IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 5, Ver. V (Sep. – Oct. 2015), PP 41-48
www.iosrjournals.org

Main Content Mining from Multi Data Region Deep Web Pages

Shaikh Phiroj Chhaware[1], Dr.Mohammad Atique[2], Dr. L. G. Malik[3]

[1]*(Research Scholar, G.H. Raisoni College of Engineering, Nagpur, India)*
[2]*(Associate Professor, Dept. of Computer Science & Engg., S.G.B Amravati University, India)*
[3]*(Professor & HOD, Dept. of Computer Science & Engg., G.H. Raisoni College of Engg., Nagpur, India)*

Abstract: *An Automatic Data extraction from the deep web pages is an endeavoring errand. Colossal web substance are gotten to as per well-known interest submitted to Web databases and the returned data records are enwrapped in reasonably made site pages. The structures of such site pages are clearer to originator of the areas. Diverse methods of insight have been tended to before for managing this issue however every one of them have necessities in light of the way that they are changing vernacular subordinate. In like route there is principal work is open in revamp data extraction from the vital page pages having single data territories considering solidifying visual fragments close to report thing model tree of colossal site page. This paper is goes for showing the framework for modified web data extraction from basic pages which has multi data area in setting of the cross breed approach. Two synchronous approaches are: First applying remembering the deciding objective to see the distinctive data locales the considered managing engraving tree and second, mining the positive data records and data things from each of the data develops self-ruling based vision based page division estimation.*

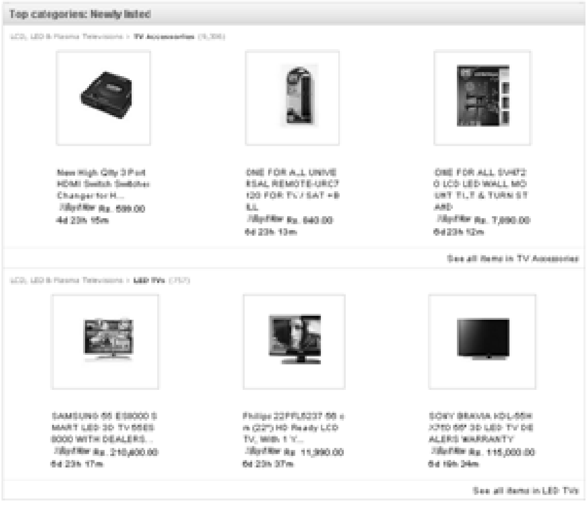
Keywords: *Web Data Extraction, Deep Web Pages, Multi Data Region Deep Web Pages, Data Record, Data Item, Wrapper Generation, Tag Tree*

I. Introduction

The World Wide Web has a large number of searchable material sources. These searchable material sources incorporate both web indexes and net databases. The accessible net databases can be sought through their web enquiry lines. The quantity of net lists has stretched out to more than a quarter century as indicated by late study [1]. All the net indexes make up the incredible net (inconspicuous net or imperceptible net). Frequently the recouped material (enquiry results) is enwrapped in net folios as information records. These unique website pages called as profound site pages are produced powerfully and are difficult to record by the conventional crawler-based web search tools, for example, Google and Yahoo.

A web structure and format fluctuates rely on upon diverse substance sort it will speak to or the essence of the fashioner styling its substance. Subsequently principle substance position or the fundamental tag containing primary substance varies in assortment of sites. Indeed, even there may be some substance in site hit that are other than one another however really in record article model (DOM) tree they are not at the same level and same folks, so discovering the primary substance here and situating components needs muddled and immoderate calculations.

The Fig. 1 shows a deep web page having multi-data region and Fig. 2 shows a deep web page having single data region.



