International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

Information Extraction from Unstructured and Ungrammatical Data Sources for Semantic Annotation

Quratulain N. Rajput, Sajjad Haider, Nasir Touheed

***Abstract*—**The internet has become an attractive avenue for global e-business, e-learning, knowledge sharing, etc. Due to continuous increase in the volume of web content, it is not practically possible for a user to extract information by browsing and integrating data from a huge amount of web sources retrieved by the existing search engines. The semantic web technology enables advancement in information extraction by providing a suite of tools to integrate data from different sources. To take full advantage of semantic web, it is necessary to annotate existing web pages into semantic web pages. This research develops a tool, named OWIE (Ontology-based Web Information Extraction), for semantic web annotation using domain specific ontologies. The tool automatically extracts information from html pages with the help of pre-defined ontologies and gives them semantic representation. Two case studies have been conducted to analyze the accuracy of OWIE.

***Keywords*—**Ontology, Semantic Annotation, Wrapper, Information Extraction.

I. INTRODUCTION

THE popularity of the World Wide Web (WWW) has resulted in an information explosion and has made it extremely difficult for users to find and utilize information in an efficient manner. Information over the web is not placed into a central repository where standard queries can be applied to access relevant information. Moreover, the web is filled with unstructured content and searching pertinent information using the existing keyword based search engines has two major limitations: (a) manual browsing of long list of retrieved links and (b) manual integration of data from different web pages. Data integration requires combining and matching information coming from different sources and resolving a variety of discrepancies [13, 15]. However, extraordinary increase in the amount of data as well as the diversity of structures in which data is stored creates tremendous complication in this process [4, 5].

Q.N. Rajput is with the Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan (phone: (92-21)111677677; fax: (92-21)9215528; e-mail: quratulain.rajput@ gmail.com).

S. Haider is with the Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan (e-mail: sajjad.haider@khi.iba.edu.pk).

N.Touheed is with the Faculty of Computer Science, Institute of Business Administration, Karachi, Pakistan (e-mail: ntouheed@iba.edu.pk).

During the last few years, semantic web technologies [22, 24] have emerged as a much needed platform that has the potential to turn the dream of data integration into reality.[16] Semantic web is an extension of the current web in which information is given well-defined meaning, thus making it possible for machines to understand web content. It consists of elements such as RDF/XML, RDF Schema, and OWL which facilitate both website developers and users in expressing formal description of concepts and their relationships. [2]

Fig. 1 Craigslist in Syntactic Web

To understand the main difference between the syntactic (existing) web and the semantic web, consider the following example. Suppose a user is interested in buying a laptop with the following characteristics: processor: Intel, price: < $1500, and RAM: 1GB. In the syntactic web, a user performs keywords based searches on websites dealing with the selling/purchasing of laptops. The search engine returns many entries as documents link which satisfy the user's criteria, either completely or partially. Fig. 1 shows the result of a query obtained through the craigslist website. Now the user has to browse this huge list of links to identify the relevant information matching his/her criteria. Because of the time taken by manual browsing, the users typically browse only top few links or select the links randomly by guessing from the titles of the links. One of the aims of Semantic web is to overcome the above mentioned problem by adding semantics to the web content which makes the task of finding and integrating relevant information from different sources/pages a lot easier. Table I shows an output of the laptop purchase query, mentioned above, if the data were organized using semantic web technology. The first row of the table indicates attribute names and the last column indicates the original link of those ads from where the data is extracted. The empty cells show that information is not available on the corresponding web pages or the annotation system failed to recognize it.

One of the main challenges in fully realizing the goal of semantic web is the handling of existing web pages. Most of the pages do not contain semantic information. . Moreover, the data on those pages is stored in diverse structure at different sources which makes data sharing extremely difficult. The goal of *semantic annotation* is to markup the web pages with semantic information that defines the meaning of contents on those pages.

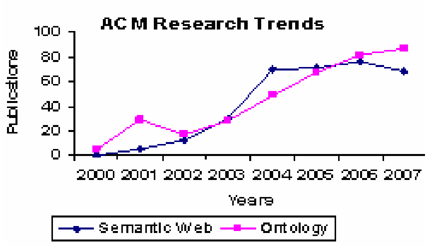
TABLE I INFORMATION EXTRACTION IN SEMANTIC WEB FROM CRAIGSLIST

Laptop	**Brand**	**Speed**	**Ram**	**HDisk**	**Size**	**URL**
001	IBM	1.6GHz	1GB	60 GB	14”	1
002	Toshiba		256MB			2
003		2.4GHz	4 GB	320 GB	17”	3

