International Journal of Electronics and Electical Engineering

Volume 3 | Issue 1 Article 4

July 2014

WEB SCALE INFORMATION EXTRACTION USING WRAPPER INDUCTION APPROACH

RINA ZAMBAD
Thadomal Shahani Engg. College, Bandra, Mumbai, India, rkbora2006@gmail.com

JAYANT GADGE
Thadomal Shahani Engg. College, Bandra, Mumbai, India, jayantrg@hotmail.com

Recommended Citation

ZAMBAD, RINA and GADGE, JAYANT (2014) "WEB SCALE INFORMATION EXTRACTION USING WRAPPER INDUCTION APPROACH," *International Journal of Electronics and Electical Engineering*: Vol. 3 : Iss. 1 , Article 4.
DOI: 10.47893/IJEEE.2014.1121
Available at: https://www.interscience.in/ijeee/vol3/iss1/4

WEB SCALE INFORMATION EXTRACTION USING WRAPPER INDUCTION APPROACH

RINA ZAMBAD[1] & JAYANT GADGE[2]

[1,2]Thadomal Shahani Engg. College, Bandra, Mumbai, India
E-mail:rkbora2006@gmail.com, jayantrg@hotmail.com

Abstract - Information extraction from unstructured, ungrammatical data such as classified listings is difficult because traditional structural and grammatical extraction methods do not apply. The proposed architecture extracts unstructured and un-grammatical data using wrapper induction and show the result in structured format. The source of data will be collected from various post website. The obtained post data pages are processed by page parsing, cleansing and data extraction to obtain new reference sets. Reference sets are used for mapping the user search query, which improvised the scale of search on unstructured and ungrammatical post data. We validate our approach with experimental results.

Keywords *– Wrapper induction, information extraction, post, reference set, parsing.*

I. INTRODUCTION

In the past few years, many approaches to Web Information systems, including machine learning and pattern mining techniques, have been proposed, with various degrees of automation. Hsu and Dung [5] classified wrappers into 4 distinct categories, including handcrafted wrappers using general programming languages, specially designed programming languages or tools, heuristic-based wrappers, and Web Information approaches. Chang [11] followed this taxonomy and compared Web Information systems from the user point of view and discriminated IE tools based on the degree of automation. They classified Information Extraction tools into four distinct categories, including systems that need programmers, systems that need annotation examples, annotation-free systems and semi-supervised systems.

The huge amounts of unstructured and ungrammatical data on the World Wide Web could be useful if the information contained within them could be extracted. Examples of such data sources are internet classified listings, such as those on Craigslist, classified ads, or internet forum postings.

The input file of an IE task may be structured, semi structured or free-text. As shown in Figure 1, the definition of these terms varies across research domains. *Soderland* [8] considered free-texts e.g. news article, that are written in natural languages are unstructured, postings on newsgroup (e.g. apartment rentals), medical records, equipment maintenance logs are semi-structured, while HTML pages are structured. However, from the viewpoint of database researchers, the information stored in databases is known as structured data; XML documents are semi-structured data for the schema information is mixed in with the data values, while Web pages in HTML are unstructured because there is very limited indication of the type of data.

2011 Mahindra Scorpio SLE for Sale - INR750000 (Aundh, Pune)

Date: 2011-12-03, 11:28PM IST

Reply to: sale-4fraa-2733943606@craigslist.org [Error when replying to ads?]

2011 Mahindra Scorpio SLE for Sale.
Brand new Scorpio SLE Purchased in January 2011.
White Colour.
70,025 kms done.
Full Servicing is just recently done. Rubber parts replaced. No creaking sounds or anything to complain in it !!!
Brand New Bridgestone Tyres installed in November 2011.
Music System.
AC.
Remote Locking, Power Windows, Power side-view mirrors.
Still feels like New Smooth !!! - See it to believe it !!!
Call Hemant Gokhale on 96897-31579 or on 94050-23228 to schedule viewing / test drive.
Car is in Sindh Colony, in Aundh.

