Ontology Based Framework for Web Page Information Extraction

Naveen Gupta, Amit Sinhal.

Abstract -- Nature of Web information is dynamic and irregular that's why it is difficult to search and integrate information from the Web. The biggest task in making WWW data accessible to users/agents is extracting the data from Web pages. We take advantage of information in existing Web pages to creating structured data semi-automatically. Extraction of information from semi-structured or unstructured documents, such as Web pages, is a useful yet complex task. Research has demonstrated that ontology may be used to achieve a high degree of accuracy in data extraction while maintaining resiliency in the face of document changes. . This paper proposes an ontology-based information extraction system and its application to online book store domain. Testing result shows that this algorithm doesn't rely on the page structure and it can increase the recall and precision of information extraction.

Keywords – DSS, Extraction rules, Information Extraction Algorithm, Ontology, Precision, Recall, and Wrapper.

1. Introduction:

The advent of the World Wide Web (WWW) has taken the availability of information to an unprecedented level. The simplicity of the Web has been a major factor in its proliferation [1]. Anyone can easily publish a document about anything or link to anyone's site. The document needs not be structured according to any particular format or even contain correct information, and the link need not be valid [2].

Making sense of the vast amount of information available on the World Wide Web has become an increasingly important and lucrative endeavor. While traditional search engines can locate and retrieve documents of interest, they lack the capacity to make sense of the information those documents contain. Traditional search engines occupy the domain of information retrieval, which is the task of identifying from a large unstructured set of documents those are most relevant to a particular user query. However, after the relevant documents have been identified and ranked, it is up to the user to browse the results and attempt to make use of their information. Often, the users need is for highly specific information buried within the documents that the search engine returns. And at the time the results may exclude relevant documents because the keyword-based algorithms of the search engine lack the sophistication to understand the user's ultimate objective.

Information extraction from various types of sources became very popular during the last decade. Owing to information overload, there are many practical applications that can utilize semantically labeled data extracted from textual sources like the web sites, emails and even conventional sources like newspaper and magazines. Extraction of information from semi-structured or unstructured documents, such as Web pages, is a useful yet complex task. Research has demonstrated that ontology may be used to achieve a high degree of accuracy in data extraction while maintaining resiliency in the face of document changes.

Ontology does not, however, diminish the complexity of a data-extraction system.

In this paper we propose a framework based on ontology for such a system. The nature of the framework allows new algorithm and ideas to be incorporated into a data extraction system without requiring wholesale rewrites of a large part of the systems source code. We demonstrate the value of the framework by providing an implementation of it, and we show that our implementation is capable of achieving accuracy in its extraction results comparable to that achieved by the legacy Ontos data-extraction system.

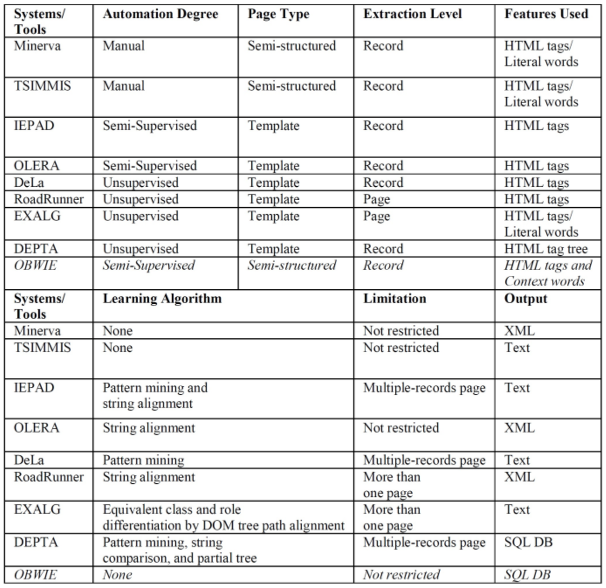
The main problems of current extraction methods (i.e. Table 1) are as follows:

1. Accuracy and robustness of Information Extraction System need to be improved.

2. The programs of information extraction rely on the structure of web pages, which makes programs can't be reused.

3. It needs to compile a new wrapper every time when a new web page comes.

Table 1: Comparison of proposed OBWIE with existing system



Systems/ Tools	Automation Degree	Page Type	Extraction Level	Features Used
Minerva	Manual	Semi-structured	Record	HTML tags/ Literal words
TSIMMIS	Manual	Semi-structured	Record	HTML tags/ Literal words
IEPAD	Semi-Supervised	Template	Record	HTML tags
OLERA	Semi-Supervised	Template	Record	HTML tags
DeLa	Unsupervised	Template	Record	HTML tags
RoadRunner	Unsupervised	Template	Page	HTML tags
EXALG	Unsupervised	Template	Page	HTML tags/ Literal words
DEPTA	Unsupervised	Template	Record	HTML tag tree
OBWIE	*Semi-Supervised*	*Semi-structured*	*Record*	*HTML tags and Context words*
Systems/ Tools	**Learning Algorithm**		**Limitation**	**Output**
Minerva	None		Not restricted	XML
TSIMMIS	None		Not restricted	Text
IEPAD	Pattern mining and string alignment		Multiple-records page	Text
OLERA	String alignment		Not restricted	XML
DeLa	Pattern mining		Multiple-records page	Text
RoadRunner	String alignment		More than one page	XML
EXALG	Equivalent class and role differentiation by DOM tree path alignment		More than one page	Text
DEPTA	Pattern mining, string comparison, and partial tree		Multiple-records page	SQL DB
OBWIE	*None*		*Not restricted*	*SQL DB*

3. Proposed Ontology Based Information Extraction

Ontology is considered as one of the key enabling technology for information processing task. Ontology identifies the entities that exist in a given domain and specifies their essential properties. It does not describe the spurious properties of these entities. On the contrary, the goal of Information Extraction is to extract factual knowledge to instantiate one or several predefined forms. The structure of the form is a matter of the ontology whereas the values of the filled template usually reflect factual knowledge that is not part of the ontology. In more powerful IE systems, the ontological knowledge is more explicitly stated in the rules that bridge the gap between the word level and text interpretation. As such, an ontology is not a purely conceptual model, it is a model associated to a domain-specific vocabulary and grammar. In the IE framework, we consider that this vocabulary and grammar are part of the ontology, even when they are embodied in extraction rules. The ontological knowledge involved in IE can be viewed as a set of interconnected and concept-centered descriptions, or "conceptual nodes". In conceptual nodes the concept properties and the relations between concepts are explicit. These conceptual nodes should be understood as ontology decomposition.

2.1 Decomposition of Ontology

Definition 1: Concept Item. Concepts can be defined as one Concept Item, if there are synonymies, near-synonymies or hyponyms between them in domain ontology. A concept item can be denoted as $CI = \{c_i | 1 \leq i \leq n\}$, where n is the amount of the concepts the Concept Item contains [6]. A Concept Item is identified by one of those elements. Take the ontology "Book" for example, concept "Price" and "Discount Price" are near-synonymies, "Author" and "Writer" are synonymies, author's "First Name" and "Last Name" are hyponyms. Therefore, the above four concepts can be classified as a Concept Item.

Definition 2: Concept Value. An instance of a concept in web pages is called a value of the concept [6]. For example,"Rs. 420" is a possible value of the concept "Price".

We have presented a simplified ORM diagram to represent the book ontology. Fig.1 is an ontology model on "Book" using ORM diagram.

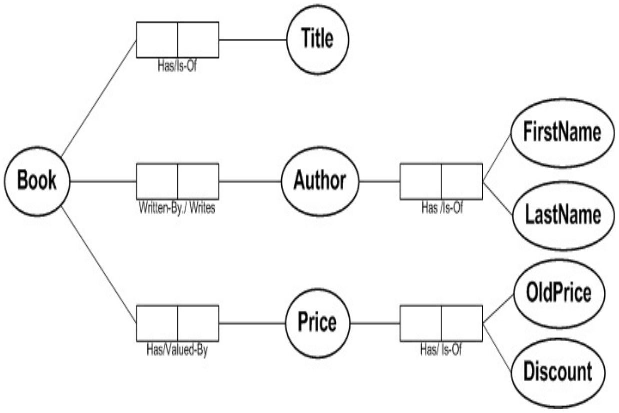


Figure 1: ORM Diagram for Book Ontology

In order to make the domain ontology more effectively guide the process of information extraction, the ontology needs to be decomposed. Ontology decomposition process is shown in Fig.2.