Top categories: Newly listed

LCD, LED & Plasma Televisions > TV Accessories

New High Qlty 3 Port HDMI Switch Switcher Changer for H... Rs. 599.00 4d 23h 15m

ONE FOR ALL UNIVERSAL REMOTE-URC7120 FOR TV / SAT +BILL Rs. 640.00 6d 23h 13m

ONE FOR ALL SV4720 LCD LED WALL MOUNT TILT & TURN STAND Rs. 7,890.00 6d 23h 12m

See all items in TV Accessories

LCD, LED & Plasma Televisions > LED TVs

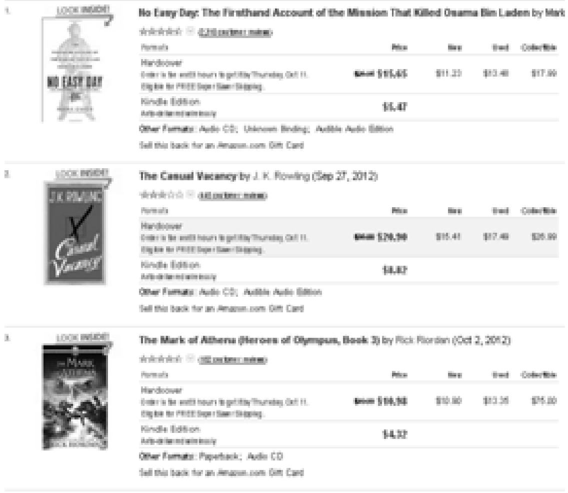
SAMSUNG 55 ES8000 SMART LED 3D TV 55ES8000 WITH DEALERS... Rs. 210,800.00 6d 23h 17m

Philips 22PFL5237 56 cm (22") HD Ready LCD TV, With 1 Y... Rs. 11,990.00 6d 23h 37m

SONY BRAVIA KDL-55HX750 55" 3D LED TV DEALERS WARRANTY Rs. 115,000.00 6d 19h 24m

See all items in LED TVs

Fig. 1: A Deep Web Page (Multi-Data Region)



1. No Easy Day: The Firsthand Account of the Mission That Killed Osama Bin Laden by Mark

Hardcover $15.65 $11.23 $13.48 $17.99

Kindle Edition $5.47

Other Formats: Audio CD; Unknown Binding; Audible Audio Edition

Sell this back for an Amazon.com Gift Card

2. The Casual Vacancy by J. K. Rowling (Sep 27, 2012)

Hardcover $15.41 $17.49 $26.99

Kindle Edition

Other Formats: Audio CD; Audible Audio Edition

Sell this back for an Amazon.com Gift Card

3. The Mark of Athena (Heroes of Olympus, Book 3) by Rick Riordan (Oct 2, 2012)

Hardcover $10.90 $13.35 $75.00

Kindle Edition $4.32

Other Formats: Paperback; Audio CD

Sell this back for an Amazon.com Gift Card

Fig. 2: A Deep Web Page (Single Data Region)

The Web content information comprise of unstructured information, for example, free messages, semi-organized information, for example, HTML archives, and a more organized information, for example, information in the tables or database produced HTML pages [1].Web content extraction is worried with removing so as to separate the applicable content from Web pages irrelevant printed clamor like promotions, navigational components, contact and copyright notes. Web creeping includes seeking an expansive arrangement space which requires a great deal of time, hard circle space and part of utilization of assets.

The exploration done in Web substance mining from two distinct perspectives: IR and DB sees. IR perspective is predominantly to help or to enhance the data discovering and sifting the data to the clients normally in view of either gathered or requested client profiles. DB see chiefly tries to show the information on the Web and to incorporate them with the goal that more advanced inquiries other than the catchphrases based hunt could be performed. The three sorts of specialists are Intelligent hunt operators, Information sifting/Categorizing specialists, Personalized web operators. A shrewd Search specialist naturally scans for data as per a specific inquiry utilizing space attributes and client profiles. Data specialists utilized number of strategies to channel information as indicated by the predefine guidelines. Customized web specialists learn client inclinations and finds records identified with those client profiles. In Database approach it comprises of very much shaped database containing constructions and characteristics with characterized areas. The calculation proposed is called Dual Iterative Pattern Relation Extraction for discovering the important data utilized via internet searchers. The substance of website page incorporates no machine lucid semantic data. Web indexes, subject registries, canny specialists, bunch investigation and gateways are utilized to discover what a client must search for.

This paper has the accompanying commitment 1) investigation of method to recognize the numerous information locales of profound site page and 2) investigation of procedures to perform the information extraction from profound site pages utilizing essentially visual elements i.e. by applying the current vision based website page division calculation.