Fig. 1 Post for Information Extraction

From our viewpoints, XML documents are considered as structured since there are DTD or XML schema available to describe the data. Free texts are unstructured since they require substantial natural language processing. For the large volume of HTML pages on the Web, they are considered as semi-structured [9] since the embedded data are often rendered regularly via the use of HTML tags. Thus, semi-structured inputs are the documents of a fairly regular structure and data in them may be presented

in HTML or non-HTML format. For many IE tasks, the input are pages of the same class, still some IE tasks focus on information extraction from pages across various web sites.

A. Wrapper Induction

There are two main approaches to wrapper generation: wrapper induction and automated data extraction. Wrapper induction uses supervised learning to learn data extraction rules from manually labeled training examples. The disadvantages of wrapper induction are

- The time-consuming manual labeling process and
- The difficulty of wrapper maintenance.

Wrapper Induction (WI)[1] systems are software tools that are designed to generate wrappers. A wrapper usually performs a pattern matching procedure (e.g., a form of finite-state machines), which relies on a set of extraction rules. Tailoring a WI system to a new requirement is a task that varies in scale depending on the text type, domain, and scenario. To maximize reusability and minimize maintenance cost, designing a trainable WI system has been an important topic in the research fields of message understanding, machine learning, data mining, etc. The task of Web IE which is concerned in this paper differs largely from traditional IE tasks. Traditional IE aims at extracting data from totally unstructured free texts that are written in natural language. Web IE, in contrast, processes online documents that are semi-structured and usually generated automatically by a server-side application program. As a result, traditional IE usually takes advantage of NLP techniques such as lexicons and grammars, whereas Web IE usually applies machine learning and pattern mining techniques to exploit the syntactical patterns or layout structures of the template-based documents.

B. Constructing Reference Sets

It's not necessary to construct the reference set manually. The machine can construct its own reference set, which it can be use for extraction and matching using the algorithms. For example, suppose an auto-parts company has an internal database of parts for cars. This company might want to join their parts database with the classified ads about cars in an effort to advertise their mechanics service. Therefore, they could use reference-set based extraction to identify the cars for sale in the classified ads, which they can then join with their internal database.

Reference set can be generated automatically from the posts themselves, which can then be used for extraction. Nonetheless, manual reference sets provide a stronger intuition into the strengths of reference-set based information extraction. Regardless of whether the reference set is constructed manually or automatically, the high-level procedure for exploiting a reference set for extraction is the same. To use the reference sets, the methods in this thesis first find the best match between each post and the tuples of the reference set. By matching a post to a member of the reference set, the algorithm can use the reference set's schema and attribute values as semantic annotation for the post (a single, uniform value for that attribute).

Internal database might have its own internal identification attribute and since this attribute would be needed to join the cars identified in the classifieds with their database, it would need to be included as annotation for the post. In this case, since none of the classified ads would already contain this attribute since it's specific to the company, it would not be possible to include it as an automatically constructed attribute, and therefore the company would have to manually create their own reference set to include this attribute. So, in this case, the user would manually construct a reference set and then use it for extraction.

C. Our Contribution

The key contribution of the project is for information extraction that exploits reference sets, rather than grammar or structure based techniques. The project includes the following contributions:

- An automatic learning system for matching and extraction of reference set.
- A method that selects the appropriate reference sets from a repository and uses them for extraction and annotation without training data.
- An automatic method for constructing reference sets from the posts themselves.
- An automatic method for web post record extraction using reference set for searching accurate information.

II. LITERATURE REVIEW

The World Wide Web has more and more online web databases which can be searched through their web query interfaces. Deep web contents are accessed by queries submitted to web databases and the returned data are enwrapped in dynamically generated web pages. Often the retrieved information (query results) is enwrapped in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index by traditional crawler based Search engines, such as Google and Yahoo. Wei Liu, Xiaofeng Meng, and Weiyi Meng[3] call this kind of special Web pages as deep Web pages. With the flourish of the deep Web, users have a great opportunity to benefit from such abundant information in it. In general, the desired information is embedded in the deep Web pages in the form of data records returned by Web databases when they respond to users' queries. Therefore, it is an

important task to extract the structured data from the deep Web pages for later processing.

