International Journal of Current Trends in Engineering & Research (IJCTER)
e-ISSN 2455–1392 Volume 2 Issue 2, February 2016 pp. 12-18
Scientific Journal Impact Factor : 3.468
http://www.ijcter.com

Ontology Based Framework for Web Page Information Extraction

Shamsher Singh[1], Chandra Prakash Singh[2]

[1,2]*SRGI JHANSI*

Abstract—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's may be used to achieve a high degree of accuracy in data extraction while maintaining resiliency in the face of document changes. Ontology's do not, however, diminish the complexity of a data-extraction system.

Keywords—Ontology, Information Extraction, Web Ontology Language, OBWIE

I. 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 [2]. Anyone can easily publish a document about anything or link to any ones site. The document needs not be structured according to any particular format or even contain correct information, and the link need not be valid [3].

The field of data extraction (also called information extraction) addresses many of the problems highlighted above. Data extraction is the activity of locating values of interest within electronic textual documents, and mapping those values to a target conceptual schema [4].

Data extraction primarily focuses on locating values of interest within documents and associating the located values with a formal structure. Extraction is performed on unstructured or semi-structured documents, whose structure does not fully define the meaning of the data, as well as structured documents, which contain sufficient structure to allow un- ambiguous parsing and recognition of information [5], but whose underlying schema is not fully known to the extraction process [6]. A natural-language narrative is an example of an unstructured text; a Web page with sentence fragments and tables of values might be classified as semi-structured, and a comma-delimited file exported from a database is an instance of a structured document.

MOTIVATION

In today scenario, Internet has become one of the most important mediums to access in- formation, while the HTML-based Web pages have a natural shortcoming: tag of HTML language is only used to display the information but cannot express any semantic in- formation. Therefore, how to extract user-interested information automatically or semi- automatically has become a useful research. The current information extraction approaches can be divided into several categories. The main problems of current extraction method 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 cannot be reused [7].

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

To date, there has been however few IE approaches that could solve all three types of the above-mentioned problems. Our work extends an IE method that is based on the so-called domain ontology's [6] so that it can exploit these three types of extraction evidence, while using adapted domain ontology as the central point in the extraction process. Extraction rules are established through domain ontology and information extraction is finished based on these rules.

The main goal of this work is to design and implement a novel framework that is able to extract relevant features from a range of semi-structured documents i.e. Web pages. The designed methodology will be able to take profit from pre-processed input when it is available in order to take the advantage of efficiently created ontology. The key point of the work is to complement the syntactical parsing with the knowledge contained in an input ontology (ideally, it should model the knowledge domain in which the posterior data analysis will be focused e.g. book related points of interest) in order to be able to: 1) identify relevant features describing a particular entity from textual data, 2) To associate, if applicable, extracted features to concepts contained in the input ontology. In this manner, the output of the system would consist of well structured features in the form of SQL schema which can be directly queried by various SQL statements.

II. RELATED WORK

Researchers have devised a number of different approaches to the problem of data extraction. Perhaps the most common solution to the data-extraction problem is the construction and use of grammar-based wrappers [8] [9] [10]. Wrappers use clues in the documents structure and syntax to locate useful data. We summarize a few of the most prominent below to establish the background for our work.

Perhaps the most common solution to the data-extraction problem is the construction and use of grammar-based wrappers [8] [9] [10]. Wrappers use clues in the documents structure and syntax to locate useful data.

The primary drawback from which most wrapper approaches suffer is their dependence upon the actual syntax of the document markup to detect the boundaries between what is and is not relevant data. This means that when the markup of a site (not the data) changes, the wrapper is invalidated. Automated generation of wrappers alleviates this problem somewhat. Another problem is that a different wrapper is required for different document syntax, so thousands of wrappers may have to be generated and managed in order to adequately extract data for a particular subject domain.

HTML-Aware Tools that really on inherent structural features of HML documents for accomplish data extraction. Before performing the extraction process, these tools turn the document into a parsing tree, a representation that reflects its HTML tag hierarchy. Following, extraction rules are generated either semi-automatically or automatically and applied to the tree. Some representative tools based on such an approach are W4F [21], XWRAP [22] and RoadRunner [23].

Natural Language Processing Tools (also called NLP-based tools) have been used by several tools to learn extraction rules for extracting relevant data existing in natural language documents. These tools usually apply techniques such as filtering part-of-speech tagging, and lexical semantic tagging to build relationship between phrases and sentences elements, so that extraction rules can be derived.

