A Vision Based Approach for Web Data Extraction Using Enhanced Cocitation Algorithm

R.Vijay[1], Dr. K. Prasadh [2]

[1]Research Scholar, M.S University
Tamil Nadu, India

[2]VK college of Engineering and Technology
Parippally, Kerala.

Abstract

Normally, the World Wide Web maintains a set of databases which can store several data records retrieved by web query interface. The information maintained in web is hidden in the database that can be retrieved through dynamic script pages are termed as deep web content. These forms of deep web contents are normally accessed by the web queries, but, extracting the structured data from web database involves complexity. To address the issue, Wei Liu et. al., presented programming language independent vision based approach that use the visual features of deep web pages for web data extraction. The vision based approach also includes the process of extraction of data record and data item. But the unsolved issues in Liu's vision based approach is that it not only process the deep web pages in one data region of the web page but also consumes additional time to extract the visual information of web pages. To address the demerit present in ViDE, a novel technique called vision based approach for deep web data extraction is presented. In this work, we describe a framework that processes the deep web pages present in multi data regions. The framework uses enhanced co-citation algorithm that, instead of developing a new set of APIs for the extraction of visual information, the algorithm retrieve the visual information of the deep web pages directly from the web database. Empirical studies with large set of database for web data extraction demonstrate that the performance of the proposed vision based approach [VBEC] are capable of offering high precision while enabling efficient and accurate recall value of similar queries with better time consumption compared to other extraction processes.

Keywords: Deep web data, vision based approach, multi data regions, co-citation algorithm, visual features, and web data extraction.

1. Introduction

Nowadays, there is a tremendous increase of usage in World Wide Web and has secured to one million searchable information sources. These searchable information sources includes both Web databases and search engine. By relocating queries to search interfaces of these information sources, practical information from them is readily accesses. For instance, the processed information is returned to the requested pages using several techniques consisting of data records, each of which are again related with the entity for example a document or a book. Data records are significantly revealed carefully on Web browsers to alleviate the utilization of human users. On the other hand, to construct the data records which are taken out from the machine process able, which is required in different applications for instance deep web crawling and meta-searching, it is highly required to be taken out from the response set of web pages.

The existing deep web page problems have come up with several solutions which are highly supported by examining the HTML source files of the response set of web pages. Even though they can realize sensibly high accuracies in the provided experimental outcomes, the present studies of this crisis have numerous limitations; HTML-based techniques suffer from the subsequent problems:

HTML itself is still growing and when novel versions or marks emerge, the earlier solutions have to be processed frequently to settle in to novel provision and original tags.

- Earlier versions on HTML-based techniques, only measured the HTML file that do not include JavaScript and CSS. As more and more web pages use more complex scripts, the applicability of the presented resolutions become poorer.
- If HTML is restored by a novel language in the prospect, then earlier resolutions have to be adjusted significantly or even discarded, and other techniques must be presented to put up the novel language.
- Finally, conventional performance procedures, precision and recall, do not completely discard the excellence of the extraction.

There are previously some parts that examine design construction of web pages. They attempt to electively symbolize and recognize the construction of web pages, which are substantial formation independent. It is recognized that web pages are employed to distribute information for humans to peruse, and not considered for computers to haul out information repeatedly. Based on such deliberation, in this paper we propose a technique to haul out data records mechanically supported with visual

illustration of web pages. Our VBEC approach follows three-steps to realize this objective.

2. Literature Review

Numerous techniques have been discussed in the literature for extracting out the information from Web pages. The most primitive techniques are the instruction manual approaches in which languages were considered to help out programmer in building wrappers to recognize and take out all the preferred data items/fields. In recent times [1], presented a new vision based technique for the extraction of web page by developing a set of API.

The crisis of extracting out the data records on the response web pages revisited from web databases or search engines remained a critical task. World Wide Web has created a demanding crisis in taking out the relevant data [2]. The huge number of techniques has been presented to deal with this crisis, but all of them contain intrinsic limitations since they are Web-page-programming-language reliant or sovereign [3]. The paper [4] studied the crisis of taking out the data records on the response web pages revisited from web databases or search engines. A new and language self-sufficient system is presented to resolve the data extraction crisis.

With the intention of develop the efficiency and decrease physical efforts, most current researches center on regular approaches in place of corporeal or semiautomatic ones. The automatic data alignment technique in [5] proposes a clustering technique to achieve alignment supported with five features of data items, as well as font of text. On the other hand, this technique is chiefly text-based and tag-structure-based, while the resented technique is mainly visual-information-based. So the efficiency of this approach is very less.

