with less reliable and untrustworthy information, contradictory statements, and fake data. Any future information extraction system that relies on the web should include some quality assessment. This would require the development of modules that will rely on intelligent techniques to filter and assess the extracted information. This would undoubtedly lead to the more challenging task of data personalization as quality is subjective and users have different needs. Users have different perception of information quality which depends on their experience, their use of the information and the risks they are ready to take in using unknown information.

Acknowledgements

The work of the first author has been partly supported by Sogang Business School's World Class University Program (R31-20002) funded by Korea Research Foundation.

References

1. Alani, H., Brewster, C.: Ontology ranking based on the analysis of concept structures. In: Proceedings of the 3rd International Conference on Knowledge Capture, pp. 51–58 (2005)
2. Aldea, et al.: An Ontology-Based Knowledge Management Platform. In: Proceedings of IJCAI 2003 Workshop on Information Integration on the Web (IIWeb 2003), Mexico, pp. 177–182 (2003)
3. Burton-Jones, A., Storey, V.C., Sugumaran, V., Ahluwalia, P.: A Semiotic Metrics Suite for Assessing the Quality of Ontologies. Data and Knowledge Engineering 55(1), 84–102 (2005)
4. Chaudhry, W., Meziane, F.: Information Extraction from Heterogeneous Sources Using Domain Ontologies. In: IEEE International Conference on Emerging Technologies, Islamabad, Pakistan, September 17-18, pp. 511–516 (2005)
5. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the 13th ACM Conference on Information and Knowledge Management, pp. 652–659 (2004)
6. Hendler, J.: Agents and the Semantic Web. IEEE intelligent Systems 16(2), 30–37 (2001)
7. Kushmerick, N., Thomas, B.: Adaptive Information Extraction: Core Technologies for Information Agents. In: Klusch, M., Bergamaschi, S., Edwards, P., Petta, P. (eds.) Intelligent Information Agents. LNCS (LNAI), vol. 2586, pp. 79–103. Springer, Heidelberg (2003)

8. Lozano-Tello, A., Gómez-Pérez, A.: OntoMetric: A method to choose the appropriate ontology. Journal of Database Management 15(2) (April-June 2004)
9. McDowell, L.K., Cafarella, M.: Ontology-Driven Information Extraction with OntoSyphon. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 428–444. Springer, Heidelberg (2006)
10. Patel, C., Supekar, K., Lee, Y., Park, E.K.: OntoKhoj: a semantic web portal for ontology searching, ranking and classification. In: Proceedings of the 5th ACM International Workshop on Web Information and Data Management, pp. 58–61 (2003)
11. Porzel, R., Malaka, R.: A Task-based Approach for Ontology Evaluation. In: ECAI Workshop on Ontology Learning and Population, Valencia, Spain (2004)
12. Simon, H.: Sciences of the artificial. MIT Press, Cambridge (1981)
13. Stephens, L.M., Huhns, M.N.: Consensus ontologies: reconciling the semantics of web pages and agents. IEEE Internet Computing 5(5), 92–95 (2001)
14. Vallet, D., Fernández, M., Castells, P.: An Ontology-Based Information Retrieval Model. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 455–470. Springer, Heidelberg (2005)
15. Yildiz, B., Miksch, S.: Motivating Ontology-Driven Information Extraction. In: International Conference on Semantic Web and Digital Libraries (ICSD 2007), Bangalore, pp. 45–53 (2007)
16. Zhang, Y., Vasconcelos, W., Sleeman, D.: Ontosearch: An ontology search engine. In: Proceedings of the 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, December 13 – 15, pp. 58–69 (2004)