The rest of the paper is organized as follows: the related works are reviewed in Section 2. Technique for identification of multiple data regions of deep web page are introduced in Section 3. Data record extraction and data items extraction are described in Section 4. Conclusion and future works are presented in Section 5.

II. Related Work

A lot of work has been done here. Liu and Grossman [2] proposed a novel system to mine information records in a page consequently which is called as MDR. This technique bargains of two sections, first is about information records on net and second is, string likeness strategy. The MDR framework is competent to uncovering both abutting and non-connecting information records. However, their answer is dialect ward and thus does not have the essential issue of covering different other usage techniques for website pages separated from HTML. W. Liu and W. Meng [3] present the ViDE procedure, profound web information extraction in light of vision based page division method. The techniques proposed here is hearty and powerful if the profound page contains the single information areas. ViDE is not able to separate the information consequently from the multi-information locale profound website page. DESP [4] presents a programmed profound extractor on profound pages for book area which can extricate information things and mark qualities in the meantime. The instance of DESP is to concentrate book's data, for example, title, writer, cost, and distributer from result pages came back from book shop sites. [5] Proposed VIPS (vision based page division) framework to mine semantic game plan of a net page. Such semantic structure is a progressive structure in which every hub is will relates to a piece. Each knob will be distributed an expense (evaluation of levelheadedness) to indicate how conceivable of the substance in the lump based on the chromatic perception. A portion of the best known apparatuses that receive manual methodologies are Minerva [6], TSIMMIS [7], and Web-OQL [8]. Clearly, they have low proficiency and are not versatile. Self-loader methods can be grouped into succession based and tree-based. The previous, for example, WIEN [9], Soft-Mealy [10], and Stalker [11], speaks to records as arrangements of tokens or characters, and creates delimiter based extraction rules through an arrangement of preparing cases. The last, for example, W4F [12] and XWrap [13], parses the archive into a various leveled tree (DOM tree), in view of which they perform the extraction process. These methodologies require manual endeavors, for instance, marking some example pages, which is work concentrated and tedious. Keeping in mind the end goal to enhance the effectiveness and decrease manual endeavors, latest looks into spotlight on programmed approaches rather than manual or self-loader ones. Some illustrative programmed methodologies are Omini [14], RoadRunner [15], IEPAD [16], MDR [17], DEPTA [18], and the technique in [19]. Some of these methodologies perform just information record extraction however not information thing extraction, for example, Omini and the technique in [19]. RoadRunner, IEPAD, MDR, DEPTA, Omini, and the technique in [9] don't create wrappers, i.e., they recognize designs and perform extraction for every Web page straightforwardly without utilizing already inferred extraction rules. The methods of these works have been talked about and analyzed in [20].

III. Identification Of Data Regions

The given below system presented 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 LBRF (layout based data record finder) 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.