A search engine revisited consequence page might hold search results that are prearranged into numerous dynamically measured sections in reply to a user query. Moreover, such a outcome page frequently contains information unrelated to the query, such as information connected to the hosting location of the search engine. In [6], the author presented a technique to robotically produce wrappers for taking out search result records from all active sections on outcome pages revisited by search engines.

Dynamic web page has a huge quantity of pages, high-value data and elevated modularity construction. Along with these feature, the paper [7] developed routine web information taking out scheme based on page clustering. Web pages are modeled by closing data values to reconstructed templates. Physical data extraction from semi supervised web pages is a complex job. The paper [8] focused on revise of different automatic web data extraction methods. OLERA [9] is semi supervised information taking out tool [10] where user can produce extraction rules consistent with training pages. But the information charge does not provide appropriate applicable web pages based on users' query. The above discussed issues are resolved in this work by implementing the vision based approach for web page extraction using enhanced co-citation algorithm under multi-data region.

3. Proposed Methodology

In this paper, the process of deep web data extraction from multi data regions is done effectively by adapting the enhanced co-citation algorithm. The proposed vision based approach for web data extraction is processed under three different phases. The first phase describes the process of obtaining visual representation of the deep web pages from the web database using enhanced co-citation algorithm. The second phase elaborates the process of retrieving data records from the web pages from multi data regions. Finally, the third phase describes the process of extracting data items from the data records and forms an efficient web data page. The architecture diagram of the vision based approach is shown in Fig. 1.

From the figure (Fig. 1), it is being noted that the process of VBEC is clearly explained with the step by step procedure. At first, the web pages are extracted from the web database using enhanced co-citation algorithm. Second, once the process of extraction of web pages is accomplished, next the data records are extracted from the multi- data regions and are further processed. Based on users' queries, the third and final step involved in VBEC is that the exact information is retrieved from the web database for an ease of use.

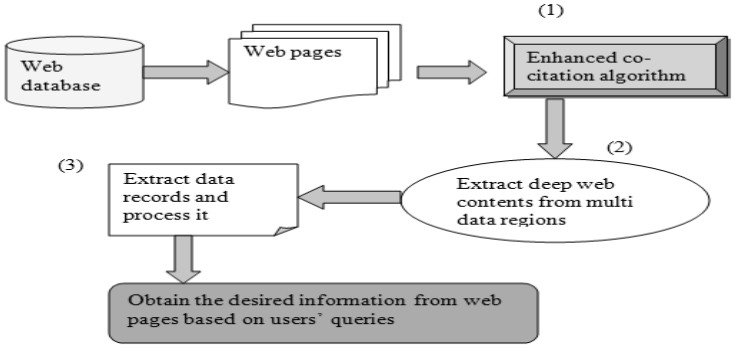

Fig. 1 Architecture diagram of VBEC

3.1 Extraction of web pages using enhanced co-citation algorithm

Web page consists of several set of information. The information present in the web page involves a text, image, video etc. Visual information of the web pages is normally related with the location, size and typeset of the web pages. The extraction of visual information of the web page is retrieved using enhanced co-citation algorithm.

The enhanced co-citation algorithm is processed based on defining some notations. Consider a set of two web pages as A and B. Let us further assume that page A is a parent of page B as page A might refer the web page B in its page. But page B has already derived from another web page which could be referred as B'. Now, these two different set of pages B and B' are said to be co-cited only if both share the same parent web page A. With this process, the degree of co-citation is measured. But some of the drawbacks related to co-citation algorithm were, the main problems like navigation links and retrieving duplicate set of pages. To address these kinds of issues related to co-citation algorithm, an enhanced co-citation algorithm is presented to retrieve the visual information of the web page directly from the web database.

The enhanced co-citation algorithm follows two strategies to extract the visual information of web pages from web database with respect to a users' query:

- Content-based, and
- Link-based

The content-based technique extracts the textual content of the users' query links and its siblings. The link-based technique utilizes only the link construction among the web pages collected for the enhanced Co-citation algorithm.

As stated above, the collection of web pages from the web database should detain the degree of union related with a topic pointed to by the links inside that group of web pages. If the topic of the link that is being focused on is restricted, then the parent webpage are expressed by the average similarity of the topics of the web page siblings based on users' query. The retrieved two topics are similar only if the sim (t1, t2) =1.

