American Journal of Applied Sciences

Original Research Paper

A Layout Based Detachment Approach for Extracting Content from Webpages

[1]Deepa Chandran and [2]Anna Saro Vijendran

[1]*Department of Information Technology, SNR Sons College, Coimbatore, India*
[2]*MCA, SNR Sons College, Coimbatore, India*

Article history
Received: 17-08-2013
Revised: 07-11-2013
Accepted: 10-07-2015

Corresponding Author:
Deepa Chandran
Department of Information Technology, SNR Sons College, Coimbatore, India
E-mail:deepaphd2013@gmail.com

Abstract: Enormous amount of useful information presented in Internet is usually formatted for the web users. But it is a really complex task to extract the relevant data from various web sources. Recently, various approaches for the extraction of data from the webpages were proposed. This study provides a simple but effective approach, named Layout Based Detachment Approach (LBDA). The proposed approach extracts the main content from the webpage by removing the irrelevant information like header-footer contents, navigation bars, advertisements and other noisy images. The proposed methodology uses the following techniques: Tag tree parsing to get the analysis structure, block acquiring page segmentation method to remove unwanted tags and data extraction to retrieve the necessary contents. The proposed approach eliminates noise and perform effective extraction of the main content blocks from the webpage and display of the essential content to the users. The performance of the proposed approach is evaluated using the performance metrics such as accuracy, precision, recall, execution time and memory usage. The implementation results obviously show that our proposed LBDA approach exhibits better performance than the existing heuristic approach.

Keywords: Webpage Content Extraction, Web Mining, DOM Tree Analysis, Web Structure Mining

Introduction

The World Wide Web (WWW) has gained popularity for dissipating and accumulating information, with the rapid growth of the Internet. Extracting useful information from webpages becomes a substantial task. Though, HTML format is designed more for human users to vision and does not comfortably lend itself to computerized processing for information extraction. An apparently simple visual block containing a few paragraphs of text is typically coded using tens of HTML nodes, some of which contain the text and others contain styling information to organize the layout. This problem is compounded by the fact that typical webpages contain significant amount of unrelated content mixed with the main content. Typically, apart from the essential content block, a webpage normally has blocked navigation bars, secrecy notices and appropriate hyperlinks, which are called noisy blocks. Even though such information components are functionally useful for human viewers for the website owners. They often blocks and essential webpage clustering, information extraction and information retrieval. Consequently, it is a crucial task to recognize and separate major content blocks from noisy blocks. A lot of research works has already been done in this field. Current repeated techniques are unacceptable as their outputs are not suitable for the query of the user.

To overcome this problem, we proposed layout based detachment approach to fully-automatically the extract content from the webpage. We are presenting an automatic approach to extract the main content of the webpage using tag tree and heuristics to filter the clutter and display the main content. The content related to query asked by the user, the title of the webpage, pop up ads, flashy advertisements, menus, unnecessary images and links are not relevant for a user querying the system for educational purposes. The conventional webpage text extractions normally have three steps: Web pretreatment such as tag adjustment, removing abortive description etc., then the webpage content extraction accedes to various text extraction algorithms, along with resulting a correction. During

Science Publications

the preprocessing stage, most of the conventional methods removes the invisible code that is irrelevant to the extraction of content.

In fact, still there is presence of lot of noises in the webpages refined by this technique. The accuracy of the text extraction algorithm will be affected by these noises.In order to resolve these problems, a novel approach is proposed for the extraction of the body text. During the preprocessing stage, the visible noise is removed to some extent along with the removal of the irrelevant invisible code. After preprocessing, more pure webpages are obtained to minimize the effect of noise interference during extraction of content from webpages. The relationship between the text length, punctuation marks and links are fully utilized for extracting the body text.In our work, a dynamic approach is proposed for the extraction of the main content of the webpage by using the tag tree parsing, block acquiring page segmentation method (BAPS) and data extraction and display the essential content. The BAPS refines the main block including the inner blocks and define the tree parsing and then verifies that parsing method has done on the basis of the algorithm.

The rest of this work is organized as follows: Section 2 describes about the related works. Section 3 summarizes our proposed layout based detachment approach for extracting content. Section 4 discussed the experiments and results achieved. Finally, we present the conclusion and future enhancements in section V.

Related Work

This section describes about some of the importantconventional research efforts for extracting content from the webpage. The web content structure is accessible for various purposes such as information retrieval, information extraction and web adaptation, as given in (Cai *et al*., 2003). The logic relationship of the web content is identified based on the visual layout information of the webpage. Then, the semantic webpage structure is efficiently represented by the web content structure, based on the logic relationship. The vision-based content structure of a webpage is obtained by using the visual cues, for successfully bridging the gap between the semantic structure and DOM structure. The page is partitioned and organized as a hierarchy relevant to guide the user for browsing the page. Xian *et al*. (2009) discussed about Most Benefit Approach (MBA) for the systematic organization of the candidate data source for the integration of internet user. The utility of specified state of the integrated system is quantified by the utility function.