Much of the research in semantic annotation has been focused on finding relevant data using information extraction techniques. Many tools have been reported in the literature based on wrapper languages and wrapper induction [3, 11, and 17], HTML-tag awareness [19, 6], natural language processing and model-based [1, 20]. [10, 14, 18] provides a detailed overview of different information extraction techniques used in semantic annotation. Another important category of tools is based on ontologies. In fact, the past few years have seen a growing interest in the use of ontology for semantic web related activities. A crude survey of the number of papers, appearing in IEEE and ACM portals since 2000, shows a dramatic increase in papers having semantic web or ontology as keywords (Fig. 2.) Ontology based tools for semantic annotation support automatic and semiautomatic annotation using domain specific ontologies. These ontologies describe data of interest, their relationship, lexical appearance, and context keywords. Some of the important ontology-based tools for semantic annotation are BYU [7-9], MnM [23], S-Cream [12], and iASA [21], ontoX [25-26].

This paper presents an ontology-based tool, named OWIE, to facilitate the semantic annotation process. At the theoretical level, the research is similar to the work done by Embley et al. [7-9] and Yildiz et al. [275-26] as it also develops ontology-based information extraction. The case studies selected in this research, however, are unique from the previous reported work as they provide a blend of highly structured/unstructured and ungrammatical source having irregular size of information.

The rest of the paper is organized as follows. Section II discusses the underlying process model of OWIE and the selected case study. Results of the experiments are presented in Section III. Finally, Section IV concludes the paper and provides future research directions.



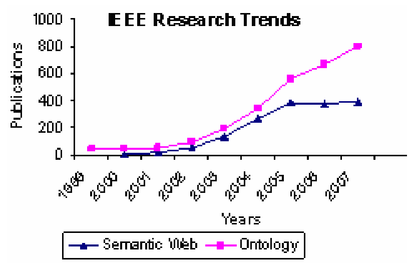


Fig. 2. Research Trends in Semantic Web and Ontology

II. OWIE: AN ONTOLOGY-BASED WEB INFORMATION EXTRACTION

Ontologies are considered as one of the key enabling technologies for semantic web. In addition to being applied to many other areas, a lot of efforts have been made in applying ontologies for information processing task, specifically in information extraction systems (IESs). Such systems extract domain-specific information from natural language text. The domain and type of information to be extracted is typically defined in advance to help in relevant information extraction. As discussed in the previous section, the focus is towards the integration of ontologies with IES to provide unambiguous and formal description of relevant information that is utilized by IES. This research also provides a methodology to integrate ontology in IESs. This section explains our proposed methodology for extracting information from unstructured content and then associating semantics to the extracted data. Ontologies are used in two different 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 semantic representation, where extracted information is stored in ontology which helps in performing conceptual queries.

This research develops a tool to automatically extract data from unstructured web sources and annotate it with semantic information. The semantic annotation enables the data to be easily accessible using standard query language. The tool is named OWIE (Ontology-based Web Information Extraction). It finds and extracts relevant information with the help of a pre-defined ontology. The graphical description of the complete process is shown in Figure. 3. The process starts with retrieving links of information of interest from explicitly provided URL(s). In an iterative manner, each link is explored

[1] Ontologies are developed in Protégé, it is open source ontology editor developed by Stanford University. downloaded from http://protege.stanford.edu

which contain ad description posted by different users. The extraction application module takes domain ontology and ad 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 words. The context words are stored in the comment section associated with each concept and data type properties. The rules are defined as regular expression to describe the appearance of the value to be extracted.

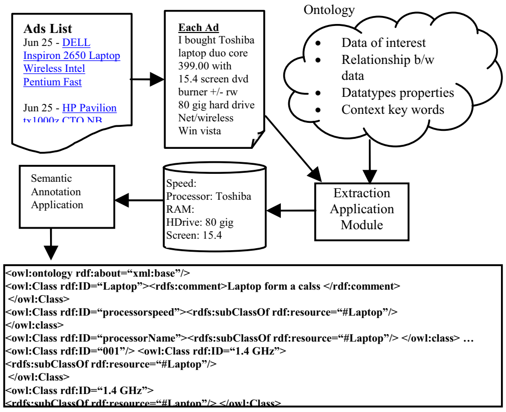


Fig. 3 Ontology-based Web Information Extraction

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 words is used. If the context word is found in ad 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 and is annotated with semantic information using OWL[2]. The semantically annotated data can then be queried for specific information. The steps involved in the extraction process are also presented in Table II.