Another class of data extractors (and the type of primary concern for this thesis) is the ontology-based extractors. These rely upon ontological descriptions of the subject domain as a basis for recognizing data of interest and for inferring the existence of objects and relationships that are not explicitly stated in the text. Representative ontology-based tools are the ones developed by the Brigham Young University Data Extraction Group [27] and X-tract [28].

Because ontology describes a subject domain rather than a document, ontology-based data-extraction systems are resilient to changes in how source documents are formatted, and they can handle documents from various sources without impairing the accuracy of the extraction.

III. PROPOSED WORK

Ontology's are considered as one of the key enabling technologies for information processing task. In this chapter, it is stated how ontology's have been applied in the process of IE from unstructured documents, specially focusing on domain specific information. 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 IE 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, 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 in section 3.1. Section 3.2 explains our proposed approach in two perspectives: (a) for information extraction, where formal description of relevant information in ontology [1] is utilized in extraction process and (b) to store information in structured representation, where extracted information is stored in database which helps in performing standard queries.

A. Proposed OBWIE Framework

Figure 1 shows our proposed framework we use to extract the data from an unstructured document and structure accordingly. The input to our framework is 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.

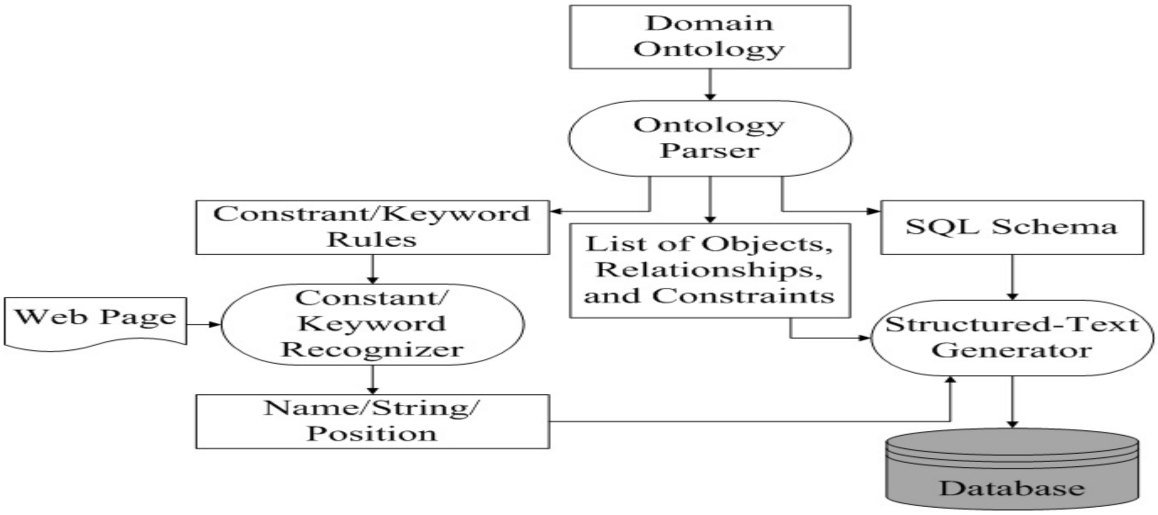
Domain Ontology
Ontology Parser
Constrant/Keyword Rules
List of Objects, Relationships, and Constraints
SQL Schema
Web Page
Constant/ Keyword Recognizer
Structured-Text Generator
Name/String/ Position
Database

Figure 1: 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.

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.

B. 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 performs 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.

Algorithm 1 : Ontology Based Information Extraction Algorithm

```
Set T=Null // use to store web page description
Set L= list of ads link
Set O= Pre-defined ontology for a domain developed in OWL Set
ContextWordList=Null
Set LexiconsOfValue[ ][ ] ={ {”\d\d*\..\d*”}, {\d*”}, {...} }
Begin
Step 1.Retrieve all ads links from the specified website. Step 2. For each ad
link L
     A. Read web page description text in T B. For each
     concept C in ontology O
         Set ContextWordList= words in comment section of C in ontology
         Create a new record R
```

```
            For each datatypeProperty D of C
        a. Append words in comment section of D in ContextWordList b. Set
        TypeOfValue=type of value of D
        c. If (TypeOfValue== float) then
            Set Rule= LexiconsOfValue[0]
            Else if(TypeOfValue== integer) then
            Set Rule= LexiconsOfValue[1]
            Else if (TypeOfValue== string) then
            Set Rule= LexiconsOfValue[2]
        d. For each context word cw in ContextWordList
            If found(cw) in T then
                Apply Rule in the neighborhood of cw and store the result in A
                //To check level of confidence a threshold is used for D
        e. For each value a of A
            If satisfies(pre-defined threshold for D) then
                Store a in R.
    C. Store R in the database
End
```