Arasu and Garcia-Molina (2003) described an EXALG algorithm to extract structured data from the webpage collection generated from a common template. EXALG utilizes the equivalence classes and differentiating roles for the discovery of template. Sharma (2010) proposed an ontology based searching approach. Construction of the ontology for a domain is performed by using the hierarchical structure of attributes values and knowledge extracted from the query result pages. This study proposed three ontology modules:

- Construction of attribute value ontology
- Construction of attribute ontology
- User query formulation

Hammer *et al*. (1997) proposed a three-step strategy, while aiming at the response pages having collective data records. Initially, the response page is computed into a visual block tree, based on the visual representation. Then, the region accommodating all data records in the tree is ascertained. Finally, the data records are extracted from the ascertained region.

Lei *et al*. (2009) discussed about automatic content extraction, for the automaticidentification of data record limits, based on a set of heuristic rules. A heuristic approach is proposed for the discovery of record limits in the web documents. Then, the structure of the document is represented as a tree of nested HTML tags. Embley *et al*. (1999) establisheda subtree having the records of interest. A candidate separator tag is recognized and an optimal consensus separator tag is obtained, based on the accumulated heuristic. Liu *et al*. (2006) proposed a fully computerized technique for extracting the search result data records from the response pages that are dynamically generated by search engines. This is done by using the visual information about the response pages. Laender *et al*. (2002) described the characterization of the web data extraction tools. This is based on main methods used by each group of tools namely HTML-aware tools, Modeling-based tools, NLP tools, Ontology tools and languages for wrapper induction tools. A new approach is proposed in (Zhai and Liu, 2005), for extracting the structured data from webpages. This method is used for the automatic accomplishment of the task. It includes two steps:

- Identification of individual data records in a page
- Aligning and extracting the data items from the identified data records

Gottlob and Koch, (2004) summarized the distinguishingtechniques for logic-based web information extraction from the parse trees of web documents. Wang and Lochovsky (2003) discussed about the Data extraction and Label assigning systems for sending queries through HTML forms. The data objects in the dynamic webpages share a mutual HTML tag structure and are listed continuously on the webpages. The user query form in the webpage provides a relational database of the website. Gupta *et al.* (2003) discussed about the document object model tree, to achieve identification, content extraction and maintenance of the original data instead of summarizing it. Following, the paper Hammouda and Kamel (2004) presented two main parts of favorably document clustering process. The first part discusses about the novel phrase-based document index model. Incremental creation of the phrase-based index of the document set is permitted by the document index graph. The second part describes the incremental document clustering algorithm. The tightness of the clusters is increased by watching the pairwise similarity allocation inside the clusters.

The comparison of the major web data extraction approaches, based on the automation degree, utilized technique and task is given in (Chang *et al.*, 2006). These parameters describes:

- Failure of the information extraction system to handle some websites of particular structures
- Classification of the information extraction systems.
- degree of computerization for information extraction systems

Fu *et al.* (2010) described a method for guiding the content extraction process, using the webpage layout information. They proposed a four aspects on some specific websites, attribute to as web bases, that afford a complex HTML search form for the users to querying the back-end databases Wang (2003). Zhang *et al.* (2011) described an improved DOM-based techniques for web information extraction.

Layout Based Detachment Approach

On the internet,the user can access extreme amount of information in the form of HTML pages. Similar presentation style is used for maximum webpages, by majority of the webistes.

The non-content blocks divide some presentation style and ordinary content and some of the main content is dissimilar in their presentation and content style. Initially the effective page is allowed through the HTML parser, for investigating a webpage for the content extraction.The HTML parser adjusts the HTML and creates a Document Object Model (DOM) tree representation of the webpage. In our proposed work, content extractor navigates the DOM tree recursively, by applying a sequence of filtering methods to delete and modify detailed nodes along with the content. The advertisement remover uses anefficient technique to remove advertisements. Since the DOM tree is parsed, then the values of "src" and "href" attributes are surveyed to conclude the servers. The node of the DOM tree contained in the link is removed, if an address matches againts a list of accepted advertisement servers. In our work, we proposed a dynamic approach to extract the main content of the webpage using the tag tree parsing, block acquiring page segmentation method, data extraction. It can eliminate noise and extract the main content blocks from webpage effectively and display the essential content.

Webpages

The process following the webpage selection is for content extraction. In our conceit, we proposed layout based detachment approach to automatically the extract content from the webpage. Then, the content related to the user query is used to remove the title of the webpage, pop up ads, flashy advertisements, menus, unnecessary images and links that are not relevant for a user querying the system for educational purposes. For that purpose we have to select the webpage dynamically.

