The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 6 || Pages || 36-45 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Scaling the Information Extraction from Unstructured and Ungrammatical Data Sources on Web

[1]Madhavi. K. Sarjare, [2]S.L.Vaikole
[1]*Department of Computers, Datta Meghe College of Engineering, Airoli, Navi Mumbai.*
[2]*Department of Computers, Datta Meghe College of Engineering, Airoli, Navi Mumbai.*

----------ABSTRACT----------

Information Extraction (IE) on the web is the task of automatically extracting knowledge from text. Web Information Extraction (WIE) systems have recently been able to extract massive quantities of relational data from online text. This massive body of text which are now available on the World Wide Web do presents an unparalleled opportunity for information extraction. However, this information extraction on the Web is challenging due to the vast variety of distinct concepts and structured expressed. The explosive growth and popularity of the worldwide web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity, diversity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching.

Information extraction is done from unstructured and ungrammatical text on the Web. They can be classified Ads, Auction listings, and web postings forums. As the data is unstructured and ungrammatical, this information extraction precludes the use of rule-based methods that rely on consistent structures within the text. It can be natural language processing techniques that rely on grammar. Posts are full of useful information, as defined by the attributes that compose the entity within the post.

Currently accessing the data within posts does not go much beyond the search of keywords. This is in particular because the unstructured and ungrammatical nature of posts makes the extraction difficult, so many a times the attributes remain hidden or embedded within the posts. Also these data sources are ungrammatical, since they do not conform to the exact rules of written language. Therefore, Natural Language Processing (NLP) based information extraction techniques are inappropriate. The ability to process and understand this information becomes more crucial, as more and more information comes online.

Also Data integration attacks this problem by letting users query heterogeneous data sources within a unified query framework; it can be combining the results to make understanding easily. However, while data integration can integrate data from structured sources such as databases, Web Services and even semi-structured sources such as that extracted data from Web pages, this leaves out the large class of useful information- Unstructured and Ungrammatical data sources.

Thus we proposed a system based Machine Learning technique to obtain the data records which are structured, from different unstructured and non-template based websites. Thus the proposed approach will be implemented by collection of known entities and their attributes, which are been referred as "reference set," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of web pages. Also it can be constructed automatically from the unstructured, ungrammatical text itself. Thus this project implements methods to exploit the reference sets for extraction using the machine learning techniques. The machine learning approach provides exact and higher accuracy extractions and also deals with ambiguous extractions, although at the cost of requiring human effort to label training data.

KEYWORDS - *Natural Language Processing, Reference set, Nested String List, Hypertrees.*

Date of Submission: 21st May 2014 Date of Publication: 10th June 2014

I. INTRODUCTION

The Internet provides access to numerous sources of useful information which are in the form of text. In recent times, there has been much interest in building systems that gather such information on a user's behalf. But because people format such information resources for use, mechanically extracting their content is not easy.

Systems using such resources typically use hand-coded *wrappers*, i.e. customized procedures for information extraction. Information extraction from unstructured, ungrammatical text on the Web such as classified ads, auction listings, and forum postings is a tough work. 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, Since the data is unstructured and ungrammatical,Posts data consists of useful information, as defined by the attributes that compose the entity within the post. For example, consider certain posts about cars from the online classified service. Each used car for sale is composed of attributes that define this car; and if we could access the individual attributes then we could include such sources in data integration systems, and answer the interesting queries. Such a query might require combining the structured database of safety ratings with the posts of the classified ads and the car review websites.

However, currently accessing the data within posts does not go much beyond search of keywords. This is specifically because the ungrammatical, unstructured nature of posts makes extraction difficult, so the attributes remain entrenched within the posts. These data sources are ungrammatical, since they do not conform to the proper rules of written language. Therefore, Natural Language Processing (NLP) based information extraction techniques are not suitable. Also the posts are unstructured since the structure can differ vastly between each listing. So, wrapper based extraction techniques will also not work either. Even if one can extract the data from within posts, you would need to assure that the extracted values map to the same value for accurate querying.