To test the capabilities and limitations of OWIE, two case studies have been conducted. The first is the selling/purchasing of laptops on the craigslist website[3]; while the other is a scholarship resource center on the scholarshipnet website[4]. The craigslist website is a centralized network of online communities, featuring free classified advertisements (with jobs, internships, housing, personals, for sale/barter/wanted services, etc.) and forums on various topics. Users can put advertisement in their own free style format. The ScholarshipNet.info is an international scholarship resource providing scholarship advertisements (at MS, PhD, or Postdoc level) and study abroad guidelines to international students from all over the world. Students interested in availing a scholarship can search available scholarship according to their requirements. The semantic annotation process includes three main steps.

A. Ontology development to capture domain knowledge.
B. Data is extraction with the aid of context words and data types defined in the ontology.
C. Extracted data is stored in semantic representation in OWL.

In this sequel, the first two steps are elaborated further.

TABLE II
ONTOLOGY-BASED INFORMATION EXTRACTION ALGORITHM

```
Set T=Null // use to store  ad description
Set L= list of ads link
Set O= pre-defined ontology for a domain developed in protégé
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 ad 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
```

A. Ontology Development

Ontology defines the concept model of a particular domain. It serves as a wrapper by defining the context information, the possibilities in which data appears over the page, and the relationship among data elements with respect to the domain knowledge. The first step of any ontology based semantic annotation system is the development of domain specific ontologies. Fig. 4 and Fig. 5 show the possible conceptualization for laptop and scholarship domains, respectively, where undirected lines indicate data type properties. In the laptop ontology, P_Speed, B_Name, D_Size, ramsize and HDsize are data type properties. The data types are defined as follows: processor speed as float, brand name as string, display size as float, and memory as integer. The scholarship ontology use string data type for all values except deadline which has date data type. These ontologies aid a user

[2] OWL is Web Ontology Language and is endorsed by W3C Consortium.
[3] www.craigslist.org
[4] www.scholarshipnet.info

to perform queries at different conceptual levels. For instance, if a user wants to know about available scholarships in physical sciences in North America, and if the data has been semantically annotated, then using the ontology of Fig. 5 the query system can return all physical sciences scholarships available in countries within the North American region without irrelevant information.

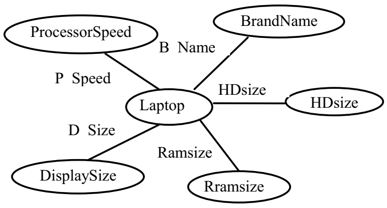


Fig. 4 Graphical view of Ontology for Laptop

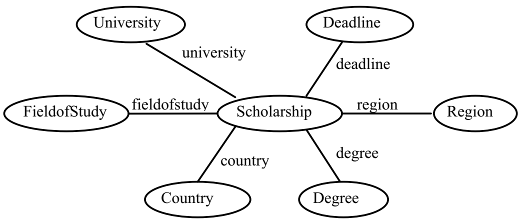


Fig. 5 Graphical view of Ontology for Scholarship

B. *Extraction of Data Element*

After successful specification of ontology the next task is that of information extraction. To extract relevant information from a list of ads, each link is accessed in an iterative manner. Most of the times, pages are accessed successfully but occasionally "Page Not Found" message appears. The primary reason for this access failure is either load on local network or deletion of the link from the corresponding website. Once a web page is found, the next task is to identify relevant data elements on the page. Tables III and Table IV show samples of ads from craigslists and scholarshipnet websites and highlight the difficulties and challenges that are present during the extraction process. The first sample ad in Table III is simply a paragraph without following the grammatical rules of the English language. It simply highlights the important features of a laptop separated by dashes (-). The second sample briefly describes the main feature of a laptop to be sold by the ad provider. The third sample provides very detailed information. It is obvious that the required information is available in all three samples but in very different and highly unstructured style, thus making it difficult for machines to understand it. The amount of information also varies significantly. The sample scholarships ads shown in Table IV are different from the laptop ads as the information is organized in a structured manner. The heading of the scholarship ad contains important data elements, such as country, degree, area of study, and university. These elements are also part of the scholarship ontology discussed earlier. It can be observed that the heading contains information in a fixed format and thus can be easily accessed by a HTML parser.