Structure Analysis

In structure investigation, the tags present in a webpage are parsed. Through the structural analysis, the tags accessible in page blocks, child tags and in addition tags over inner block are recognized. Moreover, tags are in HTML while the entire tags are analyzed and structure has been distinct. For e.g., in <Head> tag what are all tags like <Script>, <Title> etc., are parsed to find the structure of tag present in input.

Tag Tree Parsing

In tag tree parsing, a DOM tree is constructed for obtaining analysis structure and further processing of the content extraction. The XML file considered as an input of DOM tree is to be converted. The HTML file is converted into XML file. In our proposal, the hierarchical DOM tree is represented based on the XML file. While the tree is constructed it is easier to examine the structure of the webpage. The HTML parser creates independent tag trees for every webpage linked to the website. Subsequently, a procedure to integrate all the tag trees into a single tree having common features of all the webpages is accompished. This combined tree helps in the analysis of the webpage content and structure. Then, parse the structured file for creating tree and find the structure for LBPS.

Block Acquiring Page Segmentation

After obtaining the structure of the webpage, we need to remove unwanted tags. The expression unwanted tags indicate that the tag may not be closed and that may not be embrace child nodes. In our proposal, we implement block acquiring page segmentation for content extraction. It uses to clarify the main block including the inner blocks and it also defines the parsing. BAPS validate that parsing technique has one on the basis of the algorithm. It includes some essential rules.

Rule1: If a gathering node does not consist of child node then remove it.

Rule 2: If a gathering node has only 1 child node will have no gathering child node, divide this node to the archive. Set the degree value of the node to c.

Rule 3: If a gathering node has over 1 child node while have no gathering child node, divide this node to the archive. Set the degree value of the node to b.

Rule 4: If a gathering node has gathering child means then don't divide the node to achieve. Set the degree value of the node to a. If this node has child visible while not gathering node, the child node will be divided.

Algorithm: Block Acquiring Page Segmentation Method (BAPS)

```
Segmentation (node)
While (node!=null)
    if (node is acquiring node and no child)
        remove node
    else if (node is acquiring node and one child)
        if (node is not gathering node)
            node to archive
            set ranging as $b$
    else if (node is acquiring node and more child)
        if (node is gathering node)
            node to archive
            set ranging as $a$
    else
        remove node
```

In our proposal, the information after the removal of unnecessary tags in the page is removed. At present, we are going to extract the content present in the most important tag called as HTML, HEAD and BODY.

Data Extraction

Data Extraction retrieves the data out of various data sources for further processing of information. Extracting data from the unstructured sources has become a considerable challenge. In our proposed technique, the unwanted contents of the webpages are eliminated based on the segmentation algorithm. In our proposal, the most confront of wrappers must be able to locate the data of interest among other uninteresting pieces of webpages, such as advertisement regions, navigation bars and inline code. In the DOM tree representations of those webpages, each record is composed of a set of sibling subtrees i.e., consecutive or interrupted by some noisy nodes. Each sibling sub-tree is an attribute-value pair of the record. After data extraction process we need to eliminate the boundary for each and every block to obtain the required information. Figure 1 shows the overall flow of the proposed method.

Performance Analysis

In order to test our proposed work, several executions on the synthetic sequence of tasks are organized. We proposed the methodology with the model of layout based detachment approach. It differentiates noisy blocks and main content blocks.The content extraction suite can produce a wide variety of outputs, based upon the type and complexity of the webpage. The algorithm performs well on pages with large blocks of text such as news articles and mid-size of long informational passages. Most navigational bars and irrelevant elements of webpages such as advertisements and side panels are removed or reduced in size.

Figure 2 and 3 depicts an example of a typical page about the university grants commission. These are the examples of a website presented in the content rich format.

Figure 4 and 5 represents best results of the proposed were more relevant than the other works. A news page about the Bank of America comparable increases many links in the major content, that can create sufficient noise. The block of optional interpretation of the main content, however the block is contained in the div tag, moreover, it simply has the link text, that the noise block is trimmed. By the similar time, consequently numerous links in the text, reduces the weight of the text, although the punctuation density successfully supplements the weight worth of tag windows. It reveals that the proposed method consumes best webpages than the other method.

When printed out in the text format, most of the resulting text is directly related to the content of the website, while making it possible to utilize the summarization and keyword extraction algorithms efficiently and accurately. Pages with little or no textual content are extracted with varying results.