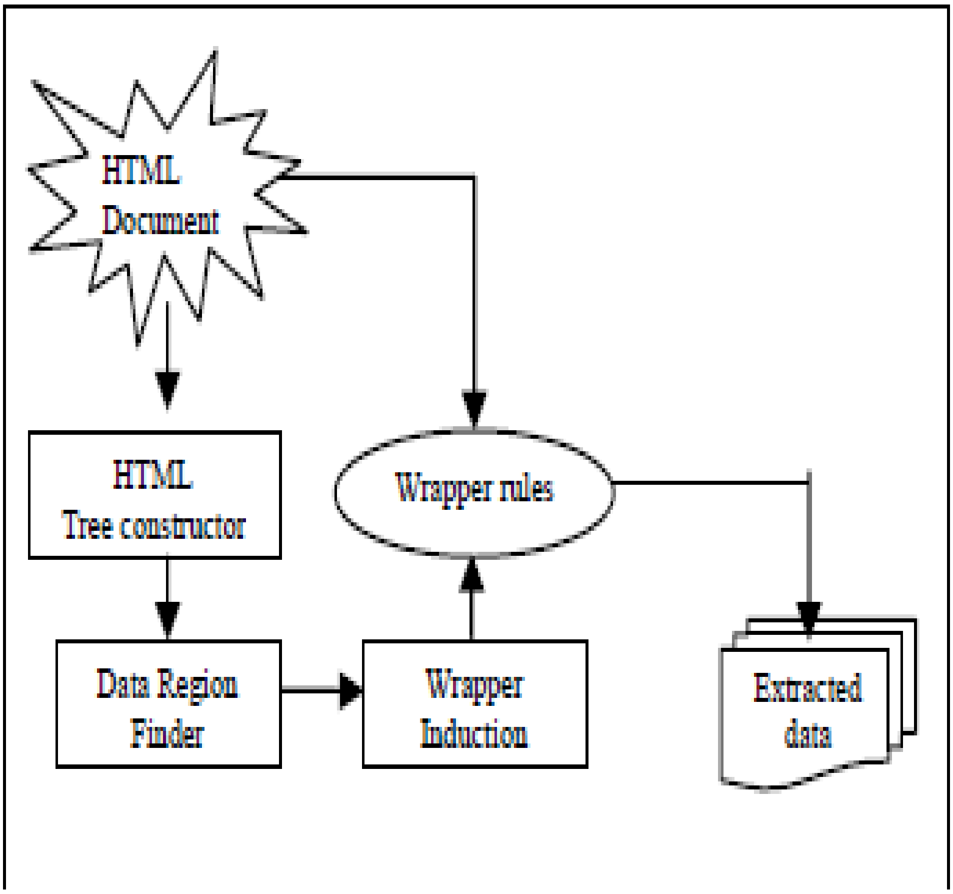


Fig. 3: Architecture of LBRF Algorithm

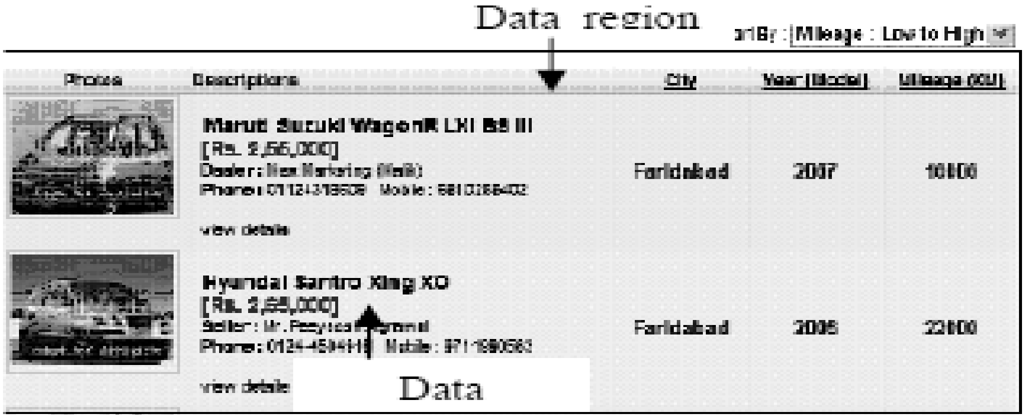
1. Tree Constructor

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.

2. Finding the Data Region Tree

LBDRF algorithm given below 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.



rtBy : Mileage : Low to High

Photos	Descriptions	City	Year (Model)	Mileage (KM)
	Maruti Suzuki WagonR LXI BS III [Rs. 2,56,000] Dealer: [illegible] Phone: [illegible] Mobile: [illegible] view details	Faridabad	2007	10000
	Hyundai Santro Xing XO [Rs. 2,58,000] Seller: Mr. [illegible] Phone: [illegible] Mobile: [illegible] view details	Faridabad	2006	22000

Fig. 4: A Subdivision of Representative List Page

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

The system presented in [4] can also be used and it is similar as that given in [21]. The work of [4] is described as below.