For the specified topic and sim functions, the notion of obtaining the content based Web Page extraction WP with URL u is obtained by identifying the similarities among the query URL's topic and the focus of the other siblings. It is specified as,

$$\text{Content based (WP)} = \frac{\sum \text{sim}(\text{topic}(u), \text{topic}(wp))}{|WP| - 1} \quad \ldots\ldots\ldots(1)$$

The content based enhanced co-cited algorithm retrieves the set of web pages to symbolize the average similarity of its similar web pages based on users' query with the obtained URL. Higher values point out stronger union with the siblings of WP concerning the topic of the query URL u. If the web page does not have includes any title tag, then the content based web page retrieval might be a complex one. In that case, a link based co-citation is utilized for extraction of web pages.

Let WD be a web database with a set of web pages wp1,w p2, . . . , wpM (as well as the query URL). The occurrence of overlap of the parent web pages are identified as P and Q, which is a sign of the union among the parent web pages on their children. We suggest that the parent web page P is measured even if more parents concur on more siblings of P. P consists of the utmost focal point in the case when all the parents related to the web pages are similar. This overlapping of web pages is normally accessed by the link based web data extraction which is expressed as,

$$\text{Link based (WP)} = \frac{\sum \frac{wp1 \cap wp2}{wp1 \cap wp2}}{|WP|} \quad \ldots\ldots\ldots(2)$$

This link-based view of web page extraction is computationally low-cost because the mandatory data is previously evaluated by the enhanced Co-citation algorithm. The pseudo code below describes the process of enhanced co-citation algorithm

Input: Web database WD, web pages WP
Identify the queries sent by the user
Identify the topic (t) of the query
    Perform content based web page extraction
If (WP has no title tag) **do**
    Perform link based web page extraction
End if
For (each parent web page p) **do**
    Customize the collected topics of the user query
End For
Rank sibling web pages based on its degree
Return (web pages based on highest degree)
End

With the queries, the web pages are efficiently extracted directly from the database based on following the subsequent steps describes in the above pseudo code.

3.2 Data record extraction

The main objective of Data record extraction is to determine the border line of data records and remove them from deep Web pages. Let us assume that the perfect record extractor should realize the subsequent properties and sees to that the following assumption are satisfied:

1) All data records present in multi data region are extracted and

2) For every extracted data record, no data item should be neglected and no erroneous data item be incorporated.

The figure (Fig. 2) below describes the general case of multi-data region of web database.