Table 1 and Fig. 6 depicts the recall values for proposed LBDA and existing heuristic approach. It represents that the result of the proposed approach was more relevant than the existing work. Accuracy depends on the accurate recovery of the webpages. Recall is estimated based on the following Equation 1:

$$Recall = \frac{No.of\ relevant\ terms\ retrieved}{No.of\ relevant\ terms\ in\ sampled\ documents} \quad (1)$$

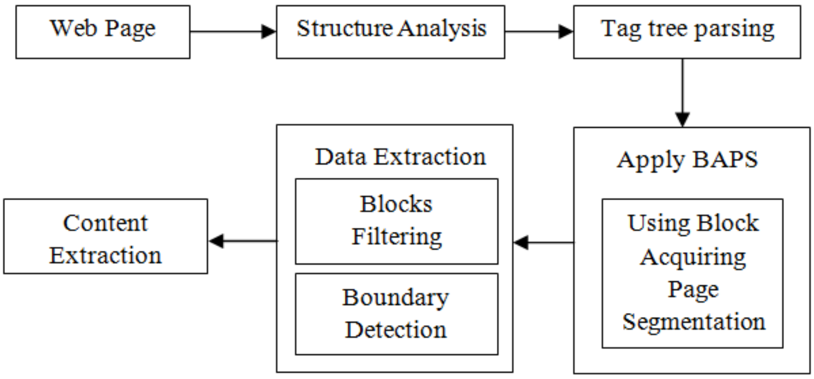
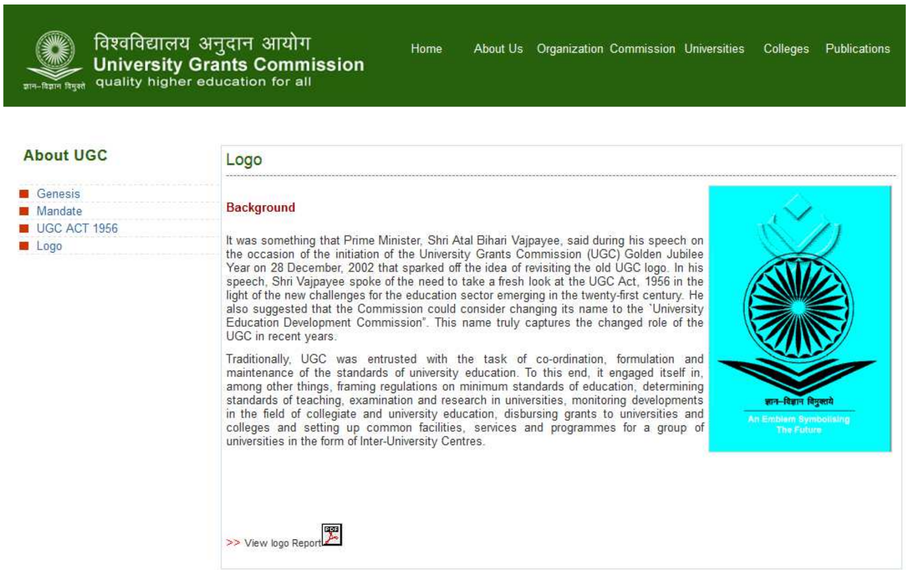


Fig. 1 Architecture of the proposed methodology



Logo

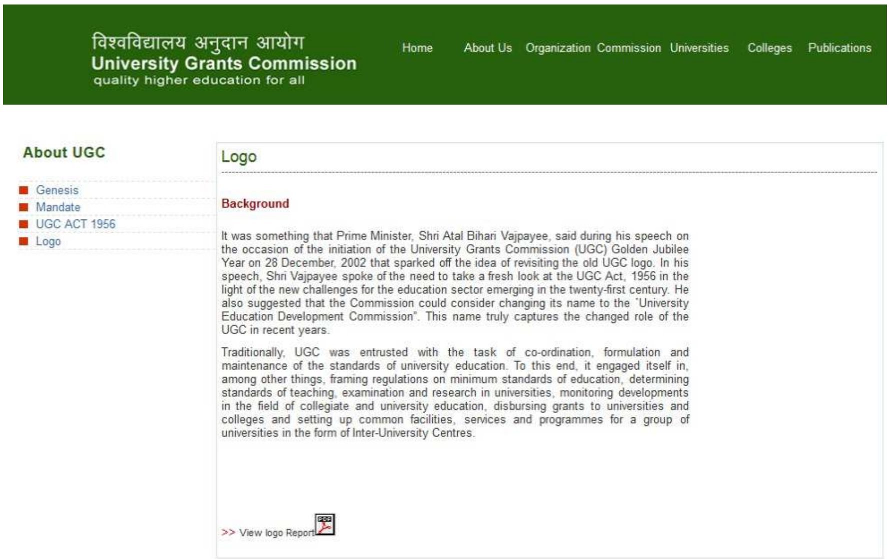
Background

It was something that Prime Minister, Shri Atal Bihari Vajpayee, said during his speech on the occasion of the initiation of the University Grants Commission (UGC) Golden Jubilee Year on 28 December, 2002 that sparked off the idea of revisiting the old UGC logo. In his speech, Shri Vajpayee spoke of the need to take a fresh look at the UGC Act, 1956 in the light of the new challenges for the education sector emerging in the twenty-first century. He also suggested that the Commission could consider changing its name to the "University Education Development Commission". This name truly captures the changed role of the UGC in recent years.