II. LITERATURE SURVEY

Existing Systems

The combination of various input documents and also variation of extraction can cause different degrees of task difficulties. Since various Information Extraction systems are designed for various IE tasks, it is not fair to compare them directly. However, analyzing what task an IE system targets and how it performs the task, can be used to evaluate this system and possibly extend to other task domains.

- TSIMMIS is one of the first approaches that gives a skeleton for manual building of Web wrappers [7]. The main component of this project is a wrapper that takes as input a specification file that declaratively states where the data of interest is located on the pages and how the data should be "packaged" into objects states (by a sequence of commands given by programmers). Each of this command is having the form of: [*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.
- WebOQL is a functional language that can be used as a query language on the Web, for semi structured data and restructuring website as well[6]. The main data structure provided by WebOQL is the *hypertree*. Hypertrees are arc-labeled ordered trees which can be used to model a relational table, a Bibtex file, a directory hierarchy, etc. The abstraction level of the data model is suitable to support collections, nesting, and ordering.
- W4F (Wysiwyg Web Wrapper Factory) is a Java toolkit to generate Web wrappers [8]. 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).

This project objective is to exploit those reference sets for extraction using both automatic techniques and machine learning techniques. The automatic technique provides 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 actually not appropriate for unstructured, ungrammatical text. Even the fully automatic case, which constructs its own reference set for automatic extraction, is competitive with the current state-of-the-art techniques that require labeled data. 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.

III. PROBLEM DEFINATION

The ability to process and understand this information which comes online becomes more and more crucial. While data integration can integrate data from structured sources such as databases, semi-structured sources such as that extracted from Web pages, and even Web Services, which ultimately leaves out a large class of useful information - Unstructured and Ungrammatical data sources. The task is to identify such unstructured, ungrammatical data as "posts". Posts are ranging in source from classified ads, forum postings, and auction listings to blog titles or paper references. The aim of the project is to structure the sources of posts, such that they can be queried and included in data integration systems.

IV. PROPOSED METHOD

To design Web-Scale Information Extraction Using Wrapper Induction Approach for an unstructured web data the following systems will be considered:

