International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Overview of Gathering Information from Unfathomable Web Pages using Language Independent and Dependent Procedures

V.Annapoorna†* and M.V.Kishore†

Anil Neerukonda Institute of Technology and Sciences (ANITS), Visakhapatnam, Andhra Pradesh, India

Accepted 10 Jan 2015, Available online 01 Feb 2015, **Vol.5, No.1 (Feb 2015)**

Abstract

World Wide Web has more and more online Web pages which can be searched through their Web query interfaces. The query results will be retrieved based on the visual information of Web pages such as the information related to Web page layout (location and size) and font and the returned data records are enwrapped in dynamically generated Web pages. Extracting structured data from unfathomable Web pages is a challenging problem due to the underlying intricate structures of such pages. Unfathomable web page means all the content of the Web That Is not direct accessible through hyper links. In particular HTML forms, Web services. Traditional web extractors focus only on the surface web while the unfathomable web page keeps expanding behind the scene. The large number of techniques has been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. In this paper, we will propose a novel and UWPE BASED (unfathomable web page extractor) approach by using the visual features of the web pages. These visual features are used to construct a visual block tree for extracting data from the unfathomable web pages .our approach is language independent.

Keywords: *Unfathomable web data Extraction, vision based approach, language-independent, dependent, search engine.*

Introduction

World Wide Web has close to millions of searchable information sources. The number of Web databases has reached 25 million according to recent survey (I. Muslea, S. Minton, and C.A. Knoblock, 2001) .These searchable information sources include both search engines and Web databases. Web database keeps expanding every day, which drives the focus on researches towards unfathomable web mining. The information in a web database can be fetched only through its web query interface. These Web databases are queried for particular information and the query result is enwrapped to form a dynamic web page called the unfathomable web page. It is almost impossible for the search engines to retrieve this information and hence this is called unfathomable web or hidden web. The result objects obtained from the query submitted, is displayed in the form of data records. The retrieved information (query results) is enwrapped on response pages returned by these systems in the form of data records(D. Cai, X. He, J.-R. Wen, and W.-Y. Ma,2004). Data records are usually displayed visually neatly on Web browsers to ease the consumption of users. These special Web pages are generated dynamically and are hard to index by traditional crawler based Search engines, such as Google and Yahoo. In this Paper, we call this kind of special Web pages unfathomable Web Pages. Each data record on the unfathomable Web pages corresponds to an object. For instance, Fig. 1 shows a typical Unfathomable Web page from Amazon.com. On this page, the Books are presented in the form of data records, and each Data record contains some data items such as title, author, Etc. In order to ease the consumption by human users, Most Web databases display Data records and data items regularly on Web browsers. However, to make the data records and data items which are needed in many Applications such as unfathomable Web crawling and Meta searching, the structured data need to be extracted from the unfathomable Web Pages.

In this paper, we study the problem of automatically extracting the structured data, including data records and Data items, from the unfathomable Web pages. The problem of Web data extraction has received a lot of attention in recent years and most of the proposed solutions are based on analyzing the HTML source code or the tag Trees of the Web pages (see Section 2 for a review of these Works). These solutions have the following main limitations: First, they are Web-page-programming-language-dependent, Or more precisely, HTML-dependent. As most Web pages are written in HTML, it is not surprising that all previous solutions are based on analyzing the HTML

*Corresponding author: **V.Annapoorna**

Figure 1: Example of unfathomable web content

source Code of Web pages. However, HTML itself is still evolving (from version 2.0 to the current version 4.01, and version 5.0 Is being drafted (daisen.cc.kyushu-u.ac,2009) and when new versions or new tags Are introduced, the previous works will have to be Amended repeatedly to adapt to new versions or new tags. Furthermore, HTML is no longer the exclusive Web page Programming language, and other languages have been Introduced, such as XHTML and XML (combined with XSLT and CSS).

