Structure based Data Extraction from Hidden Web Sources: A Review

Anuradha
Department of Computer Engineering
YMCA University of Sc. & Technology
Faridabad, India

A.K.Sharma
Department of Computer Engineering
YMCA University of Sc. & Technology
Faridabad, India

ABSTRACT

In order to extract data from the web pages of Hidden web sources, many semi-automatic and automatic techniques are proposed based on structure and tags of HTML documents. These techniques include machine learning and schema- matching approaches to solve the problem of data extraction. This paper discusses the research that has been done in the area of data extraction from Hidden Web sources. The goal of this paper is to discuss the advantages and disadvantages of currently existing techniques.

Keywords

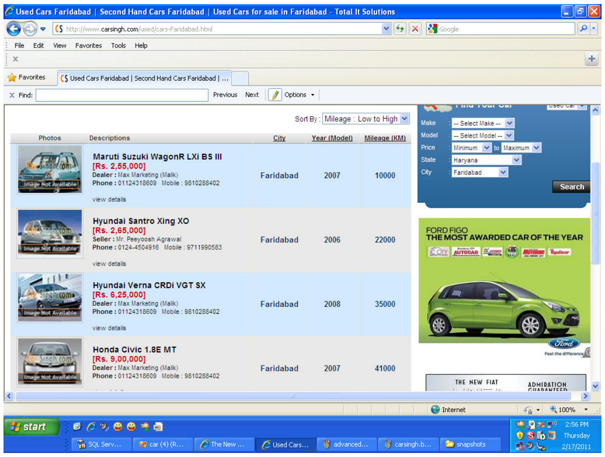
Surface Web, Hidden Web, Information Extraction.

1. INTRODUCTION

World Wide Web (WWW) is broadly divided into two categories: The Surface web that contains 1% of information content of the web. Search engine crawl along the web to extract and index text from HTML documents on the websites, then make this information searchable through keywords. Second is the Hidden web that contains 99% of information content of the web. Most of this information is contained in the databases and is not indexed by search engines. This means if we are searching for information from surface web only, we are searching through only 1% of WWW and missing 99% of it. Moreover, 95% of hidden web is free publicly accessible information.

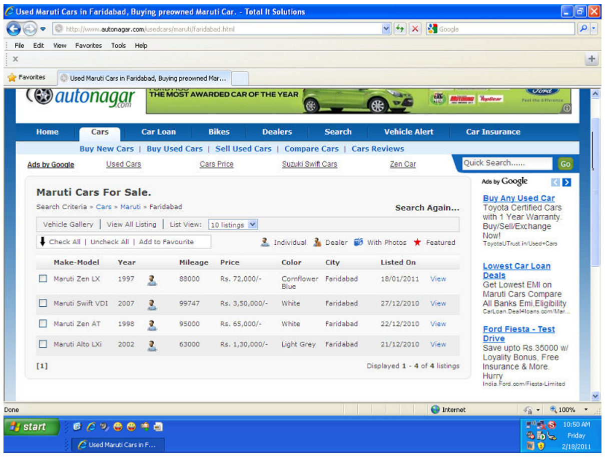
Surface Web refers to the part of WWW that contains static web pages and these pages are linked to many other pages. Traditional search engines create their indices by crawling these web pages. Hidden Web consists of web pages that are created dynamically by filling the search query forms. Traditional crawlers do not attempt to find dynamic pages that are the result of database queries due to the millions of queries that are possible. This means there is an urgent need to design or develop the system that will extract the information from hidden web sources and present them to user in integrated form. Such attempt is known as information extraction (IE) by various researchers.

Mostly the Websites contains their data in one of the three forms: unstructured, semi-structured and structured. Unstructured information does not contain any structure or they are in text form. Structured information contains defined structure and contains high quality data. They are in table form as shown in figure2. Semi-structured information lies between above two. They lack in web defined structure but in this type, similar objects grouped together in some manner as shown in figure1.