1. Learning System
2. Data Extraction System
3. User Search Query System

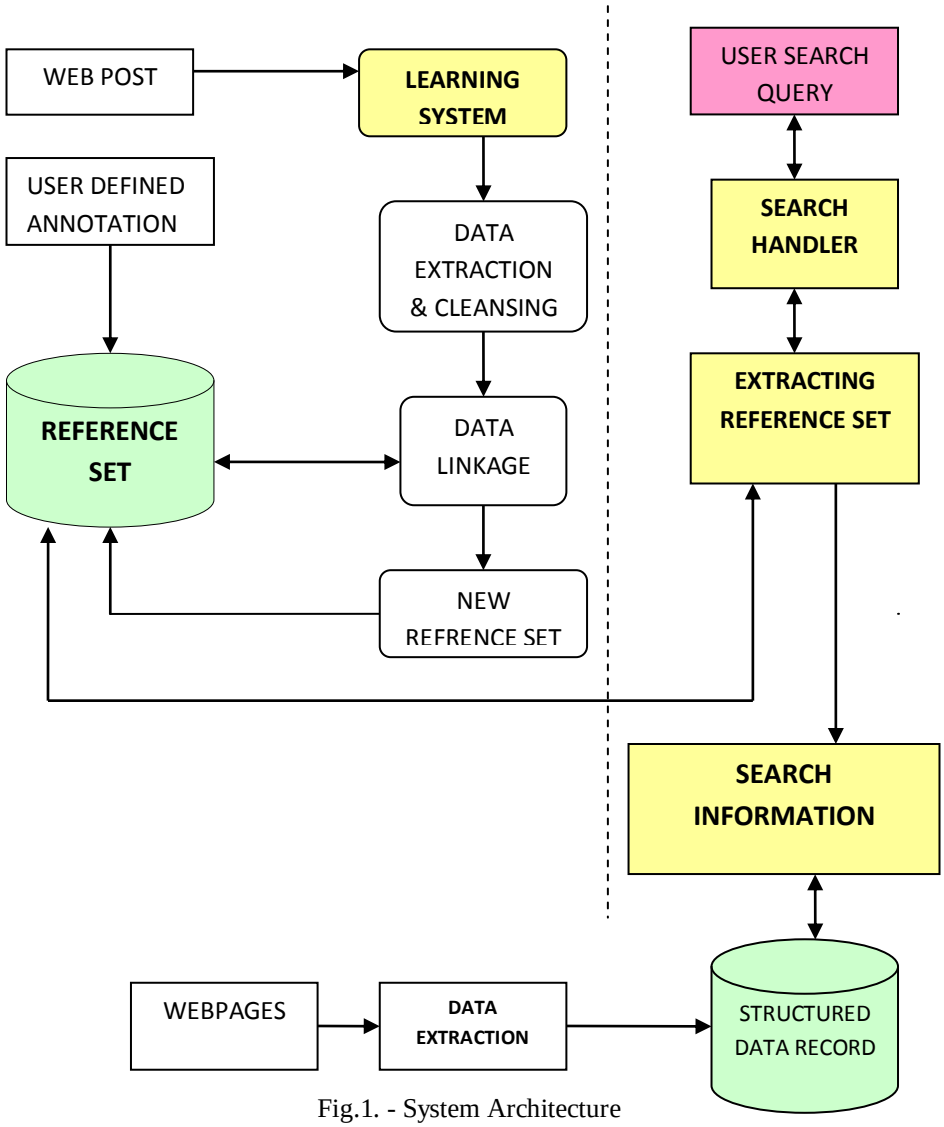


Fig.1. - System Architecture

1. Learning 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.

2. Data Extraction System

An 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.

3. User Search Query System

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.

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.

The proposed approach will be implemented by collection of known entities and their attributes, which refer as "reference set," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of Web pages. A reference set can even be constructed automatically from the unstructured, ungrammatical text itself. It follows the following methodology for information extraction from unstructured, ungrammatical data sources:

1. Automatically Choosing the Reference Sets
2. Matching Posts to the Reference Set
3. Extraction using reference sets
4. Automatically Constructing Reference Sets for Extraction
5. A Learning Approach to Reference-Set Based Extraction
6. Extracting Data from unstructured data sources.

V. RESEARCH ELABORATIONS

5.1 Implementation Methodology

Web browsers contain implementations of World Wide Web Consortium-recommended specifications, and software development tools contain implementations of programming languages.

5.1.1. 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.

5.1.2. Extracting Reference Sets:

Extracting Reference Sets implements the approach to creating relational data sets from unstructured and ungrammatical posts exploits reference sets. 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 we 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 first post from Table 1 as an example, set of tokens becomes, {"93", "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.

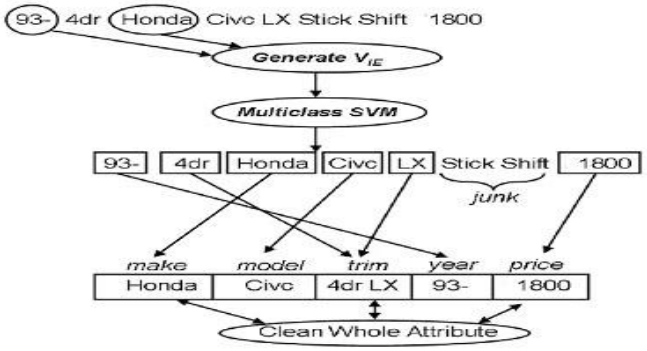


Fig 5.1: Extraction process for attributes

V_{IE}=<common scores ("civic"),
IE scores ("civic", "Honda"),
IE scores ("civic", "Civic"),
IE scores ("civic", "1993")>

More generally, for a given token, V_{IE} looks like:
V_{IE}=<common scores (token),
IE scores (token, $attribute_1$),
IE scores (token, $attribute_2$)
. . . ,
IE scores (token, $attribute_n$)>

Each V_{IE} is then passed to a structured SVM, 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.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 we can define schema elements for the post using the schema of the reference set, and we can provide standard attributes for these attributes by using the attributes from the reference set when a user queries the posts.