The previous solutions now face the Following dilemma: should they be significantly revised or even abandoned? Or should other approaches be proposed to accommodate the new languages? Second, they are Incapable of handling the ever-increasing complexity of HTML source code of Web pages. Most previous works have not considered the scripts, such as java script and CSS. In the HTML files. In order to make Web pages vivid and Colorful, Web page designers are using more and more Complex java script and CSS(J. Hammer, J. McHugh, and H. Garcia-Molina,(1997). Based on our observation from A large number of real Web pages, especially unfathomable WebPages, the underlying structure of current Web pages is More complicated than ever and is far different from their Layouts on Web browsers. This makes it more difficult for existing solutions to infer the regularity of the structure of Web pages by only analyzing the tag structures. Meanwhile, to ease human users' consumption of the information retrieved from search engines; good template designers of unfathomable Web pages always arrange the data records and the data items with visual regularity to meet the reading habits of human beings.

For example, all the data records in Fig. 1 are clearly separated, and the data items of the same semantic in different data records are similar on layout and font. In this paper, we explore the visual regularity of the data records and data items on unfathomable Web pages and propose a novel UWPE BASED approach, Unfathomable web page extractor(UWPE), to extract structured results from unfathomable Web pages automatically. UWPE is primarily based on the visual features human users can capture on the unfathomable Web pages while also utilizing some simple non-visual information such as data types and frequent symbols to make the solution more robust. UWPE consists of two main components, UWPE BASED Data Record extractor (UWPERE) and UWPE BASED Data Item extractor (UWPEIE). By using visual features for data extraction, UWPE avoids the limitations of those solutions that need to analyze complex Web page source files. Our approach employs a four-step strategy.

First, given a sample unfathomable Web page from a Web database, obtain its visual representation and transform it into a Visual Block tree which will be introduced later; second, extract data records from the Visual Block tree; third, partition extracted data records into data items and align the data items of the same semantic together; and fourth, generate visual wrappers (a set of visual extraction rules) for the Web database based on sample unfathomable Web pages such that both data record extraction and data item extraction for new unfathomable Web pages that are from the same Web database can be carried out more efficiently using the visual wrappers. To our best knowledge, although there are already some works (D. Buttler, L. Liu, and C. Pu,(2001),(D. Cai, X. He, J.-R. Wen, and W.-Y. Ma,2004), (A. Sahuguet and F. Azavant,2001), (J. Wang and F.H. Lochovsky,2003), that pay attention to the visual information on Web pages, our work is the first to perform unfathomable Web data extraction using primarily visual features.

Our approach is independent of any specific Web page programming language. Although our current implementation uses the VIPS algorithm(D. Cai, X. He, J.-R. Wen, and W.-Y. Ma,2004) to obtain a unfathomable Web page's Visual Block tree and VIPS needs to analyze

the HTML source code of the page, our solution is independent of any specific method used to obtain the Visual Block tree in the sense that any tool that can segment the Web pages into a tree structure based on the visual information, not HTML source code, can be used to replace VIPS in the implementation of UWPE. In this paper, we also propose a new measure, revision, to evaluate the performance of Web data extraction tools.

It is the percentage of the Web databases whose data records or data items cannot be perfectly extracted (i.e., at least one of the precision and recall is not 100 percent). For these Web databases, manual revision of the extraction rules is needed to achieve perfect extraction. In summary, this paper has the following contributions:

1) A novel technique is proposed to perform data extraction from unfathomable Web pages using primarily visual features. We open a promising research direction where the visual features are utilized to extract unfathomable Web data automatically.
2) A new performance measure, revision, is proposed to evaluate Web data extraction(G.O. Arocena and A.O. Mendelzon,(1998)tools. This measure reflects how likely a tool will fail to generate a perfect wrapper for a site.

Searching for information on the Web is not an easy task. Searching for personal information is sometimes even more complicated. Below are several common problems we face when trying to get personal details from the web:

• Majority of the Information is distributed between different sites.
• It is not updated.
• Multi-Referent ambiguity(C.-H. Chang, C.-N. Hsu, and S.-C. Lui,2003) – two or more people with the same name.
• Multi-morphed ambiguity which is because one name may be referred to in different forms.
• In the most popular search engine Google, one

can set the target name and based on the extremely limited facilities to narrow down the search, still the user has 100% feasibility of receiving irrelevant information in the output search hits. Not only this, the user has to manually see, open, and then download their respective file which is extremely time consuming(V. Crescenzi and G. Mecca,1998). The major reason behind this is that there is no uniform format for personal information. Maximum of the past work is based on exploiting the link structure of the pages on the web, with hypothesis that web pages belonging to the same person are more likely to be linked together.