Traditionally, UGC was entrusted with the task of co-ordination, formulation and maintenance of the standards of university education. To this end, it engaged itself in, among other things, framing regulations on minimum standards of education, determining standards of teaching, examination and research in universities, monitoring developments in the field of collegiate and university education, disbursing grants to universities and colleges and setting up common facilities, services and programmes for a group of universities in the form of Inter-University Centres.

>> View logo Report

Fig. 2. Webpage content before extraction



Logo

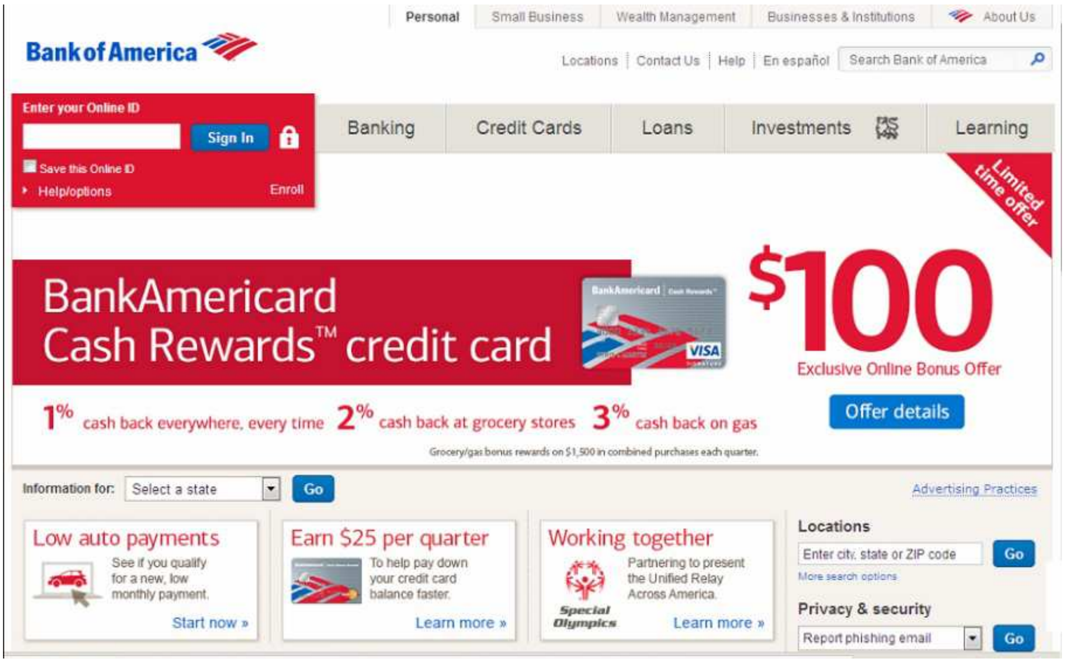
Background

It was something that Prime Minister, Shri Atal Bihari Vajpayee, said during his speech on the occasion of the initiation of the University Grants Commission (UGC) Golden Jubilee Year on 28 December, 2002 that sparked off the idea of revisiting the old UGC logo. In his speech, Shri Vajpayee spoke of the need to take a fresh look at the UGC Act, 1956 in the light of the new challenges for the education sector emerging in the twenty-first century. He also suggested that the Commission could consider changing its name to the "University Education Development Commission". This name truly captures the changed role of the UGC in recent years.

Traditionally, UGC was entrusted with the task of co-ordination, formulation and maintenance of the standards of university education. To this end, it engaged itself in, among other things, framing regulations on minimum standards of education, determining standards of teaching, examination and research in universities, monitoring developments in the field of collegiate and university education, disbursing grants to universities and colleges and setting up common facilities, services and programmes for a group of universities in the form of Inter-University Centres.