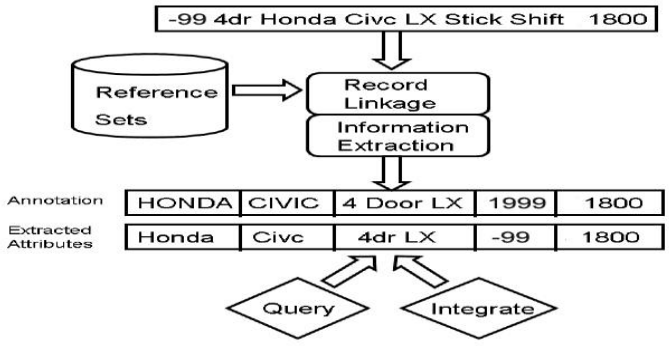


Fig 5.2- Creating relational data from unstructured sources

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.

First determine the set of matching rules for the page based on the page URL. The final rule is subsequently chosen based on the page shingle vector.

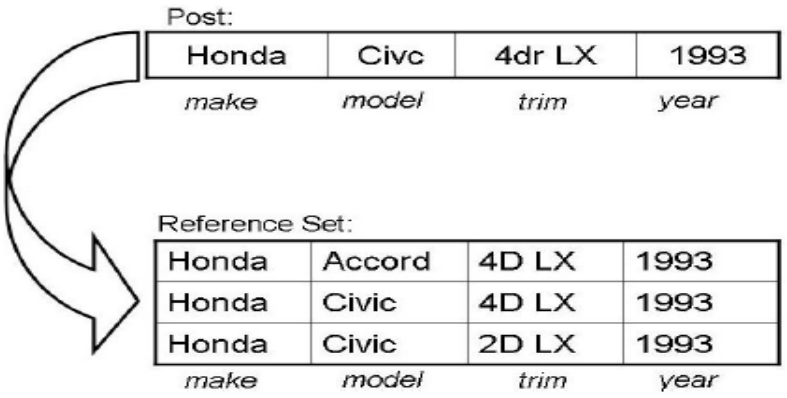


Fig 5.3- Matching a post to the reference set

5.1.3. Extracting Web Page Records:

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.

5.1.4. User Search Query:

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.

VI. RESULTS AND DISCUSSION

We will evaluate the accuracy of query and performance of the system we implemented the various data files gathered for 3 different websites. The input file will be html files.

To measure accuracy of query and performance, we compared, in terms of query error rates and accuracy results obtained by running our 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 we runs user query system where user pose a query input for searching required contents.

6.1 Extracting Reference Sets

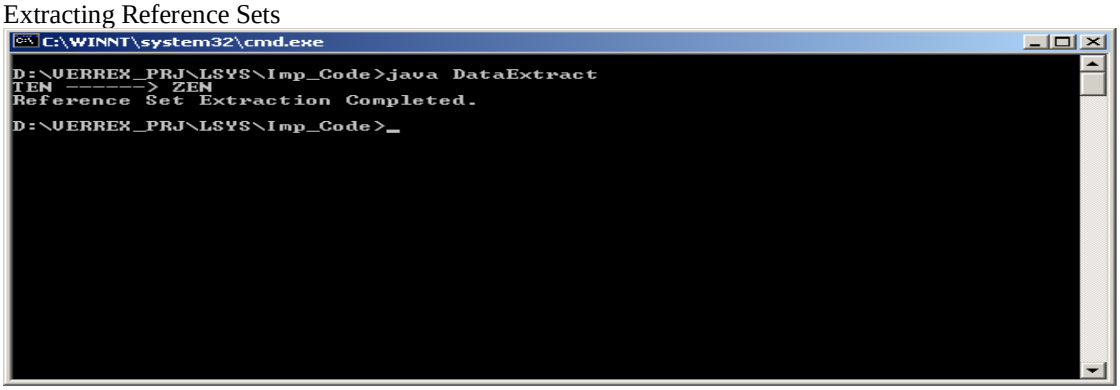


Fig – 6.1 - Extracting Reference Sets

The Fig 6.1 shows the extracting reference sets process. It shows a like a word 'TEN' which can be reference of 'ZEN'. Similarly it build a database of reference sets for minimize the query error rates.