Used Cars Faridabad | Second Hand Cars Faridabad | Used Cars for sale in Faridabad - Total It Solutions
http://www.carsingh.com/used/cars-Faridabad.html
File Edit View Favorites Tools Help
Favorites
Find: Previous Next Options
Sort By : Mileage : Low to High
Make
Model
Price
State
City
Search
FORD FIGO THE MOST AWARDED CAR OF THE YEAR
start

Photos	Descriptions	City	Year (Model)	Mileage (KM)
Image Not Available	Maruti Suzuki WagonR LXi BS III [Rs. 2,55,000] Dealer : Max Marketing (Malik) Phone : 01124318609 Mobile : 9810288402 view details	Faridabad	2007	10000
Image Not Available	Hyundai Santro Xing XO [Rs. 2,65,000] Seller : Mr. Peeyoosh Agrawal Phone : 0124-4504916 Mobile : 9711990583 view details	Faridabad	2006	22000
Image Not Available	Hyundai Verna CRDi VGT SX [Rs. 6,25,000] Dealer : Max Marketing (Malik) Phone : 01124318609 Mobile : 9810288402 view details	Faridabad	2008	35000
Image Not Available	Honda Civic 1.8E MT [Rs. 9,00,000] Dealer : Max Marketing (Malik) Phone : 01124318609 Mobile : 9810288402	Faridabad	2007	41000

Fig1: Example of Semi-structured data


Used Maruti Cars in Faridabad, Buying preowned Maruti Car. - Total It Solutions
http://www.autonagar.com/usedcars/maruti/faridabad.html
File Edit View Favorites Tools Help
autonagar.com
THE MOST AWARDED CAR OF THE YEAR
Home Cars Car Loan Bikes Dealers Search Vehicle Alert Car Insurance
Buy New Cars | Buy Used Cars | Sell Used Cars | Compare Cars | Cars Reviews
Ads by Google Used Cars Cars Price Suzuki Swift Cars Zen Car
Quick Search...... Go
Buy Any Used Car
Lowest Car Loan Deals
Ford Fiesta - Test Drive
Done Internet
start

Maruti Cars For Sale.

Search Criteria » Cars » Maruti » Faridabad

Search Again...

Vehicle Gallery | View All Listing | List View: 10 listings

Check All | Uncheck All | Add to Favourite — Individual Dealer With Photos ★ Featured

	Make-Model	Year		Mileage	Price	Color	City	Listed On	
□	Maruti Zen LX	1997		88000	Rs. 72,000/-	Cornflower Blue	Faridabad	18/01/2011	View
□	Maruti Swift VDI	2007		99747	Rs. 3,50,000/-	White	Faridabad	27/12/2010	View
□	Maruti Zen AT	1998		95000	Rs. 65,000/-	White	Faridabad	22/12/2010	View
□	Maruti Alto LXi	2002		63000	Rs. 1,30,000/-	Light Grey	Faridabad	21/12/2010	View

[1] Displayed 1 - 4 of 4 listings

Fig2: Example of Structured data

Semi-structured and structured pages also contain some unwanted information like advertisement, copyright statements, navigation patterns etc. The real content lies within the data region in which user is interested. Many IE systems are developed to extract only the useful content from the page. Most of these systems extract the data based on HTML structure and HTML tags of the webpage. In the next section, we will discuss some of these systems.

2. LBDRF (LAYOUT BASED DATA REGION FINDING)

This system[1] has three components. They are

1. *HTML Tree Constructor*- It is designed to translate the HTML file to a Tree, which is the input of Data Region Finder.

2. *Data Region Finder*- It takes the Tree as an input, and adopts the LBDRF algorithm to find data region in the list page.

3. *Wrapper Induction*- It produces the wrapper rule for data record extraction according to tag path schema.

Once the rules are constructed, data is extracted using these rules. Architecture for LBRF algorithm is shown below in fig.3.

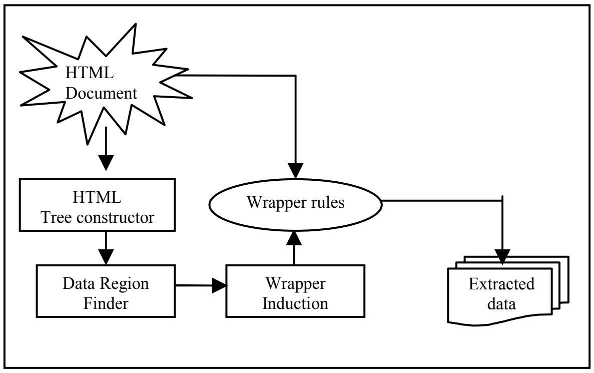
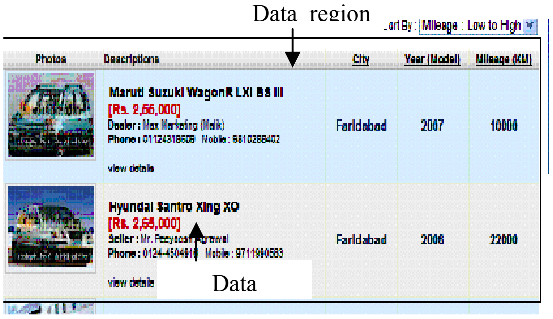