>> View logo Report

Fig. 3. Webpage content after extraction



BankAmericard Cash Rewards™ credit card

$100 Exclusive Online Bonus Offer

1% cash back everywhere, every time 2% cash back at grocery stores 3% cash back on gas

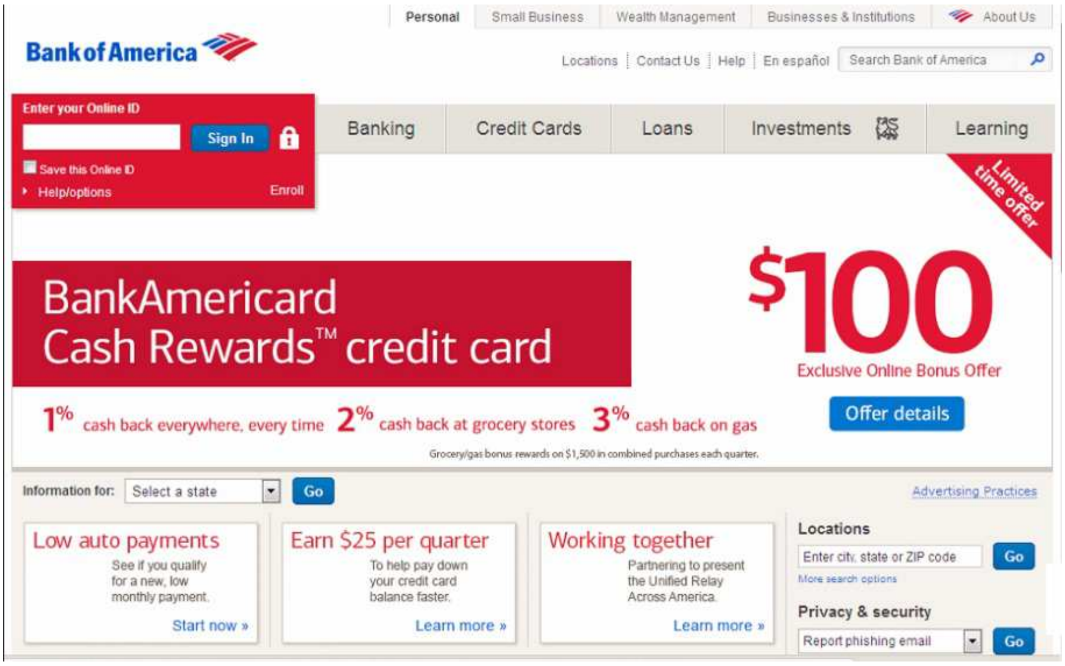
Grocery/gas bonus rewards on $1,500 in combined purchases each quarter.

Low auto payments
See if you qualify for a new, low monthly payment.

Earn $25 per quarter
To help pay down your credit card balance faster.

Working together
Partnering to present the Unified Relay Across America.

Fig. 4. Webpage before content extraction



BankAmericard Cash Rewards™ credit card

$100 Exclusive Online Bonus Offer

1% cash back everywhere, every time 2% cash back at grocery stores 3% cash back on gas

Grocery/gas bonus rewards on $1,500 in combined purchases each quarter.

Low auto payments
See if you qualify for a new, low monthly payment.

Earn $25 per quarter
To help pay down your credit card balance faster.

Working together
Partnering to present the Unified Relay Across America.

Fig.5 Webpage after content extraction

Table 1. Recall analysis

Dataset	Heuristic approach	LBDA (proposed)
1	73	82
2	75	86
3	79	89
4	83	93
5	84	96

Table 2 and Fig. 7 depicts the precision rate of our proposed work. The precision of the algorithm is based on a baseline, that is a pre-evaluated parameter so it is a normalized index. Furthermore, these indexes are compared among the traditional algorithm along with our own algorithm.