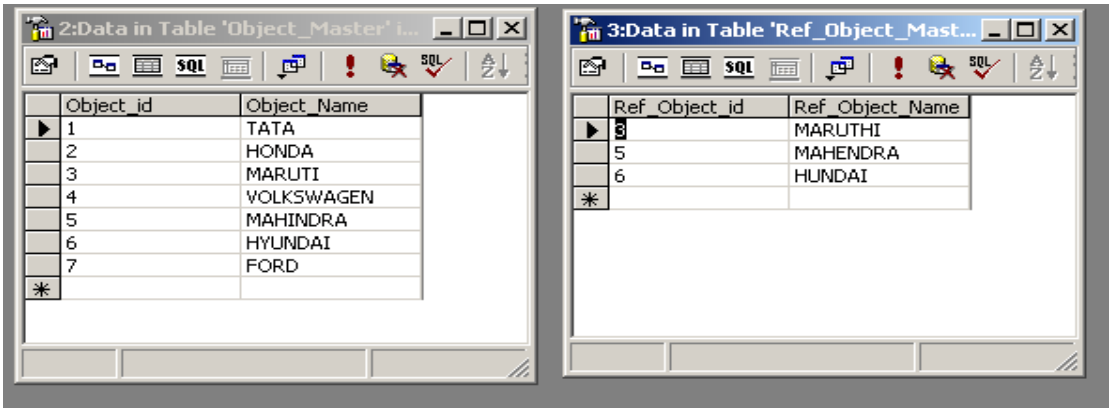


Fig – 6.2 – Data Reference and Reference Sets

The Fig 5.2 and 5.3 represent the data structure of the data reference and data reference sets. In Fig 5.2 it shows the main object reference objects data and its reference sets objects. For example, for 'MARUTI' object the reference sets may be 'MARUTHI' or for 'HYUNDAI' it may be 'HUNDAI' or 'HUNDAIE'. Similarly, for object models reference and sets are shown in Fig 6.3. For example, 'SANTRO' may be 'SANTO' or 'CITY' may be 'CITI'.

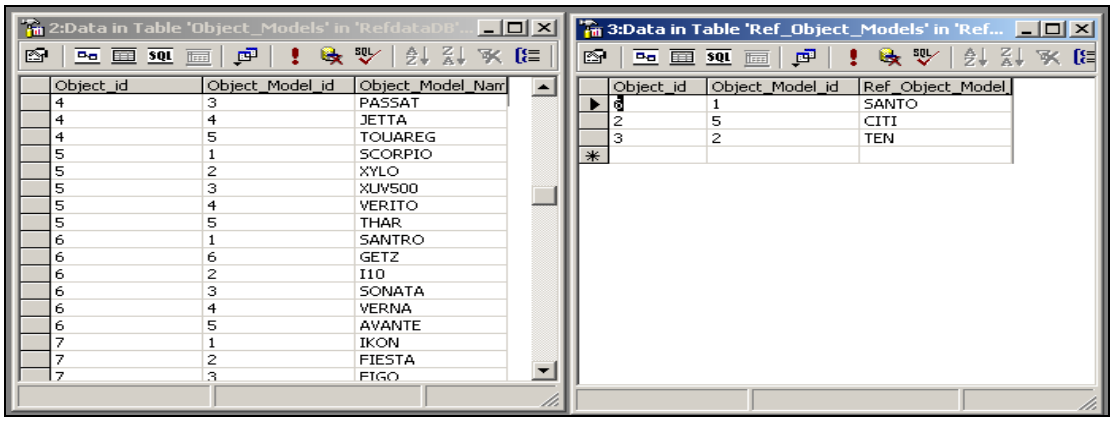
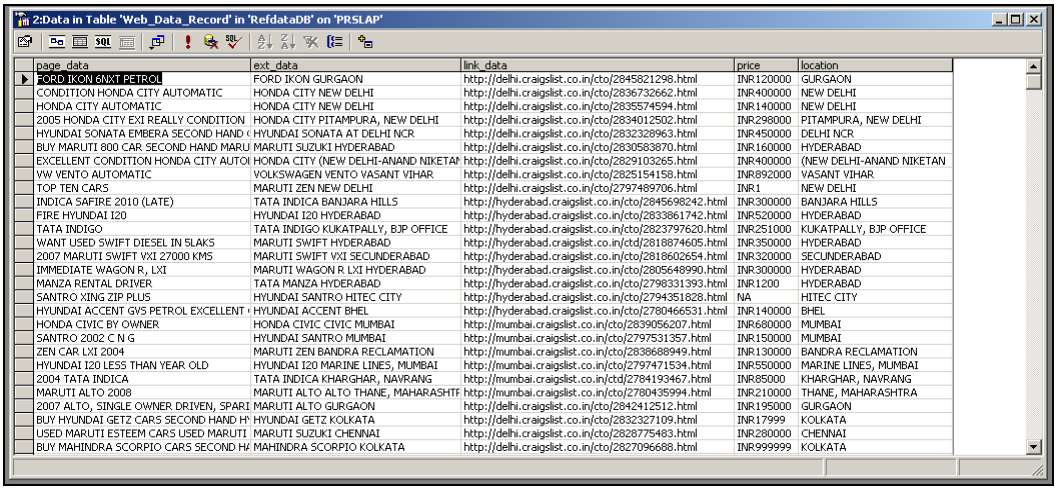