Fig 3: LBRF System Architecture

2.1 Tree Constructor:

In the first step, an HTML page is converted into a Tree where each node represents an HTML tag pair, e.g. the body object represents the body tag of the HTML page (<body>and</body>). The nested structure of HTML tags corresponds to the parent-child relationship among Tree nodes. fig 5 shows the Tree segment corresponding to the segment of the web page in fig.4.



_ort By : Mileage : Low to High

Photos	Descriptions	City	Year (Model)	Mileage (KM)
	Maruti Suzuki WagonR LXI BS III [Rs. 2,56,000] Dealer : Max Marketing (Malik) Phone : 01124318508 Mobile : 9810288402 view details	Faridabad	2007	10000
	Hyundai Santro Xing XO [Rs. 2,65,000] Seller : Mr. Peeyoosh Agrawal Phone : 0124-4504918 Mobile : 9711990583 view details	Faridabad	2006	22000

Fig 4: A segment of representative list page

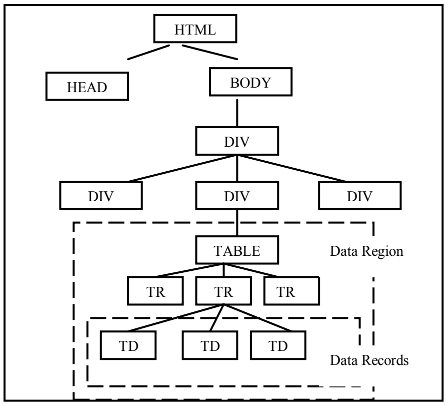


Fig 5 : Tree segment

2.2 Finding the Data Region Tree:

LBDRF algorithm given below in figure 6 selects the data region which is a sub tree with the tag *table* as its root. Each data record is displayed by a TR node, which can be seen from inner square frame. There are three functions used by this algorithm. These are as follows:

1. *GetCandidate* (Node root) is used to find the data region candidates.

2. *GetRegionNode* (ArrayList candidate) is used to find the root node of data region by comparing the length between the candidate node and the statistic and paging node.

3. *GetStatisticandPagingNode* (Body) is used to find the node which displays the statistic information and paging information.

```
Algorithm Node LBDRF (Node rootnode)
1: begin
2: ArrayList candidate=new ArrayList ();
3: GetCandidate (rootnode, candidate);
4: Node regionrootnode=GetRegionNode (candidate);
5: return regionrootnode;
6: end
```

Fig 6 : LBDRF algorithm

2.3 Extracting the Data Records in the Data Region:

After finding the data region node from the page, system extracts the data records from this data region by finding the longest common tag path for every group.

3. MDR(MINING DATA RECORDS IN WEB PAGES)

Liu and Grossman[2] proposed a novel method to mine data records in a Web page automatically which is called as MDR. It currently finds all data records formed by table and form related tags, i.e., table, form, tr, td, etc. It assumes that majority of web pages are constructed by these tags. The algorithm is based on two observations:

1. A group of data records are always presented in a contiguous region of the web page and are formatted using similar HTML tags. Such region is called a Data Region.

2. The nested structure of the HTML tags in a web page usually forms a tag tree and a set of similar data records are formed by some child sub-trees of the same parent node.

This work is presented in three main steps:

1. Building an HTML Tag tree
2. Mining Data Regions
3. Identify Data Records

3.1 Building the HTML Tag Tree :

Most HTML tags work in pairs. Each pair consists of an *opening* tag and a *closing* tag. Within each corresponding tag pair, there can be other pairs of tags, resulting in nested blocks of HTML codes. Building a *tag tree* from a Web page using its HTML code is thus natural. In this algorithm, each pair of tags is considered as one *node*. An example tag tree is shown in figure 7.