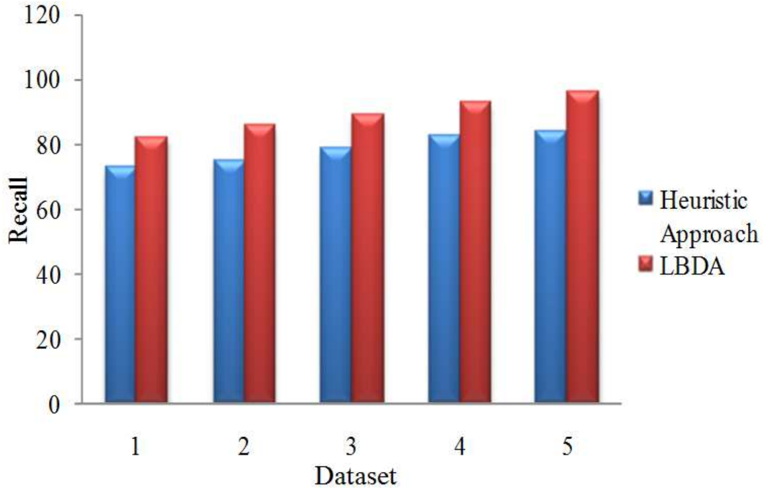


Fig. 6. Recall study of LBDA (proposed) and heuristic approach

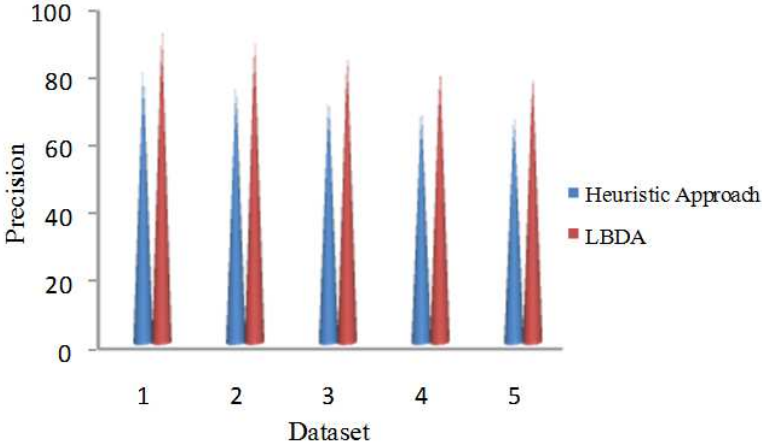


Fig. 7. Precision comparison of LBDA (proposed) and heuristic approach

Table 2. Precision analysis

Dataset	Heuristic approach	LBDA (proposed)
1	80	92
2	76	89
3	72	85
4	69	81
5	67	79

These comparisons reveal the benefits of our algorithm. It displays two page segmentation algorithm. While comparing with Heuristic approach and LBDA, LBDA gain acceptable results. These two algorithms are info based page segmentation algorithm. The precision of Heuristic approach in our experiments is 55%, however LBDA takes precedence over it and the result is 60%. This LBDA can gain the most acceptable results. In order to predict the proposed method's efficiency we calculate its overall prediction accuracy. The precision can be calculated based on the following Equation 2:

$$Precision = \frac{No.of\ relevant\ terms\ retrieved}{No.of\ terms\ retrieved\ from\ sampled\ documents} \quad (2)$$

Table 3 and Fig. 8 depicts the processing accuracy comparatively with existing and proposed work, the LBDA results better accuracy value than the existing approaches.

Execution time of the approach plays a fundamental role in the estimation of the efficacy of the webpage. Table 4 and Fig. 9 characterize the time required for both LBDA (proposed) and existing heuristic technique for webpage information. It reveals that the proposed method consumes less time than the existing heuristic method. This explicitly denotes that proposed method is faster than the existing method.

Memory usage determines the required memory to execute the techniques. Table 5 and Fig. 10 express the utilization of main memory during execution of both proposed and existing techniques. It represents that LBDA requires lesser memory than the existing heuristic approach.