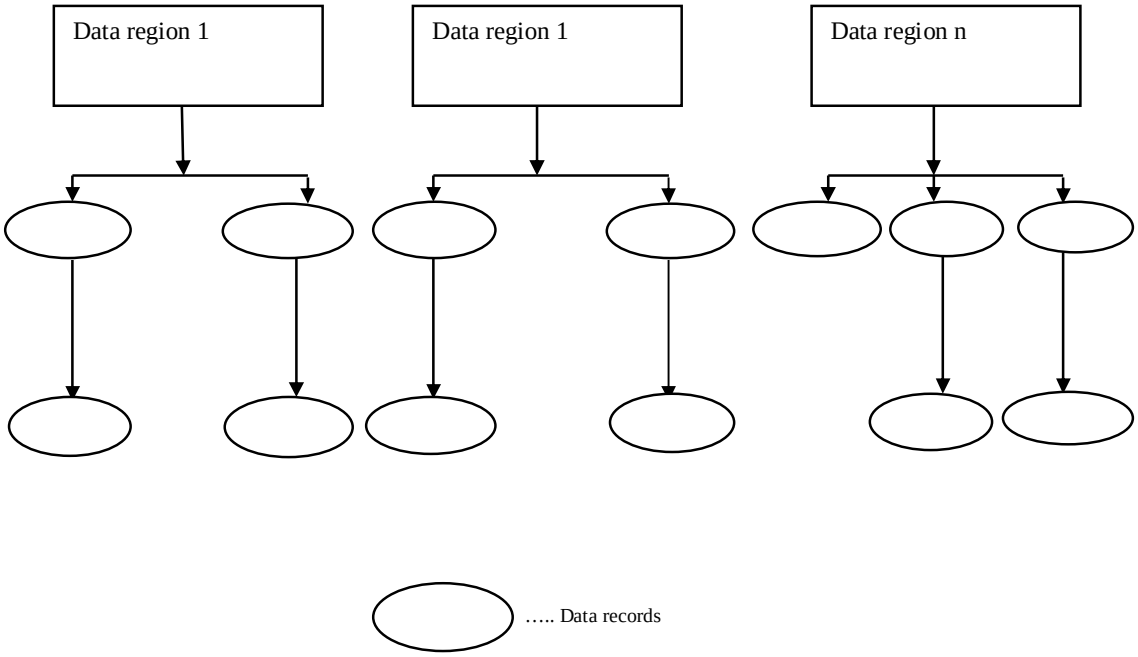

Fig. 2 Multi data region

Instead of retrieving the data records from the deep Web page straightforwardly, the data regions in the database are first established and then pull out data records from the data region. The data records are normally the prime content on the deep Web pages and the data region is centrally positioned on these web pages. The data region corresponds to a page P in the enhanced co-citation algorithm. We situate the multiple set of data regions by deciding the position of the web page that satisfies the two enhanced co-citation techniques.

In order to pull out data records from the multi-data region precisely, two facts are measured. First, there might be web pages Pi that do not fit in to any data record, for instance the geometric information. Then such web page(s) that do not fit into any data record are removed. Second, one data record might communicate to one or more web pages in the co-citation method, and the total number of web pages in which one data record hold is not permanent. In this case, the data record used more than once is referred for communication but not during the processing stage, that avoid redundancy.

3.3 Data item extraction

A data record is observed as the depiction of its equivalent data objects, which has a collection of data items and some fixed template texts. In genuine applications, these prearranged data records are gathered (often in relational tables) at data item rank and the data items of the similar semantic must be positioned beneath the similar column. When commencing the enhanced co-citation algorithm, every data record has been twisted into a series of data items during data record segmentation. Data item arrangement explain the way to support the data items of the similar by keeping the arrangement in semantic manner for the data items in every data record.

After the data regions are identified on a deep Web page WP, the data record is extracted from the respective web page and the space connecting two data records are identified. For the location data region in a new page, each data record is discovered by the visual comparison with the collected visual information using enhanced co-citation algorithm. The pseudo code below describes the process of enhanced co-citation algorithm for web data extraction in multi data region.

Input: Web database WD, Data records DR, Data items DI, Users U
Identify the queries Q sent by the user
For each query **do**
    Perform enhance co-citation algorithm described in section 3.1
End
For each extracted webpage P **do**
    Identify data region DR
    For each DR **do**
        Identify set of records DR
        Select DR which is more relevant

Identify DI present in the appropriate DR
End For
For each users' query **do**
Sort the Web pages P relevant to the user query related information
End for
End

The above algorithm describes the entire process of enhanced co-citation algorithm with vision based approach for web page extraction. The next section describes the experimental evaluation to estimate the performance of the proposed technique.

4. Experimental Evaluation

An experimental evaluation is done to estimate the performance of the vision based approach [VBEC] using enhanced co-citation algorithm. The VBEC experiments are done on a Pentium 4 1.9 GH, 512 MB PC. A large set of web database is utilized which consists of more than 10,000 entries of web database. These web databases are categorized into several domains. Several set of users submit their queries to extract the relevant information from the web database.

For each Web database, three set of user queries are presented and collect five deep Web pages holding three data records at any rate. With the set of web database, the enhanced co-citation algorithm is implemented to the database to extract the web pages directly from the web database. With the set of web pages, the data records and its corresponding data items are retrieved based on user query related information. The performance of the vision based approach [VBEC] using enhanced co-citation algorithm is measured in terms of

i) Precision
ii) Recall
iii) Time consumption

Precision measures the ratio of web data records extracted using enhanced co-citation algorithm based on the relevancy factor that is relevant to the search based on user query related information.

Recall measures the ratio of data records extracted relevant to the query that are successfully retrieved using enhanced co-citation algorithm.

Time Consumption refers to the time consumed to extract the web pages using enhanced co-citation algorithm.

Time Consumption (TC) = (Query Count)i * (Cycle Time)i

Where time consumption for a specific user is evaluated by the products of queries to be executed for n users (i=1,2,…,N) and cycle time for the specific ith user (i = 1,2,…..,N).

5. Results and Discussion

In this work, we have seen that the proposed VBEC approach efficiently extracts the web pages directly from the web database presents in multi-data region. For diverse set of web pages, the users' required data records are extracted and processed with the data items. After extracting the data records, the users' query related information is processed. The below table and graph describes the performance of the proposed vision based approach [VBEC] using enhanced co-citation algorithm.