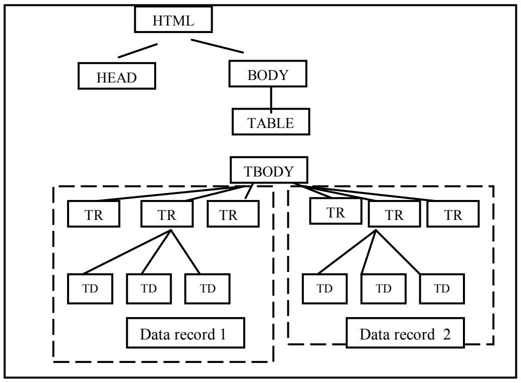


Fig 7 : Tag Tree

3.2 Mining Data Regions:

This step identifies every data region in a Web page that contains similar data records. It first mines *generalized nodes* in a page and then finds out data records. A sequence of adjacent generalized nodes forms a data region. From each data region, it identifies the data records.

A *generalized node* (or a *node combination*) of length *r* consists of *r* (*r* ≥ 1) nodes in the HTML tag tree with the following two properties:

1) the nodes all have the same parent.
2) the nodes are adjacent.

A *data region* is a collection of two or more generalized nodes with the following properties:

1) the generalized nodes all have the same parent.
2) the generalized nodes all have the same length.
3) the generalized nodes are all adjacent.
4) the normalized edit distance (string comparison) between adjacent generalized nodes is less than a fixed threshold.

For example, in figure 7, it has two generalized nodes, the first one consists of the first 5 children TR nodes of TBODY, and the second one consists of the next 5 children TR nodes of TBODY. It is important to notice that although the generalized nodes in a data region have the same length (the same number of children nodes of a parent node in the tag tree), their lower level nodes in their sub-trees can be quite different. Thus, they can capture a wide variety of regularly structured objects.

3.3 String Comparison Using Edit Distance:

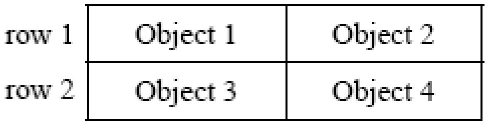
The string comparison method that we use is based on the normalized *edit distance* [1][7]. Let the two strings be *s*1 and *s*2. The time-complexity of the algorithm is $O(|s1||s2|)$ [1]. The computation can be substantially reduced as this method is only interested in very similar strings. The computation is only large when the strings are long. If $|s1| > 2|s2|$ or $|s2| > 2|s1|$, no comparison is needed because they are obviously too dissimilar.

3.4 Determining Data Regions:

This method identifies each data region by finding its generalized nodes. The algorithm basically uses the string comparison results at each parent node to find similar children node combinations to obtain candidate generalized nodes and data regions of the parent node.

3.5 Identify Data Records:

After all data regions and their generalized nodes are found from a page, we can identify data records in each region. A generalized node may not be a data record containing a single object. The actual records may be at a lower level, i.e., a generalized node may contain one or more data records. Table1 shows a data region that contains two table rows (1 and 2). Row 1 and row 2 have been identified as generalized nodes.



row 1	Object 1	Object 2
row 2	Object 3	Object 4

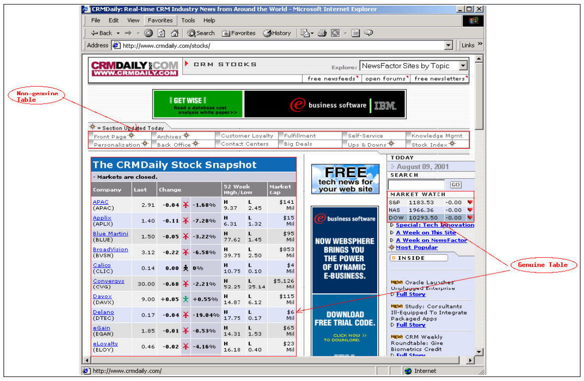
Table1: Data region containing rows

However, they are not individual data records. Each row actually contains two data (objects) records in the two cells.

4. A MACHINE LEARNING BASED APPROACH FOR TABLE DETECTION ON THE WEB BY (YALIN WANG AND JIANYING HU)

Yalin Wang and Jianying Hu [3] proposed a machine learning approach to detect data rich tables on a web page. Tables are used to represent relational information in web documents. Web designer choose <TABLE> tag not only for relational information display but also to create any type of multiple-column layout for easy viewing, thus the presence of the <TABLE> tag does not necessarily indicate the presence of a relational table. In this work, they defined *genuine* tables to be documents where a two dimensional grid is used for the logical relations among the cells. In contrast, Non-genuine tables are documents where <TABLE>

tags are used as a mechanism for grouping contents into clusters for easy viewing only. This work refers Table detection as the technique which classifies a document entity enclosed by the <TABLE></TABLE> tags as genuine or non-genuine tables. Figure 8shows examples of genuine and non-genuine tables.