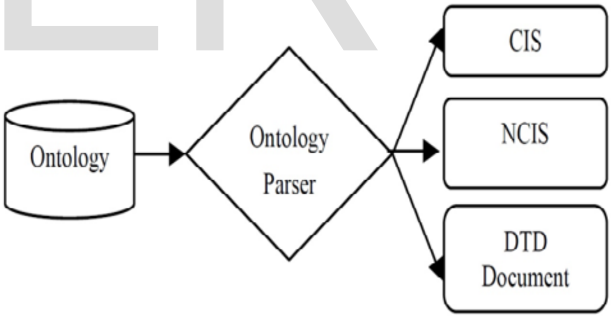


Figure 2: Ontology Decomposition Process

We take the ontology "Book" for example and explain the meanings of Concept Item Set, Necessary Concept. Item Set and DTD (Document Type Definition) Document.

Definition 3: Concept Item Set is the collection of all the concept items in domain ontology. It is represented as $CIS = \{c_i \leq i \leq s\}$, where s is the number of concept Items ontology contains [6].

Definition 4: Necessary Concept item set is the collection of the concept items that are necessary and indispensable to constitute the ontology.

It is represented as $NCIS = \{ci_i |1 \leq i \leq p\}$, where p is the number of core concept Items [6]. Necessary Concept item is manually determined. In the ontology “Book”, “Title”, “Author”, and “Price” are necessary concept items. Simultaneously, we mark

$PC = ci \times ci \times ... \times ci_n$, where $ci_i \in NCIS$.

Definition 5: DTD Document is defined based on the organizational structure of ontology [6]. Its transformation rules are as follows:

- Take lexical concept Items as the leaf elements in the DTD document;
- Quantitative relationship between concept Items in the DTD document is denoted by ”+”, ”-” or ”?”. For example, the title of book can only have one, while the price (OldPrice and DiscountPrice) can have one or more. They are reflected in the DTD document like:

  h! ELEMENT Book (Title, Price+,Author+)i.

- “Has” relationship and “is - a” relationship between concept Items are represented as “parent - child” relationship between elements in DTD document.

4. Proposed OBWIE Framework

Figure 3 shows our proposed framework we use to extract the data from an unstructured document and structure accordingly. The input to our framework is an application ontology and an unstructured document, and the output is a filtered and structured document whose data is in a database. Since all the processes and intermediate file formats are fixed in advance, our framework constitutes a general procedure that takes as input any declared ontology for an application domain of interest and an unstructured document within the application’s domain and produces as output structured data, filtered with respect to the ontology.

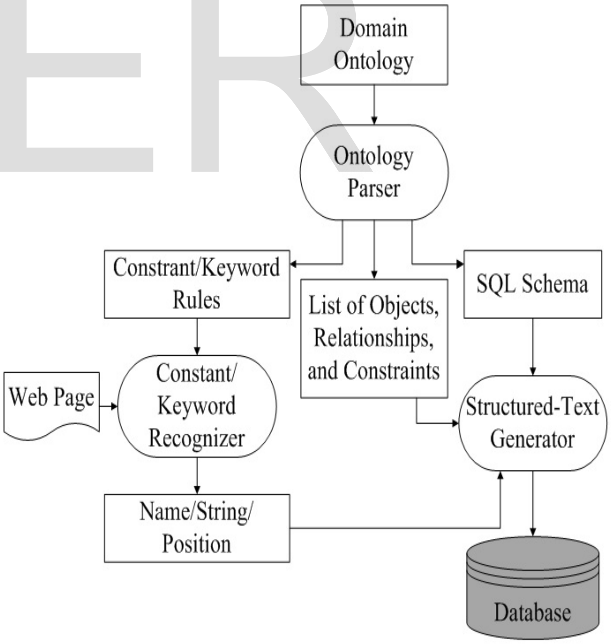

Domain Ontology
Ontology Parser
Constrant/Keyword Rules
List of Objects, Relationships, and Constraints
SQL Schema
Web Page
Constant/ Keyword Recognizer
Structured-Text Generator
Name/String/ Position
Database

Figure 3: Framework for ontology Based Information Extraction

The only step that requires significant human intervention is the initial creation of application ontology. However, once such an application ontology is written, it can be applied to unstructured documents from a wide variety of sources, so long as these documents correspond to the given application domain. Also, because our extraction is ontology-based, our approach is resilient to changes in source-document formats. For ex- ample, changes in HTML formatting codes do not affect our ability to extract and structure information from a given Web page.