Template designers of deep Web pages always arrange the data records and the data items with visual regularity to meet the reading habits of human beings. In this paper, author explore the visual regularity of the data records and data items on deep Web pages and propose a novel vision-based approach, Vision-based Data Extractor (ViDE), to extract structured results from deep Web pages automatically. ViDE is primarily based on the visual features human users can capture on the deep Web pages while also utilizing some simple nonvisual information such as data types and frequent symbols to make the solution more robust. However, there are still some issues like ViDE can only process deep Web pages containing one data region, while there is significant number of multidata-region deep Web pages.

Nam-Khanh Tran; Kim-Cuong Pham; Quang-Thuy Ha [2] presents XPath-Wrapper Induction_algorithm which leverages user queries and template-based sites for extracting structured information. Template-based sites have common properties. For example, pages within a web site have a similar structure, thus attribute values occur at fixed positions within pages. The fix positions can be defined as paths from the root to the node called XPath containing attribute values on DOM tree of the pages.

Once attribute values in a few pages of a site are identified, one can infer their positions, and then use these to extract attribute values from remaining pages of the site. However, the task of extracting attribute values from these sites remains challenging as pages within site have a lot of spurious information such as advertisement information or dynamic user generated comments, etc. This paper restricts for only detail web pages.

An Information Extraction (IE) task is defined by its input and its extraction. The input can be unstructured documents like post text that are written in natural language (e.g. Fig -1) or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. The extraction target of an IE task can be a relation of k-tuples (where k is the number of attributes in a record) or it can be a complex object with hierarchically organized data. For some IE tasks, an attribute may have zero (missing) or multiple instantiations in a record. The difficulty of an IE task can be further complicated when various permutations of attributes or typographical errors occur in the input documents.

TSIMMIS is the approach that gives a framework for manual building of Web wrappers [9]. The main component of this project is a wrapper that takes as input specification file that declaratively states (by a sequence of commands given by programmers) where the data of interest is located on the pages and how the data should be "packaged" into objects. Each command is of the form: [variables, source, pattern], where source specifies the input text to be considered, pattern specifies how to find the text of interest within the source, and variables are a list of variables that hold the extracted results. The special symbol '*' in a pattern means discard, and '#' means save in the variables.

W4F (Wysiwyg Web Wrapper Factory) is a Java toolkit to generate Web wrappers [9]. The wrapper development process consists of three independent layers: retrieval, extraction and mapping layers. In the retrieval layer, a to-be processed document is retrieved (from the Web through HTTP protocol), cleaned and then fed to an HTML parser that constructs a parse tree following the Document Object Model (DOM). In the extraction layer, extraction rules are applied on the parse tree to extract information and then store them into the W4F internal format called Nested String List (NSL).

XWrap is a system that exploits formatting information in Web pages to hypothesize the underlying semantic structure of a page [9]. It encodes the hypothetical structure and the extraction knowledge of the web pages in a rule-based declarative language designed specifically for XWrap. The wrapper generation process includes two phases: structure analysis, and source-specific XML generation. In the first phase, XWrap fetches, cleans up, and generates a tree-like structure for the page.

In addition to the categorization by input documents, an IE task can be classified according to the extraction target. For example, Sarawagi classified HTML wrappers into record-level, page-level and site-level IE tasks. Record-level wrappers discover record boundaries and then divide them into separate attributes; page-level wrappers extract all data that are embedded in one Web page, while site-level wrappers populate a database from pages of a Web site, thus the attributes of an extraction object are scattered across pages of a Web site. Academic researchers have devoted much effort to develop record-level and page-level data extraction, whereas industrial researchers have more interest in complete suites, which support site-level data extraction.

III. PROPOSED ARCHITECTURE

System architecture defines the system major component interaction model.