Fig – 6.3 – Data Model Reference and Reference Sets

6.2 Extracting Web Data Records

Fig 6.4 – Extracting Web Data Records



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

page_data	ext_data	link_data	price	location
FORD IKON 6NXT PETROL	FORD IKON GURGAON	http://delhi.craigslist.co.in/cto/2845821298.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/2835574594.html	INR140000	NEW DELHI
2005 HONDA CITY EXI REALLY CONDITION	HONDA CITY PITAMPURA, NEW DELHI	http://delhi.craigslist.co.in/cto/2834012502.html	INR298000	PITAMPURA, NEW DELHI
HYUNDAI SONATA EMBERA 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/2830583870.html	INR160000	HYDERABAD
EXCELLENT CONDITION HONDA CITY AUTOI	HONDA CITY (NEW DELHI-ANAND NIKETAI	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/2825154158.html	INR892000	VASANT VIHAR
TOP TEN CARS	MARUTI ZEN NEW DELHI	http://delhi.craigslist.co.in/cto/2797489706.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 5LAKS	MARUTI SWIFT HYDERABAD	http://hyderabad.craigslist.co.in/ctd/2818874605.html	INR350000	HYDERABAD
2007 MARUTI SWIFT VXI 27000 KMS	MARUTI SWIFT VXI SECUNDERABAD	http://hyderabad.craigslist.co.in/cto/2818602654.html	INR320000	SECUNDERABAD
IMMEDIATE WAGON R, LXI	MARUTI WAGON R LXI HYDERABAD	http://hyderabad.craigslist.co.in/cto/2805648990.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/2794351828.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/2839056207.html	INR680000	MUMBAI
SANTRO 2002 C N G	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/2838688949.html	INR130000	BANDRA RECLAMATION
HYUNDAI I20 LESS THAN YEAR OLD	HYUNDAI I20 MARINE LINES, MUMBAI	http://mumbai.craigslist.co.in/cto/2797471534.html	INR550000	MARINE LINES, MUMBAI
2004 TATA INDICA	TATA INDICA KHARGHAR, NAVRANG	http://mumbai.craigslist.co.in/ctd/2784193467.html	INR85000	KHARGHAR, NAVRANG
MARUTI ALTO 2008	MARUTI ALTO ALTO THANE, MAHARASHTF	http://mumbai.craigslist.co.in/cto/2780435994.html	INR210000	THANE, MAHARASHTRA
2007 ALTO, SINGLE OWNER DRIVEN, SPARI	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 HA	MAHINDRA SCORPIO KOLKATA	http://delhi.craigslist.co.in/cto/2827096688.html	INR999999	KOLKATA

Fig 6.5 – Web Data Records data Structure

Fig 6.4 and 6.5 shows the web data records extractions from web pages gathered from various web post sites. It extracts the link post data, link data and related attributes. The process for prepare the correct form of post data by referring the reference set data and stored in the data structure shown in Fig 6.5.