Non-genuine Table
Genuine Table
File Edit View Favorites Tools Help
Back Search Favorites History
Address
Links
CRMDAILY.COM WWW.CRMDAILY.COM
GET WISE Read a database cost analysis white paper>>
business software IBM
FREE tech news for your web site
business software NOW WEBSPHERE BRINGS YOU THE POWER OF DYNAMIC E-BUSINESS.
DOWNLOAD FREE TRIAL CODE. CLICK NOW >> TO DOWNLOAD.
GO
Internet

CRMDaily: Real-time CRM Industry News from Around the World - Microsoft Internet Explorer

http://www.crmdaily.com/stocks/

CRM STOCKS

Explore: NewsFactor Sites by Topic

free newsfeeds | open forums | free newsletters

= Section Updated Today

Front Page	Archives	Customer Loyalty	Fulfillment	Self-Service	Knowledge Mgmt
Personalization	Back Office	Contact Centers	Big Deals	Ups & Downs	Stock Index

The CRMDaily Stock Snapshot

- Markets are closed.

Company	Last	Change		52 Week High/Low		Market Cap
APAC (APAC)	2.91	-0.04	-1.68%	H 9.37	L 2.45	$141 Mil
Applix (APLX)	1.40	-0.11	-7.28%	H 6.31	L 1.32	$15 Mil
Blue Martini (BLUE)	1.50	-0.05	-3.22%	H 77.62	L 1.45	$95 Mil
BroadVision (BVSN)	3.12	-0.22	-6.58%	H 39.75	L 2.50	$853 Mil
Calico (CLIC)	0.14	0.00	0%	H 10.75	L 0.10	$4 Mil
Convergys (CVG)	30.00	-0.68	-2.21%	H 52.25	L 25.14	$5,126 Mil
Davox (DAVX)	9.00	+0.05	+0.55%	H 14.87	L 6.12	$115 Mil
Delano (DTEC)	0.17	-0.04	-19.04%	H 17.75	L 0.17	$6 Mil
eGain (EGAN)	1.85	-0.01	-0.53%	H 14.31	L 1.53	$65 Mil
eLoyalty (ELOY)	0.46	-0.02	-4.16%	H 16.18	L 0.40	$23 Mil

TODAY
August 09, 2001
SEARCH

MARKET WATCH

S&P	1183.53	-0.00
NAS	1966.36	-0.00
DOW	10293.50	-0.00

Special: Tech Innovation
A Week on This Site
A Week on NewsFactor
Most Popular

INSIDE

NEW! Oracle Launches Unplugged Enterprise
Full Story

NEW! Study: Consultants Ill-Equipped To Integrate Packaged Apps
Full Story

NEW! CRM Weekly Roundtable: Give Biometrics Credit
Full Story

http://www.crmdaily.com/

Fig 8 : Example of genuine and non-genuine tables

This method is based on feature selection. Features were designed to capture both of these aspects. Here, 16 features were developed which can be categorized into three groups: seven layout features, eight content type features and one word group feature. With statistical and heuristic functions they compare among other things the table layout, relation between rows and columns, the length of data in each cell and contents (images, links, texts) within each cell in the target page with a ground truth data collection. The problem with this approach is that an entire data collection of ground truth data is needed in order for it to work. Even though the system seems to find data-rich tables, there is no guarantee that they find the correct data-rich tables. For example, if we are looking for product information we just as might find company information, addresses or etc., and most product pages (like those containing books, computers, hardware) contain lots of images and links that would aggravate their heuristic functions used for classification. Product pages will have fundamental differences in layout, and even if you use very varying input the result is totally dependent on the ground truth data.

5. A VISION BASED APPROACH FOR DATA EXTRACTION BY P. S. HIREMATH, SIDDU P. ALGUR

This research[4] proposed a novel method to extract data items from the web pages. It comprises of two steps:

(1) Identification and Extraction of the data regions based on visual clues information.
(2) Identification of data records and extraction of data items from a data region.