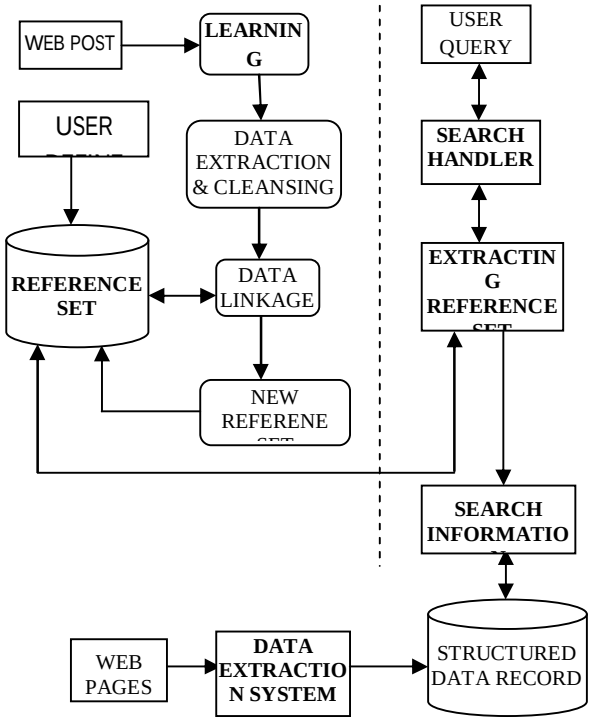

Fig. 2 System Architecture

Architecture in fig. 2 describes the component interaction for web-scale information extraction using unstructured and un-grammatical data. The source of data will be collected from various post website. The obtained post data page will be processed by page parsing, cleansing and data extraction to obtain new reference sets. Reference sets are used for mapping the user search query, which improvised the scale of search on unstructured and ungrammatical post data. To extract the search information it build a record database of post data on various web sites.

The system can be divided into three subsystems as Learning system, Data Extraction System and User Search Query System.

A learning system is responsible for learning a new set of extraction rules for specific sites. A single web site may contain pages conforming to multiple different templates, from each website all samples of pages are collected and are clustered using Shingle based signature which is computed for each web page based on html tags.

In Extraction system, the learnt rules are applied to the stream of crawled wed pages to extract records from them. For each incoming web page, the shingle based signature and page URL are used to find the matching rule for the page, which is then applied to extract the record for the page.

A Search Query System, are used to search matching records based user query. For each request query will be matched based on the rules of learning system.

A. LEARNING SYSTEM

1. Grouping Similar pages

A large set of similar structure web pages will be grouped from the website. Although Web pages within a cluster, to a large extent, have similar structure, they also exhibit minor structural variations because of optional, disjunctive, extraneous, or styling sections. To ensure high recall at low cost, we need to ensure that the page sample set that is annotated has a small number of pages and captures most of the structural variations in the cluster. One way to create a relational data set from the posts is to define a schema and then fill in values for the schema elements using techniques such as information extraction. This is sometimes called semantic annotation.

2. Extracting Reference Sets

Extracting Reference Sets implements the approach to creating relational data sets from unstructured and ungrammatical posts. A reference set consists of collections of known entities with the associated, common attributes. A reference set can be an online (or offline) set of reference documents.

First label each token with a possible attribute label or as “junk" to be ignored. After all the tokens in a post are labeled, then clean each of the extracted labels. To begin the extraction process, the post is broken into tokens. Using the post as an example, set of tokens becomes, {“11", “civic", “5speed"...}. Each of these tokens is then scored against each attribute of the record from the reference set that was deemed the match.

To score the tokens, the extraction process builds a vector of scores, V_{IE}. V_{IE} is composed of vectors which represent the similarities between the token and the attributes of the reference set.