Table 1: No. of entries in database vs. precision

No. of entries in database	Precision (%)	
	Proposed VBEC	Existing ViDE
100	85	75
200	97	77
300	89	80
400	90	85
500	91	90.5
600	92	91

The precision is made out in Table 1 based on the number of entries in the web database. The precision of the VBEC approach is compared with the existing ViDE approach.

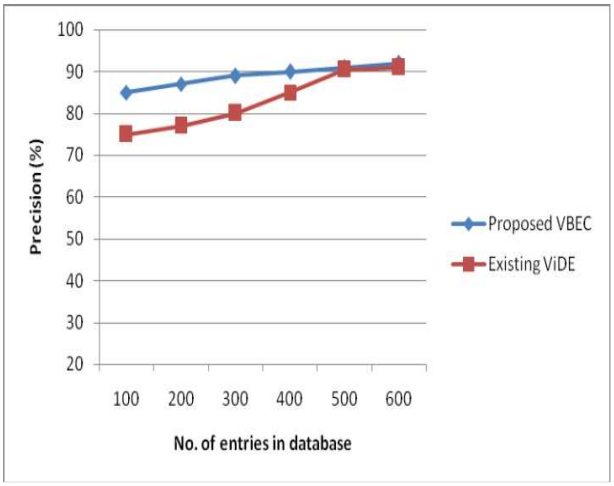

Fig. 3 No. of entries in database vs. precision

Fig. 3 describes the precision value of the extracted web pages based on the number of entries in the web database. The precision value is measured based on the rate at which the retrieved web pages are relevant to the users' queries. Compared to the existing ViDE approach, the VBEC approach provides better precision outcome since VBEC extracts the web pages directly from the web database using enhance co-citation algorithm. The enhanced co-citation algorithm efficiently filters the web pages and provides the desired outcome to the user. But in the existing ViDE approach, precision rate become less because of the presence of noise in the extracted web pages. So, the variance in the precision rate is 5-10% high in the proposed VBEC approach.

Table 2: No. of entries in database vs. recall

No. of entries in database	Recall (%)	
	Proposed VBEC	Existing ViDE
100	87	77
200	89	79.5
300	90	80
400	92	85
500	93	88.5
600	94	89

The recall is made out in Table 2 based on the number of entries in the web database. The recall of the VBEC approach is compared with the existing ViDE approach.

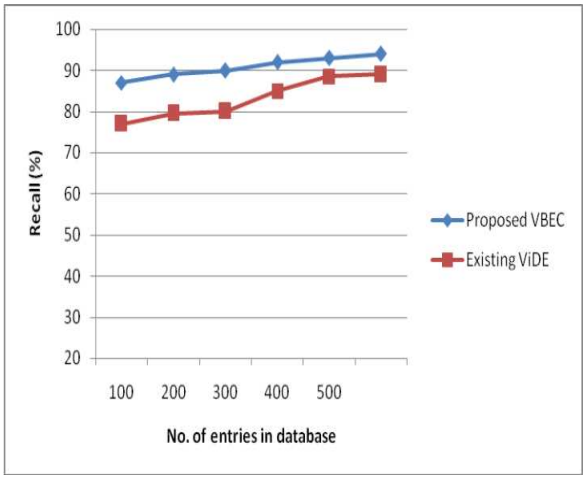

Fig. 4 No. of entries in database vs. recall

Fig. 4 describes the recall value of the extracted web pages based on the number of entries in the web database. The recall value is measured based on the rate at which the number of related web pages retrieved to the users' queries. Compared to the existing ViDE approach, the VBEC approach provides a better recall rate outcome with a variance 5-10% since VBEC uses multi-data region from the web database using enhance co-citation algorithm. But in the existing ViDE approach, recall rate become less because of the presence of noise in the extracted web pages.

Table 3: No. of entries in database vs. time consumption

No. of entries in database	Time consumption (sec)	
	Proposed VBEC	Existing ViDE
100	10	20
200	16	26
300	20	35
400	23	41
500	27	48
600	30	53

The time consumption required to extract the relevant deep web pages based on the entries of the web database is illustrated in Table 3. The time consumed for VBEC approach is compared with the existing ViDE approach.