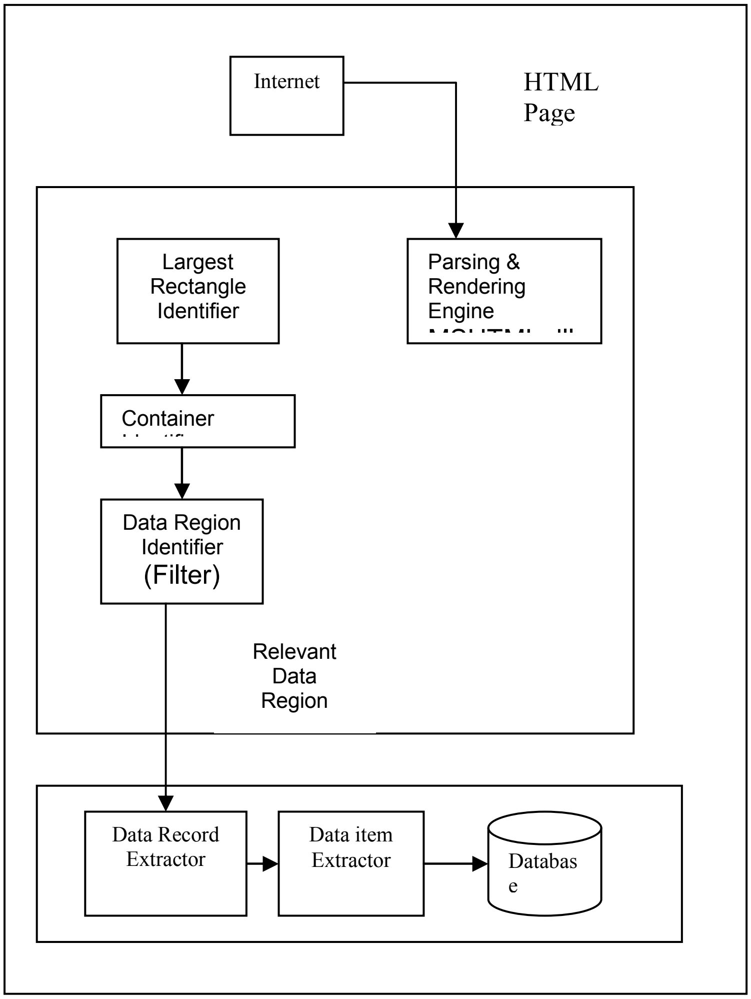


Fig 9: Framework for vision Based pproach

The framework for visual based data extraction is given below in figure 9. The HTML page is input to the first phase which extracts the relevant data regions and these data regions are input to the second phase that extracts the data item from them. The output of each component is the input for the next component.

5.1 Identification and Extraction of the data regions:

A data region is defined as the most relevant portion of a web page. Figure 10 shows an example, which is a web page that consists of a data region containing list of four books. Data Record is the full description of each book.

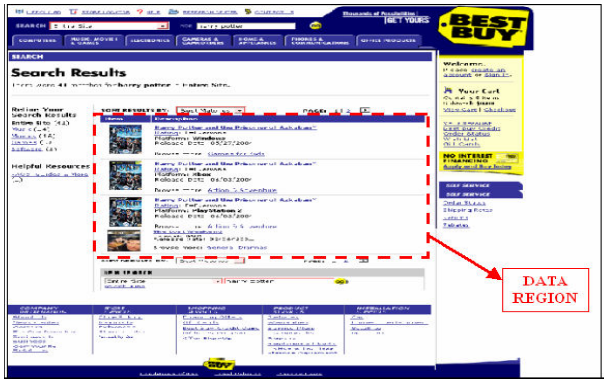


Fig 10: An example of a Data Region containing 4 data records

An algorithm for the proposed technique is given in figure 11. This algorithm works by taking the HTML document as an input. The MSHTML parsing and rendering engine is the main HTML component of the Microsoft Internet Explorer web browser. The rendering Engine of the browser produces the boundary coordinates. It is assumed that the most relevant data in that web page lies inside the largest rectangle. Based upon the coordinates returned by rendering engine, the largest rectangle amongst these bounding rectangles is determined. Line 4 in above algorithm identifies the data region which consists of the relevant data region and some irrelevant regions also. Line 5 identifies the actual relevant data region by filtering the bounding irrelevant regions.

```
Algorithm
Input: HTML Document
1. Set maxRect=NULL
2. Set dataRegion=NULL
3. FindMaxRect(BODY);
4. FindDataRegion(maxRect);
5. FilterDataRegion(dataRegion);
end
```

The Data region returned from the first algorithm shown in figure 11 is scanned for the tags. If the tag is a <table> tag then this means this is the beginning of a record and the procedure for extracting the data items from a data record is called.