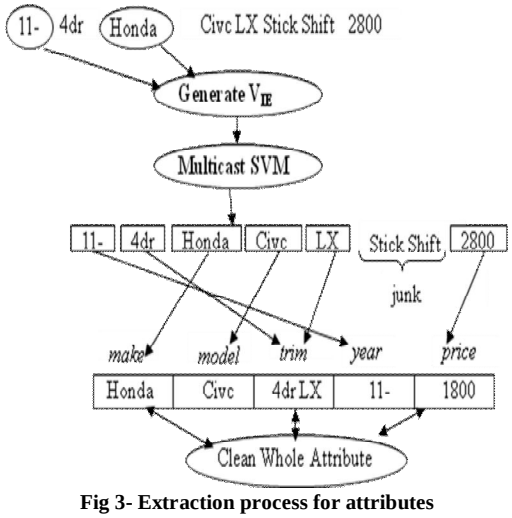

Fig 3- Extraction process for attributes

Each V_{IE} is then passed to a structured SVM(Support Vector Machine), trained to give it an attribute type label, such as make, model, or price.

Since there are many irrelevant tokens in the post that should not be annotated, the SVM learns that any V_{IE} that does associate with a learned attribute type should be labeled as “junk", which can then be ignored. Without the benefits of a reference set, recognizing junk is difficult because the characteristics of the text in the posts are unreliable.

3. Creating Relational data

To use a reference set to build a relational data set, exploit the attributes in the reference set to determine the attributes from the post that can be extracted. The first step is to find the best matching member of the reference set for the post. This is called the “record linkage" step. By matching a post to a member of the reference set schema elements for the post can be defined by using the schema of the reference set, and standard attributes for these attributes are provided by using the attributes from the reference set when a user queries the posts.

Next, perform information extraction to extract the actual values in the post that match the schema elements defined by the reference set. This step is the information extraction step. During the information extraction step, the parts of the post are extracted that best match the attribute values from the reference set member chosen.

B. EXTRACTION SYSTEM

Extracting Web Page Records implementation based on the learnt rules are applied to the stream of crawled wed pages to extract records from them. For each incoming web page, the shingle based signature and page URL are used to find the matching rule for the page, which is then applied to extract the record for the page. The extracted record will be stored in the database for user query search.

C. SEARCH QUERY SYSTEM

User Search Query System implementation for user to search matching records based user query input. For each request query will be matched based on the rules of learning system. User input may not be in appropriate syntax or semantics, the system do an auto correction of the input using learning system data set and pose an appropriate query for search.

IV. EXPERIMENTAL RESULTS

This section presents results for unsupervised approach to selecting a reference set, finding the attributes and exploiting these attributes for extraction. Before examining the results we describe the post and reference set used in testing.

Various data files are gathered from two different websites to evaluate the accuracy of query and performance of the system implemented. The input file should be html files.

To measure accuracy of query and performance, compared it, in terms of query error rates and accuracy results obtained by running the implemented system on the gathered data from various sites. The system first extract the reference sets required for query correction using Learning System mechanism, and then runs the web data records process to build web link and URL database for user query matching. In last it runs user query system where user poses a query input for searching required contents.

1. Post Sets

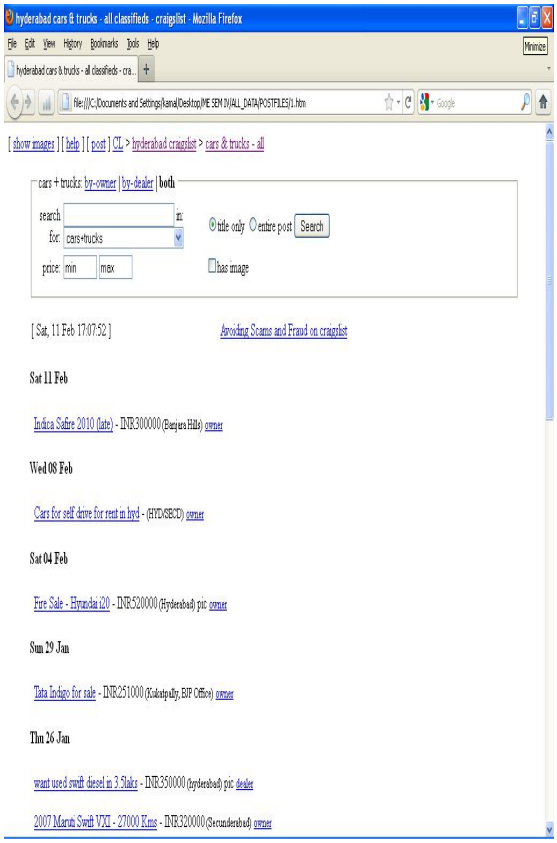

Fig. 4 Post from Craigslist

Choose the set of posts to test different cases that exist for finding the appropriate reference sets. One set of post matches only a single reference set in our collection. It contains 500 posts from Craigslist classifieds and Clickindia classified ads for cars.