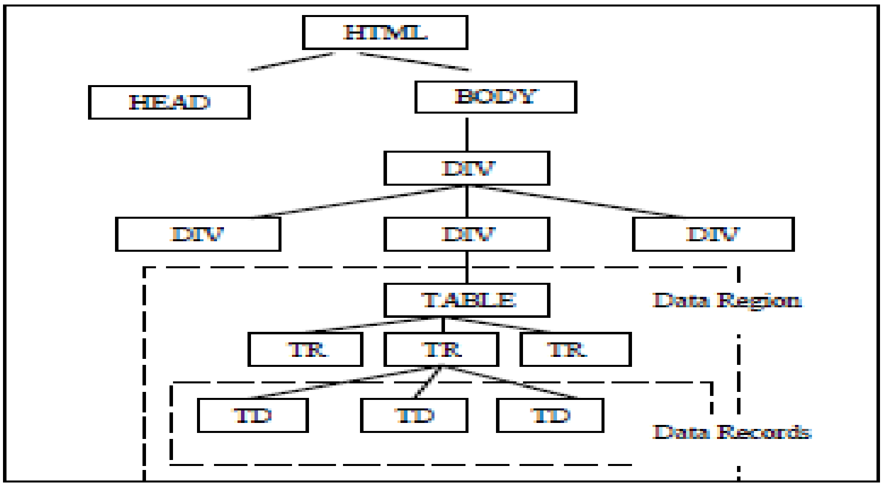


Fig. 5: Hierarchy Subdivision

3. Building the HTML Tag Tree

In this effort, we merely practice labels in string evaluation to catch data records. Utmost HTML labels work in duos. All duos comprises of an inaugural label and a concluding label. Inside every equivalent label duos, there can be additional duos of labels, causing in nested chunks of HTML programs. Constructing a label hierarchy from a net folio by its HTML program is thus normal. In our label hierarchy, all duos of labels is reflected as one nodule. An example label hierarchy is shown in Fig 6.

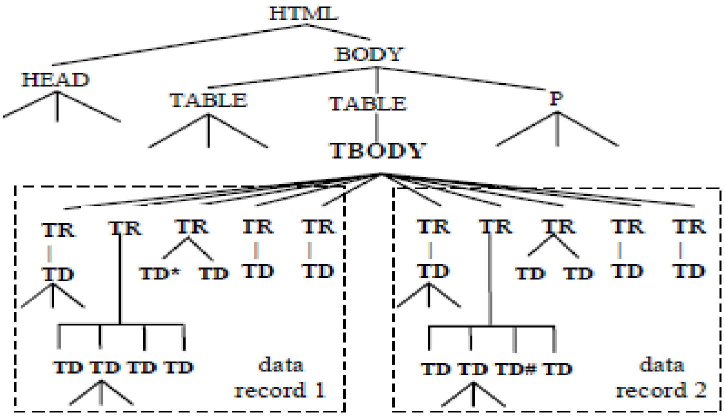


Fig. 6: Label Hierarchy of a Net Sheet

4. Mining Data Regions

This phase excavates each data area in a net sheet that comprises analogous data records. In its place of excavating data records straight, which is stubborn, we first excavate global nodules in a sheet. An arrangement of neighboring global nodules formulates a data area. After all data area, we will recognize the genuine data records.

Definition: A global nodule (or a nodule mixture) of distance r contains of r (r ≥ 1) nodules in the HTML label hierarchy with the subsequent two belongings:

1) All the nodules have the similar parent.
2) The nodules are neighboring.

The motivation that we present the global nodule is to capture the condition that an entity (or a data record) may be controlled in a limited sibling label nodules fairly than one. Note that we call all nodules in the HTML label hierarchy a label nodule to discriminate it from a global nodule.

Definition: An information area is a gathering of two or more global nodules with the subsequent belongings:

1) The global nodules all have the similar parent.
2) The global nodules all have the similar distance.
3) The global nodules are all neighboring.
4) The standardized edit distance (string assessment) among neighboring global nodules is smaller than a static threshold.