6.3 User Search Query

Fig – 6.6 User Search Query Interface

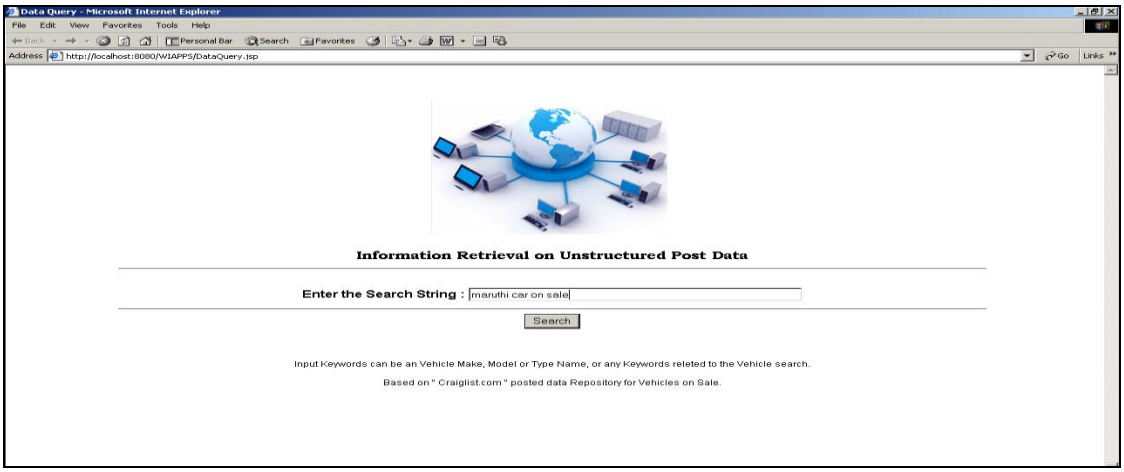


Fig – 6.7 User Search Query Interface

User needs to pose query using Search Query Interface as shown in Fig – 6.6 and Fig 6.7. The search result will be displayed in search result interface as shown below in Fig 6.8.