2. Reference sets

For this project references set for cars is used. In this cars make, model, price and location are considered. At the same time if there is any mistake in name of the cars make or model name it corrects it and then save into the reference set. For example, for ‘MARUTI’ object the reference sets may be ‘MARUTHI’ or for ‘HYUNDAI’ it may be ‘HUNDAI’ or ‘HUNDAIE’. Similarly It shows a like a word ‘TEN’ which can be reference of ‘ZEN’.

For example, 'SANTRO' may be 'SANTO' or 'CITY' may be 'CITI'.

Results Messages

	Object_id	Object_Name
1	1	TATA
2	2	HONDA
3	3	MARUTI
4	4	VOLKSWAGEN
5	5	MAHINDRA
6	6	HYUNDAI
7	7	FORD
8	8	FIAT

Object_Master

Results Messages

	Ref_Object_id	Ref_Object_Name
1	2	HANDA
2	1	TAT
3	3	MARUTHI
4	6	HUNDAI
5	7	FOR
6	5	MAHENDRA
7	4	VOLKSWAGE
8	6	HUNDAIE
9	8	FEAT

Ref_Object_Master

Fig. 5 Data Reference and Reference Sets

Once the reference set is created it extracts the record from the web pages and web-page-record table is updated which contains page-data, ext-data, link, price and location as shown in fig 6.

2.Data in Table 'Web_Data_Record' in 'RefdataDB' on 'PRSLAP'