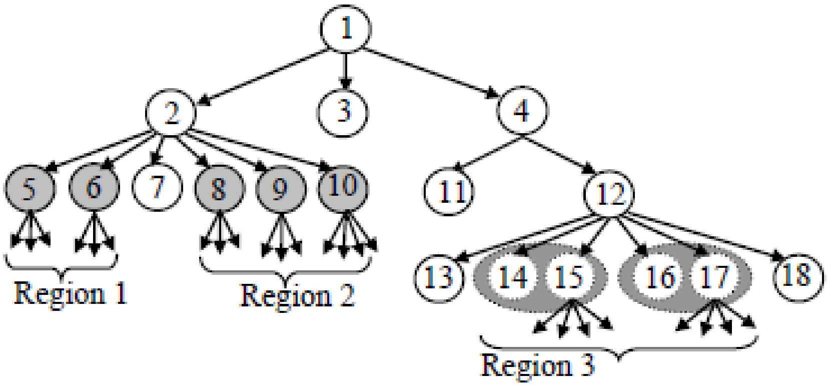


Fig. 7: An Illustration of Global Nodules And Data Areas

For instance, in Fig. 6, we can detail two worldwide hubs, the first contains initial 5 posterity TR knobs of TBODY, and the second one contains ensuing 5 posterity TR knobs of TBODY. It is noteworthy to inform that despite the fact that the worldwide knobs in an information zone have the indistinguishable separation (the comparable amount of posterity knobs of a guardian knob in the mark progressive system), their substandard level knobs in their sub-order can be genuinely different. In this way, they can capture a broad assorted qualities of as often as possible composed things.

To extra elucidate distinctive sorts of worldwide knobs and information regions; we make utilization of a false name pecking order in Fig. 7. For notational suitability, we don't utilize genuine HTML mark names however ID numbers to imply name knobs in a name order. The dappled zones are worldwide knobs. Knobs 5 and 6 are worldwide knobs of separation 1 and they made the information zone marked 1 if the alter separation condition 4) is satisfied. Knobs 8, 9 and 10 are additionally worldwide knobs of separation 1 and they formed the information territory marked 2 if the alter separation condition 4) is satisfied. The teams of knobs (14, 15) and (16, 17) are worldwide knobs of separation 2. They created the information zone marked 3 if the alter separation condition 4) is satisfied.

6. Determining Data Regions

We right now perceive each information region by finding its worldwide knobs. Here we utilization Figure 8 to embody the issues. In the blink of an eye there are 8 information records (1-8) in this sheet. Our strategy reports each column as a worldwide knob and the dash-lined holder as an information territory. The framework chiefly hones the string evaluation results at each guardian knob to find similar posterity knob blends to get wannabe worldwide knobs and information zones of the guardian knob. Three key issues are critical for making the last conclusions.

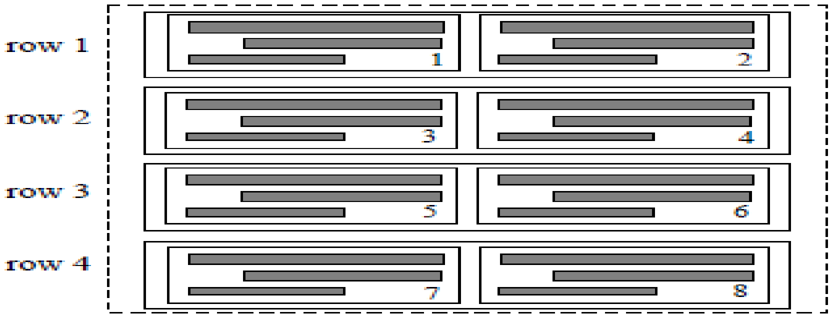


Fig. 8: A Possible Configuration of Data Records

1. On the off chance that a propelled level information territory shields a sub-par level information region, we report the propelled level information region and its worldwide knobs. Shield here implies that a mediocre level information range is inside a propelled level information region. Case in point, in Fig. 8, at a substandard level we find that cells 1 and 2 are competitor worldwide knobs and they made out of a wannabe information territory, line 1. On the other hand, they are encased by the information territory containing all the 4 lines at a propelled level. For this situation, we just report all column is a worldwide knob.

2. A stuff about practically equivalent to strings is that if a cluster of strings s1, s2, s3, .., sn, are undifferentiated from each other, then a blend of any amount of them is additionally comparable to an alternate blend of them of the indistinguishable amount. In this way, we just report worldwide knobs of the slightest separation that shield an information territory. In Fig. 8, we only report every column as a worldwide knob as opposed to a blend of two (lines 1-2, and lines 3-4).