5.2 Identification of Data Record:

Data records are identified by applying the algorithm to data region as shown below in figure 12. It identifies the tags in the HTML code of the data region which contains the relevant records and is given below:

```
Algorithm
Input: DataRegion
for each tag in DataRegion till the EOF
If Tag = <Table> tag
Extract Contents till </Table> tag
Store it in a Record
ExtractDataItems(Record)
endif
end
```

Fig 12 : Algorithm for Finding Data records

5.3 Identification of Data items:

The next step is the item extraction from each data record. The data items of each record are extracted by taking one data record with data fields at one time. The items are looked for <TD> tag and sent to the database.

6. COMPARISON OF DIFFERENT DATA EXTRACTION METHODS

Sr. No.	Method	Advantages	Disadvantages
1	**LBDRF**	Extracts the data from hidden web sources by constructing tree.	This method assumes that large majority of web data records are formed by <table>, <TR> and <TD> tags. Hence, it mines the data records by looking only at these tags. Other tags like <div>, <span> are not considered.
2.	**MDR**	If the web page contains table tags then it can mine the data records automatically.	MDR, not only identifies the relevant data region containing the search result records but also extracts records from all the other sections of the page, e.g, some advertisement records also, which are irrelevant.

Sr. No.	Method	Advantages	Disadvantages
3.	A Machine Learning Based Approach for Table Detection on the Web by (Yalin Wang and Jianying Hu)	Genuine tables are detected easily and can be used for data extraction.	The entire data collection of ground truth data is needed in order for it to work. For example, if we are looking for product information we just as might find company information, addresses or etc
4.	A Vision Based Approach for Data extraction by P. S. Hiremath, Siddu P. Algur	Identification and Extraction of the data regions are based on visual clues information. Data records can be Identified from data items of a data region. So, This method is independent of extraction rules up to some extent.	Data region is found by identifying the largest container. But there can be the cases when the result page contains one or two result then, the container will be very small. So, this method is inefficient. Secondly, again this method assumes that large majority of web data records are formed by <table>, <TR> and <TD> tags. Hence, it mines the data records by looking only at these tags. Other tags like <div>, <span> are not considered.

Table2: Comparison between existing Techniques

7.CONCLUSION AND FUTURE SCOPE :

This paper discusses the existing techniques to extract data from structured and semi-structured web pages. All these techniques use the HTML structure of the web page to find the location of desired information. After reaching at this location, they again find data region in which the real content lies. Advantages and disadvantages are discussed in this paper. By studying all these techniques, we can develop the new data extraction system which should be independent of extraction rules and also the HTML tags.

8. REFERENCES:

[1] Chen Hong-ping; Fang Wei; Yang Zhou; Zhuo Lin; Cui Zhi-Ming; Automatic Data Records Extraction from List Page in Deep Web Sources; 978-0-7695-3699- 6/09 © 2009 IEEE pages 370-373.

[2] Bing Liu, Robert Grossman, and Yanhong Zhai. Mining data records in web pages. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 601–606, New York, NY, USA, 2003.ACM Press.

[3] YalinWang and Jianying Hu. A machine learning based approach for table detection on the web. In WWW '02: Proceedings of the 11th international conference on World Wide Web, pages 242–250, New York, NY, USA, 2002. ACM Press.

[4] Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. 2003. VIPS: a Vision-based Page Segmentation Algorithm. Tech. Rep. MSR-TR-2003-79, Microsoft Technical Report.

[5] Simon, K., Lausen, G., and Boley, H. 2006. From HTML documents to web tables and rules. In ICEC, M. S. Fox and B. Spencer, Eds. ACM International Conference Proceeding Series, vol. 156. ACM, 125–131.

[6] Chang, K. C.-C., He, B., Li, C., Patel, M., and Zhang, Z. 2004. Structured databases on the web: observations and implications. SIGMOD Rec. 33, 3, 61–70.

[7] Freitag, D. 1998. Information Extraction from HTML: Application of a General MachineLearning Approach. In AAAI/IAAI. 517–523.

[8] B. Liu and Y. Zhai. NET: System for extracting Web data from °at and nested data records. In *Proceedings of the Conference on Web Information Systems Engineering*, pages 487-495, 2005.

[9] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. In Proceedings of VLDB, pages 129–138, 2001.

[10] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280(5360):98–100, 1998.