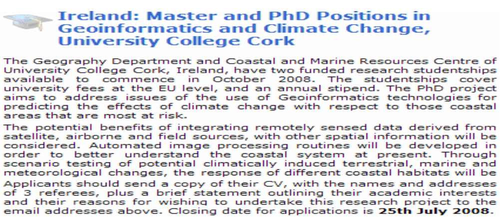
TABLE III
SAMPLES OF LAPTOP ADS

Ads	Ads Description for Laptop from Craigslist Site
1.	Used Dell Latitude CPx Laptop - Microsoft Office - Windows XP Professional - reliable, and in good operating condition - includes computer and charger/adaptor - 2 PCI network cards (1 wireless & 1 not)- extra external mo ethernet port replicator - screen is bright, colorful and clear - has CD ROM drive- PC Card slots for wireless or v network adaptors - this computer is great for uses such as Web browsing, document editing, general office or stu use, etc - not suitable for the newest high powered video games. good battery, charger. 650 htz P3 Processor 256 Mb
2.	i bought toshiba laptop duo core 399.00 with 15.4 screen dvd burner +/- rw 80 gig hard drive net /wireless win vista evrything working good nothing wrong now i 'm saling that tho who pay more 399.00 i will sale email to me
3.	Brand New! Never Been Opened! Never Used! In Original Retail Box! Dell Latitude D420 12.1" Ultra-Portable Laptop With 3 Years Dell Warranty! Dell's smallest business laptop Slim, lightweight design; full-size keyboard Only about 3 lbs! Super Loaded!! Latitude D420, Intel Core Duo U2500, 1.20GHz ULV, 533Mhz 2M L2 Cache much faster than solo processor Three Years Dell Warranty!! 12.1 inch Wide Screen WXGA LCD 1.0GB, DDR2-667 SDRAM (512MB Integrated) 60GB Hard Drive 8MM, 4200RPM Windows XP Professional, SP2 with media Super D/BAY plus 24X CDRW/DVD with Cyberlink Power DVD for Vista Basic/Busines 6-Cell/42 WHr Primary Battery for Dell all Latitude D420 Super Dell Wireless 1390 WLAN (802.11g,54Mbps) Mini Card Latitude

The extraction of relevant information is accomplished by assuming a particular sequence of values in ads, such as country, degree, field of study and finally university. The colon and comma are considered as a separator to separate these values. Alternatively, a dictionary can be maintained for all instances of a property which can then be use to extract the relevant information. Another important data element "scholarship deadline", however, is typically available in the body of the ad and its extraction from unstructured text is a challenging task. Like other data elements in the laptop ontology, this element is handled by the use of context words defined in scholarship ontology.

After getting ads description, the next task is to recognize individual data elements and assign attribute names to them with the aid of ontology. It should be mentioned that for humans it is easy to recognize the data by viewing advertisements but this recognition process is not as simple for machines. For example, if we want to find the size of a hard disk, our automated tool must have an understanding of all the possible ways hard disks are mentioned in advertisements. Moreover, it should be able to distinguish among similar values belonging to different data elements. Thus, the quality of the extraction depends upon the specification level of the built ontology - the more specific and more detailed the ontology is the better are the extraction results.

TABLE IV
SAMPLES OF SCHOLARSHIP ADS

Ads	Ads Description for Scholarship from Scholarshipnet Site
1.	 International Scholarship Resources SCHOLARSHIP, POSTDOC, & GRANTS INFORMATION LISTINGS scholarshipnet HOME | ABOUT | FAQS Norway: PhD Research Fellowships in Evolutionary Ecology, University of Oslo *Please kindly mention ScholarshipNet when applying for this position* PhD RESEARCH FELLOWSHIP (STUDENTSHIP) IN EVOLUTIONARY ECOLOGY available at the Department of Biology, Faculty of Mathematics and Natural Sciences, University of Oslo, Norway. We seek to employ a motivated PhD candidate in a project on life-history variation and climate change on carabid beetles in alpine environments.
2.	 HOME | ABOUT | FAQS Italy: PhD Scholarships in Economics, Università Ca' Foscari di Venezia The Advanced School of Economics (in Italian: Scuola Superiore di Economia, or SSE for short) runs two doctoral programs: one in Economics and one in Business. The Doctoral Program in Economics (DEC) and the Doctoral Program in Business (DEA) emanate from the Department of Economics and the Department of Business Management at the University of Venice. The School coordinates these two programs, that share several first-year courses and a common philosophy. Since 2008, the School has taken over over the activities of a third doctoral program in Economics and Organisation that used to be run as part of the School for Advanced Studies in Venice. The School is based in Venice and offers only courses at a doctoral level.
3.	 Ireland: Master and PhD Positions in Geoinformatics and Climate Change, University College Cork The Geography Department and Coastal and Marine Resources Centre of University College Cork, Ireland, have two funded research studentships available to commence in October 2008. The studentships cover university fees at the EU level, and an annual stipend. The PhD project aims to address issues of the use of Geoinformatics technologies for predicting the effects of climate change with respect to those coastal areas that are most at risk. The potential benefits of integrating remotely sensed data derived from satellite, airborne and field sources, with other spatial information will be considered. Automated image processing routines will be developed in order to better understand the coastal system at present. Through scenario testing of potential climatically induced terrestrial, marine and meteorological changes, the response of different coastal habitats will be Applicants should send a copy of their CV, with the names and addresses of 3 referees, plus a brief statement outlining their academic interests and their reasons for wishing to undertake this research project to the email addresses above. Closing date for applications is **25th July 2008**.