3. An alter separation edge is required to pick whether two strings are comparable to. A gathering of preparing sheets is use to determine it.

```
Algorithm: FindDRs(Node, K, T)
    1.  If TreeDepth(Node) => 3 then
            a.  Node.DRs = IdenDRs(!,K,T)
            b.  tempDRs = ∞
    2.  For each child € Node.Childre do
            a.  FindDRs(Child, K, T)
            b.  tempDRs = tempDRs V UnCoveredDRs(Node, Child)
    3.  Node.DRs = Node.DRs V tempDRs
```

Fig. 9: Finding All Data Areas in the Label Hierarchy

IV. Data Record And Data Item Extraction

Now once the various data regions are identified, the VIPS tree or visual block tree will be constructed of the same web page to identify the data records within each data region and then getting the data items within each data records. This procedure is defined underneath in detail.

A. Data Record Extraction

Data record extraction aims to discover the boundary of data records and extract them from the deep Web pages. An ideal record extractor should achieve the following: 1) all data records in the data region are extracted and 2) for each extracted data record, no data item is missed and no incorrect data item is included.

Data record extraction is to discover the boundary of data records. That is, we attempt to determine which blocks belong to the same data record. We achieve this in the following three phases:

1. **Phase 1:** Filter out some noise blocks.
2. **Phase 2:** Cluster the remaining blocks by computing their appearance similarity.
3. **Phase 3:** Determine data record borderline by reorganizing chunks.

B. Data Item Extraction

An information record can be viewed as the depiction of its comparing article, which comprises of a gathering of information things and some static layout writings. In genuine applications, these extricated organized information records are put away (regularly in social tables) at information thing level and the information things of the same semantic must be put under the same section. We as of now said that there are three sorts of information things in information records: compulsory information things, discretionary information things, and static information things. We extricate every one of the three sorts of information things. Note that static information things are frequently annotations to information and are valuable for future applications, for example, Web information annotation. Beneath, we concentrate on the issues of portioning the information records into a grouping of information things and adjusting the information things of the same semantics together. Update that information thing mining is divergent from information record mining; the past accentuations on the leaf knobs of the Visual Block chain of command, while the second accentuations on the tyke squares of the information region in the Visual Block progressive system.

C. Vision Wrapper Generation

Since all profound Web pages from the same Web database have the same visual format, once the information records and information things on a profound Web page have been separated, we can utilize these removed information records and information things to produce the extraction wrapper for the Web database so that new profound Web pages from the same Web database can be prepared utilizing the wrappers rapidly without reapplying the whole extraction process.

D. Vision Based Data Record Wrapper

Given a profound Web page, vision-based information record wrapper first finds the information area in the Visual Block tree, and afterward, removes the information records from the youngster pieces of the information locale.

E. Vision Based Data Item Wrapper

The fundamental thought of our vision-based information thing wrapper is depicted as takes after: Given a succession of properties {a1, a2 . . . , an} acquired from the specimen page and an arrangement of information things {item1, item2, . . .,, itemm} got from another information record, the wrapper forms the information things keeping in mind the end goal to choose which trait the present information thing can be coordinated to. For thing and aj, on the off chance that they are the same on f, l, and d, their match is perceived. The wrapper then judges whether itemiþ1 and ajþ1 are coordinated next, and if not, it judges itemi and ajþ1. Rehash this procedure until all information things are coordinated to their right properti

Comparisons of different Data Extraction methods

Sr. No.	Method	Advantage	Disadvantage
1	**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.
2	**ViDE**	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.
3	**LBDRF**	Mines the data from unseen net sources by building hierarchy.	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.
4	**Hybrid (MDR + ViDE)**	Can mine the data from multi data region web pages based visual cues	Multiple passes are required to process the deep web page.

Table1: Comparison between Existing and Hybrid Technique

V. Conclusion and future work

This paper examines the current procedures to separate the information from profound site pages. Some system just procedures HTML label tree and dialect subordinate. The illustration framework we secured here MDR and LBDRF. The other promising method is ViDE which is extricating the profound web information taking into account visual signs of profound website pages alongside label tree however just equipped for mining single information locales profound pages. The half breed approach which is introduced in this paper is great arrangement towards extraction of web information from profound site pages. The execution of this framework is as of now under procedure with the mean to evacuate its disservices by applying delicate processing methodologies.