page_data	ext_data	link_data	price	location
FORD DOON GIVT PETROL	FORD DOON GURGAON	http://delhi.craigslist.co.in/cto/2845822298.html	INR120000	GURGAON
CONDITION HONDA CITY AUTOMATIC	HONDA CITY NEW DELHI	http://delhi.craigslist.co.in/cto/2836732662.html	INR400000	NEW DELHI
HONDA CITY AUTOMATIC	HONDA CITY NEW DELHI	http://delhi.craigslist.co.in/cto/2836574594.html	INR140000	NEW DELHI
2005 HONDA CITY EXT REALLY CONDITION	HONDA CITY PITAMPURA, NEW DELHI	http://delhi.craigslist.co.in/cto/2834012502.html	INR298000	PITAMPURA, NEW DELHI
HYUNDAI SONATA-DYNERA SECOND HAND	HYUNDAI SONATA AT DELHI NCR	http://delhi.craigslist.co.in/cto/2832328963.html	INR450000	DELHI NCR
BUY MARUTI 800 CAR SECOND HAND MARU	MARUTI SUZUKI HYDERABAD	http://delhi.craigslist.co.in/cto/2832083870.html	INR160000	HYDERABAD
EXCELLENT CONDITION HONDA CITY AUTO	HONDA CITY (NEW DELHE-ANAND NIKETA	http://delhi.craigslist.co.in/cto/2829103265.html	INR400000	(NEW DELHI-ANAND NIKETAN
VW VENTO AUTOMATIC	VOLKSWAGEN VENTO VASANT VIHAR	http://delhi.craigslist.co.in/cto/2825154150.html	INR892000	VASANT VIHAR
TOP TEN CARS	MARUTI ZEN NEW DELHI	http://delhi.craigslist.co.in/cto/2797469706.html	INR1	NEW DELHI
INDICA SAFIRE 2010 (LATE)	TATA INDICA BANJARA HILLS	http://hyderabad.craigslist.co.in/cto/2845698242.html	INR300000	BANJARA HILLS
FIRE HYUNDAI I20	HYUNDAI I20 HYDERABAD	http://hyderabad.craigslist.co.in/cto/2833861742.html	INR520000	HYDERABAD
TATA INDIGO	TATA INDIGO KUKATPALLY, BJP OFFICE	http://hyderabad.craigslist.co.in/cto/2823797620.html	INR251000	KUKATPALLY, BJP OFFICE
WANT USED SWIFT DIESEL IN SLAWS	MARUTI SWIFT HYDERABAD	http://hyderabad.craigslist.co.in/cto/2818874605.html	INR350000	HYDERABAD
2007 MARUTI SWIFT VXI 27000 KMS	MARUTI SWIFT VXI SECUNDERABAD	http://hyderabad.craigslist.co.in/cto/2818602054.html	INR320000	SECUNDERABAD
IMMEDIATE WAGON R, LXI	MARUTI WAGON R LXI HYDERABAD	http://hyderabad.craigslist.co.in/cto/2808544990.html	INR300000	HYDERABAD
MANZA RENTAL DRIVER	TATA MANZA HYDERABAD	http://hyderabad.craigslist.co.in/cto/2798331393.html	INR1200	HYDERABAD
SANTRO XING ZIP PLUS	HYUNDAI SANTRO HITEC CITY	http://hyderabad.craigslist.co.in/cto/2794351826.html	NA	HITEC CITY
HYUNDAI ACCENT GVS PETROL EXCELLENT	HYUNDAI ACCENT BHEL	http://hyderabad.craigslist.co.in/cto/2780466531.html	INR140000	BHEL
HONDA CIVIC BY OWNER	HONDA CIVIC CIVIC MUMBAI	http://mumbai.craigslist.co.in/cto/2839656207.html	INR680000	MUMBAI
SANTRO 2002 CNG	HYUNDAI SANTRO MUMBAI	http://mumbai.craigslist.co.in/cto/2797531357.html	INR150000	MUMBAI
ZEN CAR LXI 2004	MARUTI ZEN BANDRA RECLAMATION	http://mumbai.craigslist.co.in/cto/2836868949.html	INR130000	BANDRA RECLAMATION
HYUNDAI I20 LESS THAN YEAR OLD	HYUNDAI I20 MARINE LINES, MUMBAI	http://mumbai.craigslist.co.in/cto/2797471534.html	INR350000	MARINE LINES, MUMBAI
2004 TATA INDICA	TATA INDICA KHARGHAR, NAVRANG	http://mumbai.craigslist.co.in/cto/2794193467.html	INR185000	KHARGHAR, NAVRANG
MARUTI ALTO 2008	MARUTI ALTO ALTO THANE, MAHARASHTR	http://mumbai.craigslist.co.in/cto/2780435994.html	INR210000	THANE, MAHARASHTRA
2007 ALTO, SINGLE OWNER DRIVEN, SPARE	MARUTI ALTO GURGAON	http://delhi.craigslist.co.in/cto/2842412512.html	INR195000	GURGAON
BUY HYUNDAI GETZ CARS SECOND HAND H	HYUNDAI GETZ KOLKATA	http://delhi.craigslist.co.in/cto/2832327109.html	INR17999	KOLKATA
USED MARUTI ESTEEM CARS USED MARUTI	MARUTI SUZUKI CHENNAI	http://delhi.craigslist.co.in/cto/2828775483.html	INR280000	CHENNAI
BUY MAHINDRA SCORPIO CARS SECOND H	MAHINDRA SCORPIO KOLKATA	http://delhi.craigslist.co.in/cto/2827096688.html	INR999999	KOLKATA

Fig. 6 Web Data Records data Structure

If the query is given it performs keyword search and give the appropriate results. Whenever click on a particular link it will shows the actual web page from where that data is extracted.