This section presents an information extraction algorithm that can locate and extract the data of interest from web pages across different sites. Our approach clearly departs from the ones in the previous literature in the following respects:

- Our algorithm is designed for the record-level extraction tasks that discover record boundaries, divide them into separate attributes, and associate these attributes with their respective values automatically. The algorithm does not rely on training examples, and does not require any interaction with the users during the extraction process.
- It works without the requirements that the web pages need to share the similar template or multiple records need to occur in a single web page. The algorithm can treat a single web page containing only one record. If a set of keywords used to describe the data of interest is collected, the extraction is fully automated, and it is easy to move from one application domain to another.
- This approach works for pages from many distinct web sources belonging to the same application domain. It also has the capacity of continuing to work properly when the format features of the source pages change, thus it is completely insensitive to changes in web-page format.

IV. RESULT ANALYSIS OF PROPOSED WORK

Our experimental data comes from the web, and we believe that the framework we pro- pose here is useful for extracting and structuring web page information. Once we populate our database after extracting interested information from web pages, we can query the in- formation using standard query languages such as SQL.

This section describes the performance of proposed OBWIE system on selected case studies (1 &2). For case study-1, 31712 book information are extracted for the search by “the secret” keyword. But for the purpose of this report

we limited ourselves to 20 randomly selected books. The information extracted from web site,”www.bookadda.com,” is shown in Table 3 below. Columns 2-5 of the table describe the four data elements about which information is extracted. The extracted values are matched against the ones obtained through manual browsing. Out of 20 selected items, OBWIE extracted 19 values with

95% accuracy for second column, 20 values with 100% accuracy for third, fourth and fifth column. On average, OBWIE extracted information with 98.75% accuracy.

Table 1: Information Extracted from BOOKADDA site

S.N.	Book Title	Author	Old Price(Rs)	Discount Price(Rs)
1	Past Secrets	Cathy Kelly	325	310
2	The Secret Life of Objects	Dawn Raffel Sean Evers	973	798
3	Dermatology Secrets In Color	Fitzpatrick	945	929
4	Gi/Liver Secrets, 4/E	Peter R Mcnally	910	885
5	The Secret Rahasya	Rhonda Byrne	285	224
6	Secrets And Sins	Jaishree Misra	299	283
7	The Secret Of Guidance	F. B. Meyer	457	390
8	Lucinda’s Secret	Holly Black Tony Diterlizzi	199	160
9	The Secret Of The Runes	Guido List Stephen E. Flowers	856	625
10	∗ Ayurveda Secrets Of Healin	Element Books Ltd.	1315	1144
11	Her Secret Lover	Sara Bennett	350	
12	The Darkest Secret	Gena Showalter	457	398
13	Christian’s Secret Of Happy Life	Hannah Smith	457	374
14	The Secretary’s Secret	Michelle Douglas	285	210
15	One Secret Thing	Sharon Olds	971	706
16	State Secrets (Famous Firsts)	Linda Lael Miller	285	267
17	Barbie: A Fairy Secret (Barbie)	Mary Man-Kong	228	187
18	A Secret Love (Cynster Novels)	Stephanie Laurens	315	285
19	The Secret Art Of Dr. Seuss	Maurice Sendak	2005	1550
20	Day Of Ahmed’s Secret	Florence H. Parry	400	346

V. CONCLUSION & FUTURE WORK

In this work, we proposed 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. We can use extracted information to make knowledge base. That knowledge base can be used for making decisions. If we consider the same case studies discussed to understand the scenario, extracted book information form knowledge base. We will implement decision support system that will use knowledge base to make decisions like recommending books on basis of user’s criteria.

REFERENCES

[1] Junfang Shi,LiLiu, ”Web Information Extraction based on News Domain Ontology Theory,” In Web Society (SWS), 2010 IEEE 2nd Symposium, Page(s): 416 - 419, Aug 2010.