The proposed OWIE tool uses regular expressions to describe values of the data type properties. These expressions are defined once in the extraction application. The location identification is performed with the aid of the pre-defined context words. It happens, however, that in many occasions the tool fails to distinguish between different data elements. For instance, the appearance of hard disk size and RAM size is very similar, such as 1GB RAM, 40GB hard drive, etc. In both cases, the last digit ends with GB/MB. In some cases, ads do not even use the context words such as memory, RAM, Hard Drive, etc. To handle such situations, rules have been defined that test the values against a pre-defined threshold. If the value is greater than the threshold value than the value belongs to hard disk otherwise it belongs to RAM.

Occasionally ads contain duplicated information in different format. For example the third ad in Table III first describes the RAM size as 1GB but later clarifies that there are two 512MB SDRAM. Thus, the RAM size occurs twice and it creates difficulty to make the right choice while extracting in such situations. Similarly, some ads contain processor speed as well as bus speed, both of which are provided in GHz. This makes it difficult to pick the right value. The situation demands the incorporation of sophisticated rules to have better accuracy in the extraction process. Furthermore, users provide information in their own free style; this increases the likelihood of spelling mistakes, typos, different abbreviation. These mistakes can be easily handled by humans but not by an automated tool.

The ads on the scholarshipnet website are comparatively more structured than the ads on craigslist website. Hence, most of the problems discussed above such as spelling mistakes and ambiguity do not occur while extracting relevant data from the scholarshipnet website. Occasionally, however, the university names are written in languages other than English (sample 2 of Table IV) which creates problem and it treats the university name as missing. The following list in Table V categorizes the main challenges that arise during the information extraction phase in both case studies.

TABLE V
CHALLENGES FOR CRAIGSLIST SITE AND SCHOLARSHIPNET SITE

Challenges	**Craigslist (Laptop)**	**Scholarshipnet**
URL unrecognized	X	X
Unstructured information	X	X
Ungrammatical/ Spelling mistakes	X	
Variable Size of information	X	X
Appearance	X	X
Unrecognized	X	X

- **URL unrecognized:** To process the available information, the links of all relevant documents have to be retrieved. During this retrieval phase, links are found to be deleted from the corresponding web sites or are unavailable due to network problem.
- **Unstructured information:** Data is typically not organized in a specific order. This is specially true for laptop ads, where users enter information in a variety of format. Thus, location of information is not fixed. The same is also true for scholarship web site, but to a lesser degree.
- **Ungrammatical/spelling mistakes:** On the craigslist website, information is not available in proper sentence form. Some ads use abbreviations of different terms and some use different conventions for similar data elements. This leads to higher chances of typing mistakes.
- **Variable Size of information:** On the craigslist website, ads' sizes vary tremendously depending upon the information provided by the user. Some users provide very detailed information including photographs of the item, while some users simply write a phrase highlighting the most important features of the item. At the scholarship website information is typically available in fixed size.
- **Appearance:** Different data elements with similar appearances and same data elements with different appearances lead to the identification problems. For example, in ads on the craigslist web site, RAM and hard drives have same appearances, such as 20MB, 20GB, etc. This ambiguity can be resolved by adding some rules during the information extraction process but the solution might not be so easy in some other domains. On the scholarship website, the scholarship deadline is sometimes referred to as "Closing Date" and at times as "Application Deadline".
- **Unrecognized:** Sometimes the required information is available in very unique format which may not be easily

recognizable.

It is obvious that the ads layout on the craigslist web site depends upon the inputs provided by ordinary user and thus creates more challenges as compared to the scholarship website where the ads are provided by different universities in a formal style.

III. RESULT

This section describes the performance of the OWIE tool on the selected case studies. For the laptop case study, 1000 ads were extracted from the craigslist website but for the purpose of this report we limited ourselves to the 30 randomly selected laptop ads. The information extracted from these ads is shown in Table VI. Columns 2-6 of the tables describe the five data elements about which information is extracted. The extracted values are matched against the ones obtained through the manual browsing of these ads by a human being. The highlighted cells indicate incorrect information extraction. This could be due to classifying a non-missing value as missing or vice versa. Out of 30 selected ads, OWIE extracted information from 21 ads with 100% accuracy, 8 ads with 80% accuracy and 1 ad with 60% accuracy. On average, OWIE extracted information with 93% accuracy.