Methodology

Information from unfathomable web pages can be extracted as shown in the figure 1. One of the key challenges that needs to be overcome to make the project functionality a reality, is to build an advance query system that is capable of reaching high disambiguation quality. The project work is targeted to design an advance version of the unfathomable extraction tool using Clustering Algorithm. In this research work, the focus is mainly on querying for personal information of scientists and researchers. The user has to set the proper target name for search, which when completed, the user will receive complete PDF and image files based on the key of the search. Each group of information items (cluster) will be defined by its key and the user make the choice. The result page will be produced from the chosen clusters.

For making the search operationally accurate, we will assume the usage of IEEE doc files as they carry a standard format of name, e-mail ID, publication, images, and links to the full images. The visual information of Web pages, which has been introduced above, can be obtained through the programming interface proUWPEd by Web browsers. A Visual Block tree is actually a segmentation of a Web page. The root block represents the whole page, and each block in the tree corresponds to a rectangular region on the Web page. The leaf blocks are the blocks that cannot be segmented further, and they represent the minimum semantic units, such as continuous texts or images.

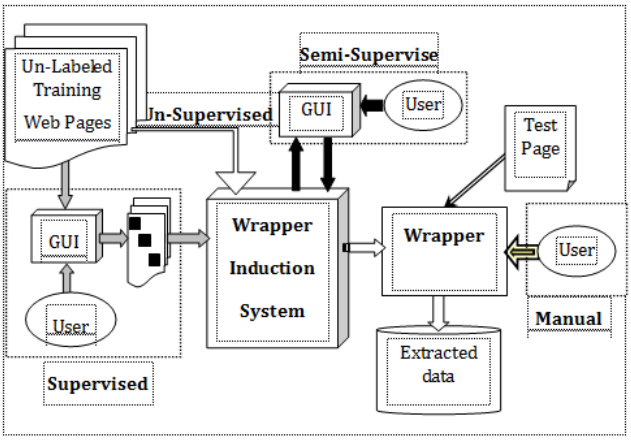


Figure 2: Data Extraction process from unfathomable web

Visual Information of Web Pages

The information on Web pages consists of both texts and images (static pictures, flash, UWPEo, etc.). The visual information of Web pages used in this paper includes mostly information related to Web page layout (location and size) and font.

Web Page Layout

A coordinate system can be built for every Web page. The origin locates at the top left corner of the Web page. The X-axis is horizontal left-right, and the Y-axis is vertical top down. Suppose each text/image is contained in a minimum bounding rectangle with sides parallel to the axes. Then, a text/image can have an exact coordinate (x, y) on the Web page. Here, x refers to the horizontal distance between the origin and the

left side of its corresponding rectangle, while y refers to the vertical distance between the origin and the upper side of its corresponding box. The size of a text/ image is its height and width. The coordinates and sizes of texts/images on the Web page make up the Web page layout.

Font

The fonts of the texts on a Web page are also very useful visual information, which are determined by many attributes as shown in Table 1. Two fonts are considered to be the same only if they have the same value under each attribute. We have different font formats they are:

1. True type fonts (TTF)
2. Open type fonts (OTF)
3. The Web Open Font Format (WOFF)
4. SVG Fonts
5. Embedded Open Type Fonts (EOT)

Table 1 Font attribute and example

Font factor	Example	Font factor	Example
Size	A (10pt)	underline	A
face	A(Sans Serif)	italic	*A*
color	A (red)	weight	**A**
strikethrough	~~A~~	frame	A

The VIPS Algorithm

In the VIPS algorithm, the UWPE BASED content structure of a page is deduced by combining the DOM structure and the visual cues.