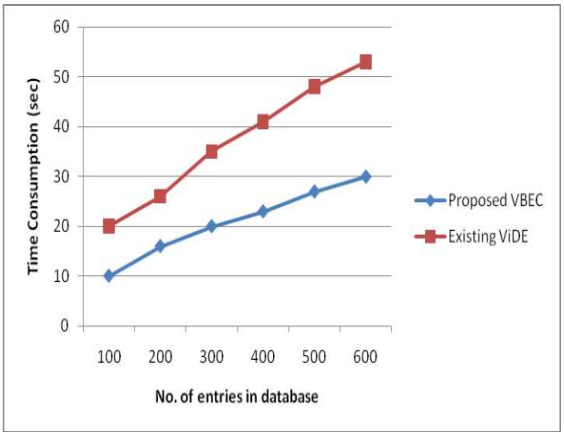

Fig. 5 No. of entries in database vs. time consumption

Fig. 5 describes the consumption of time required to extract the relevant deep web pages based on the entries of the web database. Time consumption is measured in terms of seconds. Compared to the existing ViDE approach, the proposed VBEC technique consumes less time to extract the web pages. Because the existing ViDE approach used IE as APIs for extracting the web pages from the web database. This consumes more time since for developing the IE tools for web page extraction. But in the proposed VBEC approach, the deep web pages are

extracted directly from the database instead of developing the APIs. So, the proposed VBEC approach consumes less time for web page extraction. The variance in the time consumption is 40-50% less in the VBEC approach for web page data extraction.

Finally, it is being depicted that the proposed VBEC approach efficiently extract the user relevant information from the web database based on their query by adapting the enhanced co-citation algorithm, data record extraction and data item extraction process.

6. Conclusion

In VBEC approach, we have discussed the problem of efficient usage of web database. Especially, we addressed this issue by enhanced co-citation algorithm and extracting the deep web pages directly from the database instead of using APIs. With the VBEC approach, users achieved a great chance of extracting the abundant information from the web database effectively. So, the users obtain the desired information in the deep Web pages returned by Web databases based on their queries. In this paper, we focused on extracting the structured Web data, as well as data record extraction and data item extraction by adapting the enhanced co-citation algorithm. At first, the existing work limitations is analyzed and implement the enhanced co-citation algorithm with vision based approach to obtain the visual information of Web pages. Based on extracted web pages, a vision-based approach is presented to haul out structured data from deep Web pages. Experimental evaluation is efficiently conducted with the sample set of web database and processes the users' queries. Performance evaluation revealed that the proposed VBEC approach is better with 5-10% higher in precision rate, with 5-10 % increase in recall and time consumption reduced from 40-50% compared to the existing ViDE approach.

References

[1] Wei Liu, Xiaofeng Meng, Weiyi Meng," ViDE: A Vision-Based Approach for Deep Web Data Extraction", Ransactions On Knowledge And Data Engineering, Vol. 22, 2010
[2] D. Raghu, Sridhar Reddy, Raja Jacob, "Dynamic Vision-Based Approach in Web Data Extraction", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2, No. 6, 2011, pp. 2734-2736
[3] S.Vasantha Kumari M.K. Chandrasekaren ,"Deep Web Data Extraction Using A Vision based Approach", IJART, Vol. 2, No 2, 2012,
[4] W. Liu, X. Meng, and W. Meng, "Vision-Based Web Data Records Extraction," Proc. Int'l Workshop Web and Databases (WebDB '06), June 2006, pp. 20-25.
[5] Y. Lu, H. He, H. Zhao, W. Meng, and C.T. Yu, "Annotating Structured Data of the Deep Web," Proc. Int'l Conf. Data Eng. (ICDE), 2007, pp. 376-385
[6] H. Zhao, W. Meng, and C.T. Yu, "Automatic Extraction of Dynamic Record Sections from Search Engine Result Pages," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2006, pp. 989-1000.
[7] Tianqi Yang , Taofen Qiu, "A method of automatic web information extraction based on page clustering", 9th World Congress on Intelligent Control and Automation (WCICA), 2011.
[8] Devika K, Subu Surendran, "An Overview of Web Data Extraction Techniques", International Journal of Scientific Engineering and Technology, Vol. 2, No. 4, 2013.
[9] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," Proc. International Conference on Information and Knowledge Management (CIKM), 2005.
[10] C.-H. Chang and S.-C. Kuo. OLERA: A semi-supervised approach for web data extraction with visual support. IEEE Intelligent Systems (SCI, EI), Vol. 19, No. 6, 2004, pp. 56–64.