TABLE VI
INFORMATION EXTRACTED FROM CRAIGSLIST SITE

S.no	Speed	Name	RAM	HDive	Screen
1	1.400 GHz	DELL	512 MB	20GB	
2	1.6GHz	IBM	1GB	60GB	14"
3	500 ghz	Dell		20GB	
4	2.4GHz		4GB	320GB	17 Inch
5	2.4GHz		4GB	320GB	17 Inch
6	2.4GHz		4GB	320GB	17 Inch
7	1.60GHz	Dell	512MB		
8	1.5ghz	Dell	1GB	160GB	13.3"
9			512MB	40GB	15.4"
10	1.9GHZ	DELL	1GB	320GB	19"
11			2GB		
12	2.4GHZ		2GB	160GB	
13	1.80GHz			64GB	
14	3.06GHz	Toshiba	1GB	60GB	15.4"
15				60GB	
16	1.8GHZ	DELL	512MB	20GB	014"
17		Toshiba	256MB		
18			1GB		
19		dell		40GB	
20	2.2Ghz		54Mb	10GB	15.4"
21	1.73Ghz	Toshiba	1GB	120GB	
22	3.4GHZ		1GB	300GB	
23					
24		HP	2038MB		
25		Dell			
26			256mb		
27					
28		Dell	192MB	11GB	
29	1.5 GHz		1GB	80GB	12"
30	1.5ghz	Toshiba		40GB	
%	93.3%	100%	80%	100%	93.3%

For the scholarshipnet website, we have again randomly selected 30 scholarship advertisements for our analysis. The information extracted from these ads is shown in Table VII. The relevant data elements form the header of Columns 2-6. Out of 30 selected ads, OWIE extracted information from 26 ads with 100% accuracy and from 4 ads with 80% accuracy. On average, OWIE extracted information from scholarship net with 97% accuracy.

Tables VIII and IX show the *precision* and *recall* values for data elements extracted by OWIE for laptop and scholarship case studies, respectively. In the current context, recall is defined as the ratio of the number of relevant document retrieved to the total number of relevant documents, while precision is defined as the ratio of the number of relevant documents retrieved to the total number of documents retrieved.

TABLE VII
INFORMATION EXTRACTED FROM SCHOLARSHIPNET SITE

Sno	Country	Degree	Field of Study	University	DeadLine
1	USA	Postdoctoral	Biomedical Informatics	Columbia University	
2	Norway	PhD	Short Range Sensing	Localization and Wireless Communication	June 10, 2008
3	UK	PhD	Statistics	University of Bristol	1st August 2008
4	Ireland	PhD	Physics of Nanostructures	Tyndall National Institute	
5	Norway	PhD	Informatics	University of Oslo	13 June 2008
6	UK	PhD	Bioproduction	NewcastleUniversity	31st July 2008
7	Norway	PhD	Mathematics	University of Bergen	6 June 2008
8	Italy	PhD	Economics	Universit	May 20, 2008
9	Ireland	MA	European Studies		
10	France	PhD	Color Content-Aware Image Processing		June 1, 2008
11	France	PhD	Aroma and Perfume Research	University of Nice-Shophia Antipolia	Thursday 22 May 2008
12	Australia	PhD	Bioinformatics		
13	UK	Master	Islamic Studies	Al Maktoum	16 May 2008

				Foundation	
14	Ireland	PhD	Civil Engineering	University College Dublin	
15	UK	PhD		University of Strathclyde	20th June 2008
16	Germany	PostDoc	Stochastic Modeling of Cell Populations		
17	Ireland		Software Appliance Anomaly Detection		
18	Spain	PhD	BioInformatics	Rovira i Virgili university of Tarragona	May 12th, 2008
19	Netherlands	PhD	Biocatalysis	University of Groningen	
20	Australia	PhD	CFD of Biofuel Engines	University of New South Wales	31 July 2008
21	Denmark	PhD		Faculty of Engineering	May 6, 2008
22	Ireland	PhD	Optical Switching Network Modelling and Optimisation	Dublin City University	July 31st 2008
23	Belgium	PhD	Empirical Study of Social Embodied Music Interaction	Ghent University	
24	UK	PhD		University College London	October 2008
25	South Korea	Phd	Control Area	Gyeongsang National University	
26	Sweden	PhD	Mathematical Ecology	Umea University	
27	Sweden	PhD	Mathematical Statistics	Umea University	May 8, 2008
28	Ireland	PhD	Software Engineering	Lero Graduate School in Software Engineering	
29	Norway	PhD	Nanopositioning	Norwegian University of Science and Technology	June 15, 2008
30	Czech	PhD	Chemical Engineering	Institute of Chemical Technology	
%	100%	100%	100%	90%	96%

To measure the efficiency of OWIE, recall and precision of each data element is computed with respect to three possibilities: correct value (V), correct missing value (M), and wrong value (W). It can be seen from Table VIII that Recall (V) is high for all attributes except RAM because of the diversity in which RAM information is stored. The precision (V) of each attribute is 100%, which shows that whenever a value is extracted it is extracted with very high accuracy. The recall value of missing elements, Recall (M), is also 100% which shows that missing values are extracted as missing quite accurately. The precision (M) values vary for different data elements. The recall and precision values for scholarship websites (Table IX) are quite high. This is mainly due to the way the information is stored in a highly structured manner.