First, DOM structure and visual information, such as position, back ground colour, font size, font weight, etc., are obtained from a web browser. Then, from the root node, the visual block extraction process is started to extract visual blocks of the current level from the DOM tree(www.w3.org,2009)based on visual cues. Every DOM node is checked to judge whether it forms a single block or not. If not, its children will be processed in the same way. When all blocks of the current level are extracted, they are put into a pool. Visual separators among these blocks are identified and the weight of a separator is set based on properties of its neighbouring blocks. After constructing the layout hierarchy of the current level, each newly produced visual blocks is checked to see whether or not it meets the granularity requirement. If no, this block will be further partitioned. After all blocks are processed, the final UWPE BASED content structure for the web page is outputted. Below we introduce the visual block extraction, separator detection and content structure construction phases respectively.

Visual Block Extraction

In this phase, we aim at finding all appropriate visual blocks contained in the current sub-tree. In general, every node in the DOM tree can represent a visual block. However, some "huge" nodes such as

<TABLE> and <P> are used only for organization purpose and are not appropriate to represent a single visual block. In these cases, the current node should be further diUWPE and replaced by its children. On the other hand, we may not extract all leaf nodes in the DOM(www.w3.org,2009)tree due to their high volume. At the end of this step, for each node that represents a visual block, its DoC value is set according to its intra visual difference. This process is iterated until all appropriate nodes are found to represent the visual blocks in the web page. Some important cues are used to produce heuristic rules in the algorithm are:

• Tag cue: Tags such as <HR> are often used to separate different topics from visual perspective. Therefore we prefer to diUWPE a DOM node if it contains these tags.
• Colour cue: We diUWPE a DOM node if its background colour is different from one of its children's.
• Text cue: If most of the children of a DOM(Z. Nie, J.-R. Wen, and W.-Y. Ma,2007) node are Text nodes (i.e., no tags surround them), we do not diUWPE it.
• Size cue: We prefer to diUWPE a DOM node if the Standard deviation of size of its children is larger than a threshold.

Visual Separator Detection

After all blocks are extracted, they are put into a pool for visual separator detection. Separators are horizontal or vertical lines in a web page that visually cross with no blocks in the pool. From a visual perspective, separators are good indicators for discriminating different semantics within the page. A visual separator is represented by a 2-tuple: (Ps, Pe),where Ps is the start pixel and Pe is the end pixel. The width of the separator is calculated by the difference between these two values.

Data record extraction aims to discover the boundary of data records and extract them from the unfathomable 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. Instead of extracting data records from the unfathomable Web page directly, we first locate the data region, and then, extract data records from the data region. PF1 and PF2 indicate that the data records are the primary content on the unfathomable Web pages and the data region is centrally located on these pages. The data region corresponds to a block in the Visual Block tree. We locate the data region by finding the block that satisfies the two position features. Each feature can be

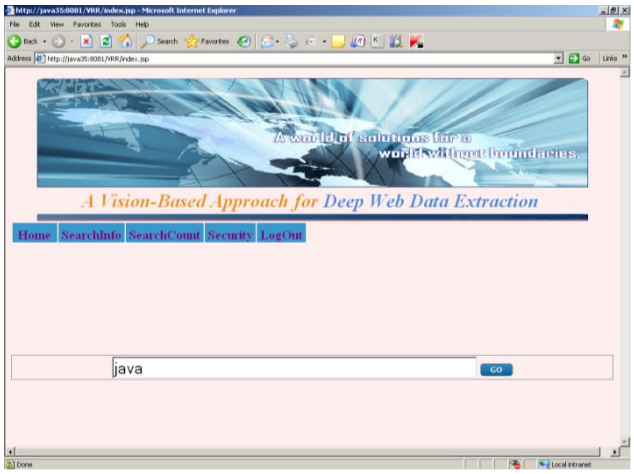
considered as a rule or a requirement. The first rule can be applied directly, while the second rule can be represented by areab=areapage> Tregion, where areab is the area of block b, areapage is the area of the whole unfathomable Web page, and Tregion is a threshold. The threshold is trained from sample unfathomable Web pages. If more than one block satisfies both rules, we select the block with the smallest area. Though very simple, this method can find the data region in the Visual Block tree accurately and efficiently. We have implemented an operational unfathomable Web data extraction system for UWPE based on the techniques we proposed.