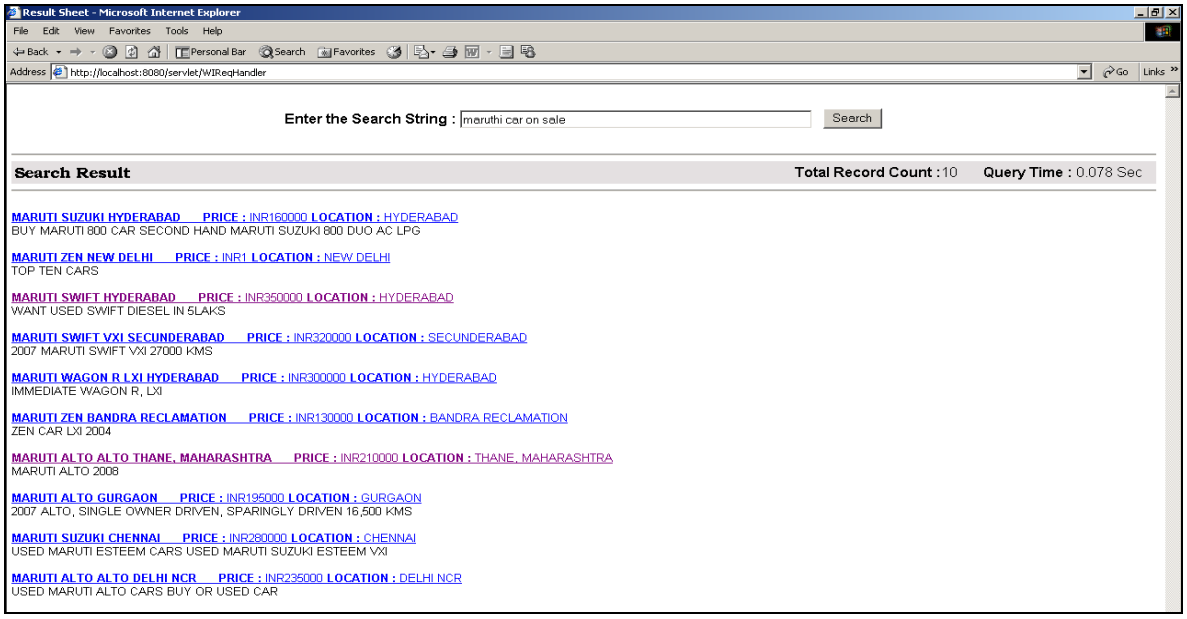


Fig – 6.8 User Search Result Interface

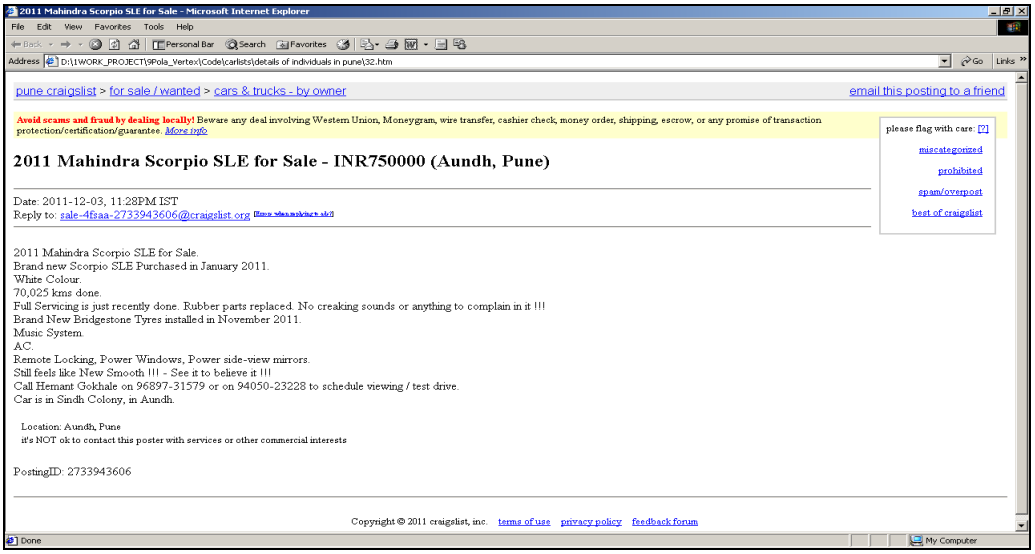


Fig – 6.9 User Search Result Interface

The search result can be viewed by clicking the link. It will be displayed the post data from online as shown in Fig 6.9.

VII. CONCLUSION

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] Hsu, C.-N. and Dung, M., Generating finite-state transducers for semi-structured data extraction from the web.

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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 Artificial Intelligence Research 31 (2008)*, Page(s):543-590

[8] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W. Ma. Simultaneous record detection and attribute labeling in web data extraction. In *SIGKDD*, 2006.

[9] Y. Zhai and B. Liu. Web data extraction based on partial tree assignment. In *WWW*, 2005.

[10] Riloff, E., Automatically constructing a dictionary for information extraction tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93), pp. 811-816, AAAI Press/The MIT Press, 1993.

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

[12] 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.

[13] 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.

[14] Arocena, G. O. and Mendelzon, A. O., WebOQL: Restructuring documents, databases, and Webs. Proceedings of the 14th IEEE International Conference on Data Engineering (ICDE), Orlando, Florida, pp. 24-33, 1998.

[15] Hammer, J., McHugh, J. and Garcia-Molina, Semistructured data: the TSIMMIS experience. In Proceedings of the 1st East-European Symposium on Advances in Databases and Information Systems (ADBIS), St. Petersburg, Rusia, pp. 1-8, 1997.

[16] Saiiuguet, A. and Azavant, F., Building intelligent Web applications using lightweight wrappers. Data and Knowledge Engineering 36(3): 283-316, 2001.