TABLE VIII
EXTRACTED RESULT FROM CRAIGSLIST SITE

	Processor Speed	Processor Name	RAM	Hard Drive	Screen Size
Recall(V)	90%	100%	75%	100%	92.3%
Precision(V)	100%	100%	100%	100%	100%
Recall(M)	100%	100%	100%	100%	100%
Precision(M)	83.3%	100%	75%	100%	94.7%

TABLE IX
EXTRACTED RESULT FROM SCHOLARSHIPNET SITE

	Country	Degree	Fieldof study	University	Deadline
Recall(V)	100%	96.6%	100%	88.4%	94.4%
Precision(V)	100%	100%	100%	100%	100%
Recall(M)	0%	100%	100%	100%	100%
Precision(M)	0%	100%	100%	100%	92.3%

IV. MAJOR SHORTCOMINGS OF OWIE

During the information extraction phase, OWIE picks the first occurrence of a data element matching a pattern specified in the corresponding regular expression. For example, if information about RAM is specified at two places (such as 1GB and 2x512MB), OWIE picks the first occurrence. In case of scholarships, if the ad says PhD/PostDoc then OWIE considers this ad as a PhD scholarship ad. Similarly, if multiple fields of studies are mentioned in the ad, OWIE only picks the first phrase. This is the limitation of this version of OWIE tool which can be resolve in the later version. In addition to this, currently OWIE can handle only English language alphabets. If a university name involves other characters beside the regular English alphabets, OWIE treats is as a missing value. Furthermore, some laptop ads contain information about more than one laptop but OWIE extracts only one value against each data element which could lead to incorrect information extraction.

V. Conclusion

The paper presented an ontology-based automated tool for information extraction. The tool, named OWIE, has been designed to facilitate the semantic annotation process. The typical semantic annotation process includes three main steps. Firstly, ontology is developed that describes the domain knowledge. Secondly, data is extracted through rules with the aid context words and data types available in the above mentioned ontology. Finally, semantics are assigned to the extracted data and this semantically annotated data is stored in a database. This annotated data becomes machine readable and can be use by machines for further processing.

Two case studies, a laptop selling/purchasing site and a scholarship site, were selected to analyze the performance of OWIE. The OWIE achieve high recall and precision values. Due to the unstructured and free text nature of the laptop website, OWIE does not perform as good as it could performed on structured text. The removal of the shortcomings, identified in Section IV, can further enhance the performance of OWIE. Moreover, a more exhaustive ontology specification supported by a sophisticated rule-based system can also improve its performance. The future research will focus on incorporating these enhancements.

Acknowledgment

The first author is grateful to Mr. Abdul Wajid for her help in the development of the OWIE tool.

References

[1] Adelberg, B.: NoDoSE A Tool For Semi-Automatically Extracting Structured And Semistructured Data From Text Documents. In Proceedings of the ACM SIGMOD International Conference on Management of data, Seattle Washington (1998)

[2] Antoniou, G., Harmelen, F.V.: A Semantic Web Primer. 2nd Edition. MIT Press (2004)

[3] Arocena, G.O., Mendelzon, A.O.: WebOQL: Restructuring Documents, Databases and Webs. In Proceedings of the 14th International Conference on Data Engineering, Florida (1998)

[4] Berendt, B., Hotho, A., Mladenic, D., someren, M.V., Spiliopoulou M., Stumme G.: A Roadmap for Web Mining: from Web to Semantic Web. Lecture Notes in Computer Science European Web Mining Forum (EWMF), Springer-Verlag Berlin Heidelberg (2004)

[5] Berendt, B., Hotho, A., Stumme, G.: Towards Semantic Web Mining. In Proceedings of the 1st International Semantic Web Conference (ISWC), Sardinia Italy (2002)

[6] Crescenzi, V., Mecca, G., and Merialdo, P.: RoadRunner: Towards Automatic Data Extraction From Large Web Sites. In Proceedings of the 26th International Conference on very large Data Bases, Rome Italy (2001)