Web Crawler forms the back-bone of applications that facilitate Web Information Retrieval. In this paper we have presented the architecture and implementation details of our crawling system which can be deployed on the client machine to browse the web concurrently and autonomously. It combines the simplicity of asynchronous downloader and the advantage of using multiple threads. It reduces the consumption of resources as it is not implemented on the mainframe servers as other crawlers also reducing server management. The proposed architecture uses the available resources efficiently to make up the task done by high cost mainframe servers. A major open issue for future work is a detailed study of how the system could become even more distributed, retaining though quality of the content of the crawled pages.

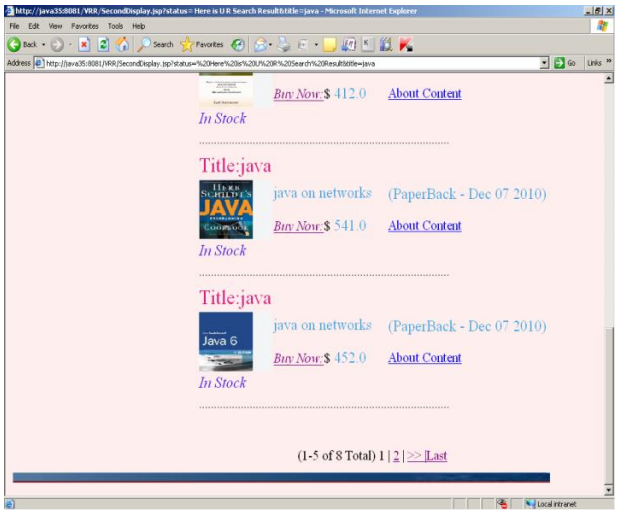
Due to dynamic nature of the Web, the average freshness or quality of the page downloaded need to be checked, the crawler can be enhanced to check this and also detect links written in JAVA scripts or VB scripts and also provision to support file formats like XML, RTF, PDF, Microsoft word and Microsoft PPT can be done.

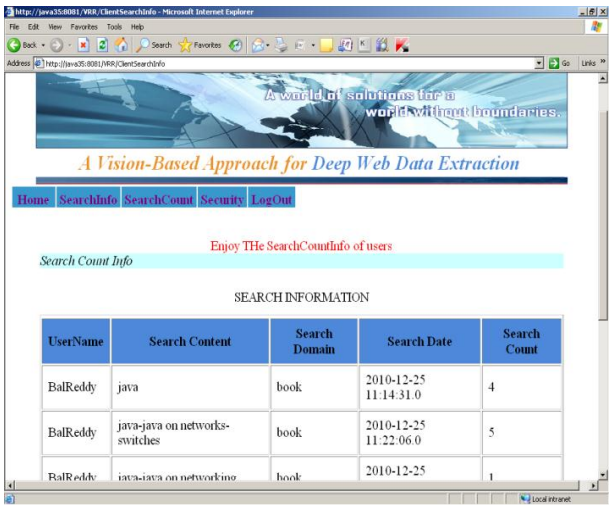
Data Analysis

User window as shown in below. For example user enters java keyword to search in search engine.



Then search engine extracts webpage links and data present in unfathomable web also as shown in below figures:





Conclusion

The part of the World Wide Web that is not discoverable by means of standard search engines, including password-protected or dynamic pages and encrypted networks. Such a web is called as unfathomable web. It becomes harder to find relevant and recent information from the unfathomable web due to the Underlying intricate structure and dynamic nature of the Web; we present a new approach using multiple HTTP navigation to WWW to access the unfathomable web content.

Initially, the user specifies the start URL from the GUI UWPE. It starts with a URL to visit. As the crawler visits the URL, it identifies all the hyperlinks in the web page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited and it stops when it reaches more than five level from every home pages of the websites visited and it is concluded that it is not necessary to go unfathomable than five levels from the home page to capture most of the pages actually visited by the people while trying to retrieve information from the internet. This approach uses the "VIPS ALGORITHM", which includes "visual block extraction, visual separator detection" to acquire structured data from the

unfathomable web pages and thus retrieves the unfathomable web content.