[2] 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.

[3] T. Berners-Lee, J. A. Hendler, and O. Lassila, ”The semantic web,” Scientific Amer- ican, pages 2831, May 2001.

[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.

[6] Embley, D.W., C. Tao, S.W. Liddle, ”Automatically extracting ontologically specified data from HTML tables with unknown structure,” InProceedings of the 21st Interna- tional Conference on Conceptual Modeling, Tampere, Finland, 7-11 October 2002, 322.

[7] YouLi,PangHongshen, ”The Overview of foreign Research on web Information Ex- traction,” Library Journal,2008,30(5):13-15

[8] Hammer, Joachim, Hector Garcia-Molina, Svetlozar Nestorov, Ramana Yerneni, Marcus Breunig, Vasilis Vassalos, ”Template-based wrappers in the Tsimmis system,” In Proceedings of the ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, 13-15 May1997. .

[9] Kushmerick, N., D. Weld, R. Doorenbos, ”Wrapper induction for information extrac- tion,” In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Nagoya, Japan, 23-29 August 1997.

[10] Ashish, N. and C. Knoblock, ”Wrapper generation for semi-structured Internet sources,” In Proceedings of the Workshop on Management of Semistructured Data, Tucson, Arizona, May 1997.

[11] Crescenzi, Valter, Giansalvatore Mecca, Paolo Merialdo, ”RoadRunner: Towards automatic data extraction from large Web sites,” In Proceedings of the 27th Interna- tional Conference on Very Large Data Bases, Roma, Italy, 11-14 September 2001.

[12] Eikvil, Line, ”Information extraction from the World Wide Web: A survey,” Techni- cal Report 945, Norwegian Computing Center, 1999.

[13] Davulcu, H., S. Mukherjee, I.V. Ramakrishnan, ”Extraction techniques for mining services from Web sources,” In Proceedings of the IEEE International Conference on Data Mining (ICDM), Maebashi, Japan, 9-12 December 2002, 601-604.

[14] Engels, Robert, ”Del 7: CORPORUM OntoWrapper: Extraction of structured information from web based resources,” On-to-Knowledge Consortium, 2002. At http://www.ontoknowledge.org.

[15] D. Konopnicki and O. Shmueli, ”A query system for the world wide web,” In Pro- ceedings of the Twenty-first International Conference on Very Large Data Bases, pages 5465, Zurich, Switzerland, 1995.

[16] L.V.S. Lakshmanan, F. Sadri, and I.N. Subramanian, ”A declarative language for querying and restructuring the web,” In Proceedings of the 6th International Work- shop on Research issues in Data Engineering, RIDE96, New Orleans, Louisiana, 1996.

[17] A.O. Mendelzon, G.A. Mihaila, and T. Milo, ”Querying the world wide web,” Inter- national Journal on Digital Libraries, 1(1):5467, April 1997.

[18] Crescenzi, V. and Mecca, G.,”Grammars have exception,” Information Systems, vol. 23 (8), pp. 539-565, 1998.

[19] Hammer, J.; Breunig, M.; Garcia-Molina, H.; Nestorov, S.; Vassalos, V. and Yerneni, R., ”TemplateBased Wrappers in the TSIMMIS System,” Proceedings of Twenty- Third ACM SIGMOD International Conference on Management of Data, pp.532-535. Tucson, Arizona, USA, 1997

[20] Arocena, G. O. and Mendelzon, A. O., ”WebOQL: Restructuring documents, databases and webs”, Proceedings of 14th International Conference on Data Engi- neering, pp.24-33. Orlando, Florida, USA, 1998.

[21] Sahuguet, A. and Azavant, F., ”Building Intelligent Web Applications using lightweight Wrappers,” Data and Knowledge Engineering, vol. 36 (3), pp. 283-316, 2001.

[22] Liu, L.; Pu, C. and Han, W., ”An XML-enabled wrapper construction system for web information sources,” Proceedings of 16th International Conference on Data En- gineering (ICDE’00), pp.611-621. San Diego, CA, USA, 2000.

[23] Crescenzi, V.; Mecca, G. and Merialdo, P., ”RoadRunner: Towards Automatic Data Extraction from Large Web Sites,” Proceedings of 27th International Conference on Very Large Data Bases, pp.109-118. Rome, Italy, 2001.