[7] Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.k., Smith, R.D.: Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. Journal of Data and Knowledge Engineering, Vol.31(3), (1999) 227-251

[8] Embley, D.W., Tao, C., Liddle, S.W.: Automating the Extraction of Data from HTML Tables with Unknown Structure. Journal of Data & knowledge Engineering. Vol. 54(1), (2005) 3-28

[9] Embley D.W., Ding Y., Liddle S. W., and Vickers M.: Automatic Creation And Simplified Querying Of Semantic Web Content. In Proceedings of First Asian Semantic Conference (ASWC), Beijing China (2006)

[10] Fiumara, G.: Automatic Information Extraction from Web Sources: A Survey. In Proceedings of the Workshop between Ontologies and Folksonomies (BOF). Michigan USA (2007)

[11] Garcia-Molina, H., Hammer, J., McHugh, J.: Semistructured Data: The Tsimmis Experience. In Proceedings of First East-European Workshop on Advances in Database and Information Systems (ADBIS). St. Petersburg Russia (1997)

[12] Handschuh, S., Staab, S., Ciravegna, F.: S-CREAM Semi-automatic CREAtion of Metadata. In Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), Siguenza Spain (2002)

[13] Hieu, L.Q.: Integration of Web Data Sources: A Survey of Existing Problems. In Proceedings of 17th GI-Workshop on the Foundations of Databases, Wörlitz in Saxony-Anhalt Germany (2005) 78-82

[14] Laender, A.H.F., Ribeiro-Neto, B.A., da Silva A.S., Teixeira J.S.: A Brief Survey of Web Data Extraction Tools. In ACM SIGMOD Record, Vol. 31(2) (2002) 84-93

[15] Madhavan, J., Jeffery, S., Cohen, S., Dong, L., Ko, D., Yu, C., Halevy, A.: Web-scale Data Integration: You can only afford to Pay As You Go. In Proceedings of Third Biennial Conference on Innovative Data Systems Research (CIDR), Pacific Grove California (2007)

[16] Mika, P., Social Networks and the Semantic Web Series: Semantic Web and Beyond. Springer, (2007)

[17] Musela, I., Minton, S., Knoblock, C.: Hierarchical Wrapper Induction For Semistructured Information Sources. Journal of Autonomous Agents and Multi-Agent systems. Vol. 4(1-2) (2001) 93-114

[18] Reeve, L., Han, H : Survey of Semantic Annotation Platforms. In Proceedings of the 20th Annual ACM Symposium on Applied Computing, Web Technologies and Applications track, Santa Fe New Mexico (2005)

[19] Sahuguet, A., Azavant, F.: Building Light-Weight Wrappers for Legacy Web Data-Sources Using W4F. In Proceeding of 25th International Conference on Very Large Databases (VLDB). Edinburgh Scotland (1999)

[20] Soderland, S.: Learning Information Extraction Rules For Semi-Structured and Free Text. Machine Learning. Vol. 34 (1-3). (1999) 233–272

[21] Tang, J., Li, J., Lu, H., Liang, B., Huang, X., Wang, K.: IASA: Learning to Annotate the Semantic Web. Journal on Data Semantics. Vol. 4. (2005) 110-145

[22] Tjoa, A., Wagner, R., Andjomshoa, A., Shayeganfar, F.: Semantic Web: Challenges and New Requirements. In Proceedings. Sixteenth International Workshop on Database and Expert Systems Application (DEXA). Copenhagen Denmark (2005) 1160 – 1163

[23] Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F: MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. In Proceedings of The 13th International Conference on Knowledge Engineering and Management. Seguenza Spain (2002)

[24] Wilson, M., Matthews, B.: The Semantic Web: Prospects And Challenges. In Proceeding of 7th International Baltic Conference on Databases and Information Systems. Vilnius Lithuania (2006)

[25] Yildiz, B., Miksch, S.: Motivating ontology-driven information extraction. In Prasad, A., Madalli, D., eds.: International Conference on Semantic Web and Digital Libraries. Indian Statistical Institute Platinum Jubilee Conference Series (2007) 45–53

[26] Yildiz Burcu, Miksch Silvia. ontoX - A Method for Ontology-Driven Information Extraction. In: Computational Science and Its Applications (ICCSA 2007), LNCS 4707, Springer-Verlag, 2007, S. 660 - 673.

First A. Author (M’76–SM’81–F’87) and the other authors may include biographies at the end of regular papers. Biographies are often not included in conference-related papers. This author became a Member (M) of **IEEE** in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state or country, and year degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests ends the paragraph.

The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the **IEEE**. Finally, list any awards and work for **IEEE** committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.