As shown in Figure 3, there are three main components in our framework: an ontology parser, a constant/keyword recognizer, and a structured-text generator. The input is application ontology and a set of unstructured documents/web pages, and the output is a populated relational database. A main program invokes the parser, recognizer, and generator in the proper sequence. The ontology parser is invoked only once at the beginning of execution, while the recognizer and generator are repeatedly invoked in sequence for each unstructured document to be processed. These three main components are described further in detail in following sections.

5. Web Page Information Extraction Process

OBWIE finds and extracts relevant information with the help of a predefined ontology. The process starts with retrieving links of information of interest from explicitly provided URL(s). In an iterative manner, each link is explored which contain relevant information to extracted. The extraction module/framework takes domain ontology and web page description as input and perform extraction using rules by exploiting knowledge stored in ontology. This knowledge is stored in the form of concepts, relationships among concepts, data type properties, and context key words. The context words are stored in the comment section associated with each concept and data type properties. The extraction rules are defined as regular expression to describe the appearance of the value to be extracted.

The data type properties define the data type of a value such as integer, string, float, etc. Regular expressions are defined for each data type used in ontology and these rules are then used with context keywords defined in ontology to extract relevant information from ads description. Considering the unstructured nature of ads considered in the experiments, the location of relevant information is not fixed. To handle this issue, a list of context key words is used. If the context key word is found in web page description then this implies that the relevant information must be in the nearby position. Thus the relevant regular expression is applied in that region to extract the required information. The extracted data is then stored in the form of a table in relational database.

6. Experimental Result And Evaluation

The performance of the OBWIE framework evaluated on two selected case studies. Prototype system built on this framework. Experiments conducted, given high recall and precision near about 99% on test data.

The evaluation consists of two different case studies that use two different web resources (i.e. Websites) and same domain ontology (i.e. Book Ontology). The parameters for performance measure of information extraction are explained below.

Precision: - A measure of the ability of a system to present only relevant items.

$$\textbf{Precision} = \frac{\textbf{Number of relevant items retrieved}}{\textbf{Total Number of Iteams Retrieved}}$$

Recall: - A measure of the ability of a system to present all relevant items.

$$\textbf{Recall} = \frac{\textbf{Number of relevant items retrieved}}{\textbf{Number of relevant items in collections}}$$

7. Conclusion And Future Work

This paper proposes Ontology Based Framework for Web Page Information Extraction (OBWIE). OBWIE extracts the relevant information found in web pages across different sites describing same domain and structure those information in database. Our typical extraction process includes three steps: Firstly, ontology is developed that describes the domain knowledge. Secondly, data is extracted through rules with the help of context key words and data type available in the developed ontology. Finally, extracted data is stored in database. Except for Ontology creation, the processes in our framework are automatic and do not require user intervention. The advantage of OBWIE over other methods is that it does not rely on page structure. Once the domain ontology is built successfully, the IE process no longer need to be adjusted due to the change in structure of web pages. When new web pages appear, it does not need to rewrite source code; only extraction rules need to be changed.

We can extend our research by associating the concept of decision support system. We can use extracted information to make knowledge base. That knowledge base can be used for making decisions.

8. References

[1] M. Koivunen and E. Miller, " W3C Semantic Web activity," In E. Hyvonen, editor, Se- mantic Web Kick-Off in Finland, pages 2744, Helsinki, Finland, May 2002. Helsinki Institute for Information Technology, HIIT Publications.

[2] T. Berners-Lee, J. A. Hendler, and O. Lassila, "The semantic web," Scientific American, pages 2831, May 2001.

[3] Maedche A. "Ontology Learning," University of Karlsruhe, Germany: Kluser Academic Publishers, 2002.

[4] Laender, A.H.F., B.A. Ribeiro-Neto, A.S. da Silva, J.S. Teixeira, "A brief survey of Web data extraction tools," SIGMOD Record, Volume 31, Number 2, June 2002.

[5] Abiteboul, Serge, "Querying semi-structured data," In Proceedings of the 6th Inter- national Conference on Database

Theory, Delphi, Greece, 8-10 January 1997, 1-18.