V. CONCLUSION AND FUTURE WORK

Keyword search over semi-structured and structured data offers users great opportunities to explore better-organized data. Our approach, reference-set based extraction exploits a reference set. By using reference sets for extraction, instead of grammar or structure, our technique free the assumption that posts require structure or grammar. This project investigates information extraction from unstructured, ungrammatical text on the Web such as web postings. Since the data is unstructured and ungrammatical, this information extraction precludes the use of rule-based methods that rely on consistent structures within the text or natural language processing techniques that rely on grammar. Our work describes extraction using a reference set, which define as a collection of known entities and their attributes. The project implements an automatic technique to provide a scalable and accurate approach to extraction from unstructured, ungrammatical text. The machine learning approach provides even higher accuracy extractions and deals with ambiguous extractions, although at the cost of requiring human effort to label training data. The results demonstrate that reference-set based extraction outperforms the current state-of-the-art systems that rely on structural or grammatical clues, which is not appropriate for unstructured, ungrammatical text. Reference-set based extraction from unstructured, ungrammatical text allows for a whole category of sources to be queried, allowing for their inclusion in data integration systems that were previously limited to structured and semi-structured sources.

Textual characteristics of the posts make it difficult to automatically construct the reference set. One future topic of research is a more robust and accurate method for automatically constructing reference sets from data when the data does not fit the criteria for automatic creation. This is a larger new topic that may involve combining the automatic construction technique in this thesis with techniques that leverage the entire web for extracting attributes for entities. Along these lines, in certain cases it may simply not be possible for an automatic method to discover a reference set.

REFERENCES

[1] Gulhane, P.; Madaan, A.; Mehta, R.; Ramamirtham, J.; Rastogi, R.; Satpal, S.; Sengamedu, S.H.; Tengli, A.; Tiwari, C.; Web-scale_information_extraction_with_vertex. *Data Engineering (ICDE), 2011 IEEE 27th International Conference on_Digital Object Identifier* Publication Year: 2011, Page(s): 1209 - 1220

[2] Nam-Khanh Tran; Kim-Cuong Pham; Quang-Thuy Ha; XPath-Wrapper Induction for Data Extraction *Asian Language Processing (IALP), 2010 International Conference on Digital Object Identifier:* Publication Year: 2010 , Page(s): 150 - 153.

[3] Wei Liu; Xiaofeng Meng; Weiyi Meng;_ViDE: A Vision-Based Approach for Deep Web Data Extraction *Knowledge and Data Engineering, IEEE Transactions on Volume: 22* Publication Year:2010, Page(s): 447 – 460

[4] Matthew Michelson michelso@isi.edu, Craig A. Knoblock knoblock@isi.edu University of Southern California Information Sciences Institute; Creating Relational Data from Unstructured and Ungrammatical Data *Journal of Arti_cial Intelligence Research 31 (2008),* Page(s):543-590

[5] Chang, C.-H., Hsu, C.-N., and Lui, S.-C. Automatic information extraction from semi-Structured Web Pages by pattern discovery. *Decision Support Systems Journal*, 35(1): 129-147, 2003

[6] Matthew Michelson; Craig A. Knoblock Unsupervised Information Extraction from Unstructured, Ungrammatical Data Sources on the World Wide Web; *International Journal of Document Analysis and Recognition (*2007); Page(s) 1-13

[7] Matthew Michelson; Craig A. Knoblock Constructing Reference Sets from Unstructured, Ungrammatical Text; *Journal of Artificial Intelligence Research* (2010) Page(s):189-221

[8] Soderland, S., Learning information extraction rules for semi-structured and free text. *Journal of Machine Learning*, 34(1-3): 233-272, 1999.

[9] Laender, A. H. F., Ribeiro-Neto, B., DA Silva and Teixeira, A brief survey of Web data extraction tools. *SIGMOD Record 31(2):* 84-93, 2002.

[10] Arocena, G.O.; Mendelzon, A.O. WebOQL: restructuring documents, databases and Webs; *Data Engineering, 1998. Proceedings. 14th International Conference;* Publication Year: 1998, Page(s): 24 – 33

[11] Chang, C.-H.; Kayed, M.; Girgis, R.; Shaalan, K.F. ; A survey of Web Information Extraction System; *Knowledge and Data Engineering, IEEE Transactions Volume: 18, Issue: 10* ; Publication Year: 2006 , Page(s): 1411 – 1428