References

Wei Liu, Xiaofeng Meng, Member, IEEE, and Weiyi Meng,Member, IEEE,(2010), ViDE: A Vision-Based Approach for Deep Web Data Extraction, *IEEE transactions on knowledge and data engineering*, Vol. 22, No. 3.

G.O. Arocena and A.O. Mendelzon,(1998), WebOQL:Restructuring Documents, Databases and Webs, *Proc.Int'l Conf. Data Eng*. *ICDE*, pp. 24-33.

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

D. Cai, X. He, J.-R. Wen, and W.-Y. Ma,(2004), Block-Level Link Analysis, *Proc. SIGIR*, pp. 440-447.

D. Cai, S. Yu, J. Wen, and W. Ma,(2003), Extracting Content Structure for Web Pages Based on Visual Representation, *Proc. Asia Pacific Web Conf. (APWeb),* pp. 406-417.

C.-H. Chang, M. Kayed, M.R. Girgis, and K.F. Shaalan, (2006),A Survey of Web Information Extraction Systems, *IEEE Trans. Knowledge and Data Eng.*, vol. 18, no. 10, pp. 1411-1428

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

V. Crescenzi and G. Mecca,(1998),Grammars Have Exceptions, *Information Systems*, vol. 23, no. 8, pp.539-565.

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

D.W. Embley, Y.S. Jiang, and Y.-K. Ng,(1999) Record-Boundary Discovery in Web Documents, *Proc. ACMSIGMOD*, pp. 467-478.

W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krpl, and B. Pollak,(2007),Towards Domain Independent Information Extraction from Web Tables, *Proc. Int'l World Wide Web Conf. (WWW*), pp. 71-80.

J. Hammer, J. McHugh, and H. Garcia-Molina,(1997),Semistructured Data: The TSIMMIS Experience,*Proc.East-European Workshop Advances in Databases and Information Systems (ADBIS*), pp.1-8, 1997.

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

http://daisen.cc.kyushu-u.ac.jp/TBDW/, 2009.

http://www.w3.org/html/wg/html5/, 2009.

N.Kushmerick,(2000),Wrapper Induction: Efficiency and Expressiveness,*Artificial Intelligence*, vol. 118, nos. 1/2, pp. 15-68.

W. Liu, X. Meng, and W. Meng,(2006),UWPE BASED Web Data Records Extraction, *Proc. Int'l Workshop Web and Databases (WebDB* '06), pp. 20-25, June.

L. Liu, C. Pu, and W. Han,(2000),XWRAP: An XMLEnabled Wrapper Construction System for Web Information Sources,*Proc. Int'l Conf. Data Eng. (ICDE)*, pp. 611-621

Y. Lu, H. He, H. Zhao, W. Meng, and C.T. Yu, (2007),Annotating Structured Data of the Unfathomable Web,*Proc.Int'l Conf. Data Eng. (ICDE*), pp. 376-385.

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

I. Muslea, S. Minton, and C.A. Knoblock, (2001),Hierarchical Wrapper Induction for Semi-Structured Information Sources," *Autonomous Agents and Multi-Agent Systems,* vol. 4, nos. 1/2, pp. 93-114.

Z. Nie, J.-R. Wen, and W.-Y. Ma,(2007), "Object-Level Vertical Search,*Proc. Conf. Innovative Data Systems Research (CIDR*), pp. 235-246.

A. Sahuguet and F. Azavant,(2001),Building Intelligent Web Applications Using Lightweight Wrappers,*Data and Knowledge Eng.,* vol. 36, no. 3, pp. 283-316

K. Simon and G. Lausen,(2005),ViPER: Augmenting Automatic Information Extraction with Visual Perceptions,*Proc. Conf.Information and Knowledge Management (CIKM)*, pp. 381-388.

R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, (2004),Learning Block Importance Models for Web Pages, *Proc. Int'l World Wide Web Conf. (WWW*), pp. 203-211.

J. Wang and F.H. Lochovsky,(2003),Data Extraction and Label Assignment for Web Databases,*Proc. Int'l World Wide Web Conf. (WWW*), pp. 187-196.

(2006),A survey of Web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411-1428.