References

[1] J. Madhavan, S.R. Jeffery, S. Cohen, X.L. Dong, D. Ko, C. Yu, and A. Halevy, “Web-Scale Data Integration: You Can Only Afford to Pay As You Go,” Proc. Conf. Innovative Data Systems Research (CIDR), pp. 342-350, 2007.

[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.J. Clerk Maxwell, *A Treatise on Electricity and Magnetism,* 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68-73.

[3] Wei Liu, Xiaofeng, member IEEE, and Weiyi Meng, Member, IEEE, “ViDE: A Vision-Based Approach for Deep Web Data Extraction”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. X, XXXXXXX 2010.

[4] Ji Ma, Derong Shen and TieZheng Nie, “DESP: An Automatic Data Extractor on Deep Web Pages”, Web Information Systems and Applications Conference (WISA), 2010 7th Publication Year: 2010, Page(s): 132 – 136.

[5] 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.

[6] V. Crescenzi and G. Mecca, “Grammars Have Exceptions,” Information Systems, vol. 23, no. 8, pp. 539-565, 1998.

[7] J. Hammer, J. McHugh, and H. Garcia-Molina, “Semi-structured Data: The TSIMMIS Experience,” Proc. East-European Workshop Advances in Databases and Information Systems (ADBIS), pp. 1-8, 1997.

[8] G.O. Arocena and A.O. Mendelzon, “WebOQL: Restructuring Documents, Databases, and Webs,” Proc. Int’l Conf. Data Eng. (ICDE), pp. 24-33, 1998.

[9] N. Kushmerick, “Wrapper Induction: Efficiency and Expressiveness,” Artificial Intelligence, vol. 118, nos. 1/2, pp. 15-68, 2000.

[10] C.-N. Hsu and M.-T. Dung, “Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web,” Information Systems, vol. 23, no. 8, pp. 521-538, 1998.

[11] I. Muslea, S. Minton, and C.A. Knoblock, “Hierarchical Wrapper Induction for Semi-Structured Information Sources,” Autonomous Agents and Multi-Agent Systems, vol. 4, nos. 1/2, pp. 93-114, 2001.

[12] A. Sahuguet and F. Azavant, “Building Intelligent Web Applications Using Lightweight Wrappers,” Data and Knowledge Eng., vol. 36, no. 3, pp. 283-316, 2001.

[13] L. Liu, C. Pu, and W. Han, “XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources,” Proc. Int’l Conf. Data Eng. (ICDE), pp. 611-621, 2000.

[14] D. Buttler, L. Liu, and C. Pu, “A Fully Automated Object Extraction System for the World Wide Web,” Proc. Int’l Conf. Distributed Computing Systems (ICDCS), pp. 361-370, 2001.

[15] V. Crescenzi, G. Mecca, and P. Merialdo, “RoadRunner: Towards Automatic Data Extraction from Large Web Sites,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 109-118, 2001.

[16] C.-H. Chang, C.-N. Hsu, and S.-C. Lui, “Automatic Information Extraction from Semi-Structured Web Pages by Pattern Discovery,” Decision Support Systems, vol. 35, no. 1, pp. 129-147, 2003.

[17] B. Liu, R.L. Grossman, and Y. Zhai, “Mining Data Records in Web Pages,” Proc. Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 601-606, 2003.

[18] Y. Zhai and B. Liu, “Web Data Extraction Based on Partial Tree Alignment,” Proc. Int’l World Wide Web Conf. (WWW), pp. 76-85, 2005.

[19] D.W. Embley, Y.S. Jiang, and Y.-K. Ng, “Record-Boundary Discovery in Web Documents,” Proc. ACM SIGMOD, pp. 467- 478, 1999.

[20] C.-H. Chang, M. Kayed, M.R. Girgis, and K.F. Shaalan, “A Survey of Web Information Extraction Systems,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 10, pp. 1411-1428, Oct. 2006.

[21] 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 c 2009 IEEE pages 370-373.