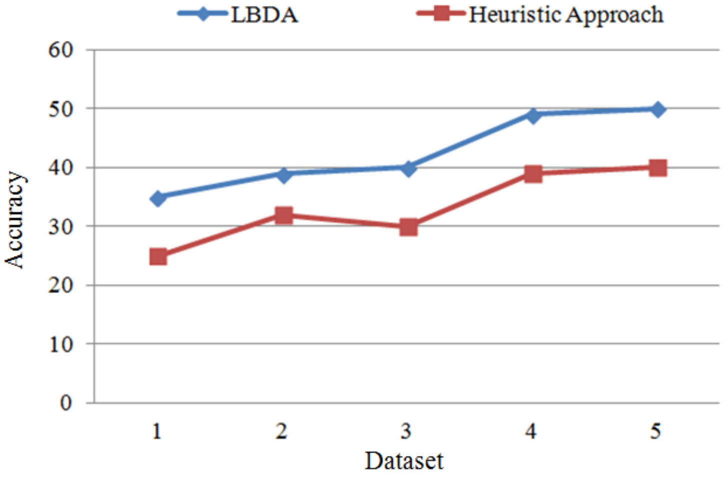


Fig. 8. Accuracy for LBDA (proposed) and heuristic approach

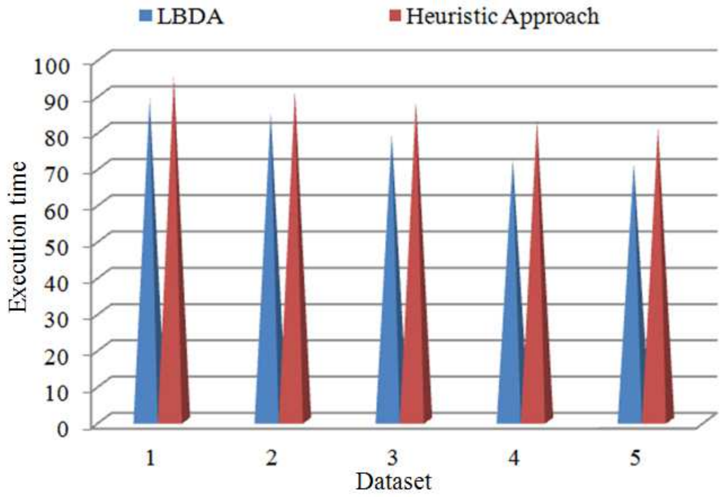


Fig. 9. Execution time for LBDA (proposed) and heuristic approach

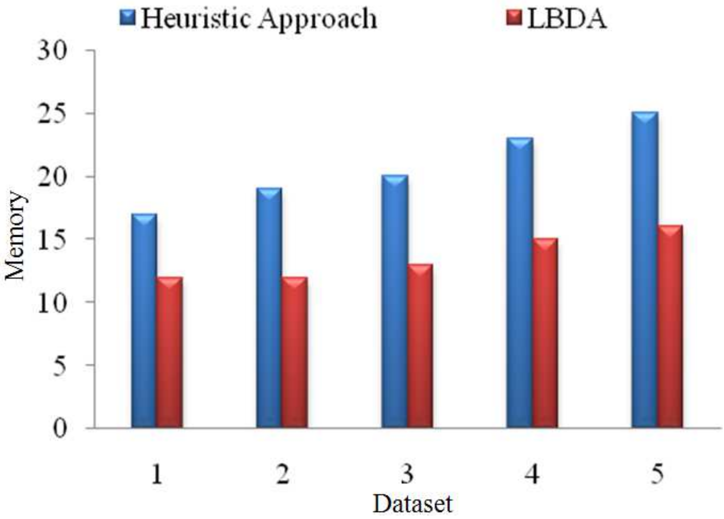


Fig. 10. Memory usage for LBDA (proposed) and heuristic approach

Table 3. Accuracy analysis

Dataset	Heuristic approach	LBDA (proposed)
1	25	35
2	32	39
3	30	40
4	39	49
5	40	50

Table 4. Execution time analysis

Dataset	Heuristic approach	LBDA (proposed)
1	95	89
2	91	85
3	88	79
4	83	72
5	81	71

Table 5. Memory usage analysis

Dataset	Heuristic approach	LBDA (proposed)
1	17	12
2	19	12
3	20	13
4	23	15
5	25	16

Conclusion and Future Work

In this study, we proposed a novel methodology for a layout based detachment approach to extract the content from the webpages. Webpage content extractions are more vital to retrieve the contents of the webpages, particularly in unstructured web. The proposed technique uses the DOM tree parsing and segmentation algorithm to remove the noise and irrelevant information. Experimental results show that the webpage produces better and effective results after implementation of LBDA. Many web applications such as information retrieval, automatic page adaptation and information extraction can benefit from this method. The proposed approach is implemented on the Java platform and it produces better recall, precision and accurate results when compared with the existing heuristic approach. Also, it utilizes less execution time and less memory usage than the heuristic approach.

Our future plan is to investigate the applicability of our suggested procedure for improving the data extraction accuracy and adding the noise filter for blocking the advertisement. Moreover, wrapper generation method is for improving specialized programs, which identify data of interest and map them to some relational form.

Acknowledgement

The authors would like to express their gratitude to Managing Trustee, SNR TRUST, Coimbatore, Tamilnadu, India for permitting to do this research work.

Funding Information

The authors have no funding information to report.

Author’s Contributions

The LBDA work was designed, directed and coordinated by Dr. Anna Saro Vijendran and Ms. Deepa Chandran. Ms. Deepa Chandran proposed and conducted the LBDA approach, prepared the manuscript and performed copy editing with the guidance of Dr. Anna Saro Vijendran.

Ethics

This article is original and it contains the unpublished material. The corresponding author confirms that the other authors have read and approved the manuscript. No other ethical issues are involved.

References

Arasu, A. and H. Garcia-Molina, 2003. Extracting structured data from webpages. Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 09-12, San Diego, CA, USA, pp: 337-348. DOI: 10.1145/872757.872799

Cai, D., S. Yu, J.R. Wen and W.Y. Ma, 2003. Extracting content structure for webpages based on visual representation. Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications, (WAT’ 03), Springer, pp: 406-417.

Chang, C.H., M. Kayed, R. Girgis and K.F. Shaalan, 2006. A survey of web information extraction systems. IEEE Trans. Know. Data Eng., 18: 1411-1428. DOI: 10.1109/TKDE.2006.152

Embley, D.W., Y. Jiang and Y.K. Ng, 1999. Record-boundary discovery in Web documents. ACM SIGMOD Record, 28: 467-478. DOI: 10.1145/304181.304223

Fu, L., Y. Meng, Y. Xia and H. Yu, 2010. Web content extraction based on webpage layout analysis. Proceedings of the 2nd International Conference on Information Technology and Computer Science, Jul. 24-25, IEEE Xplore Press, Kiev, pp: 40-43. DOI: 10.1109/ITCS.2010.16

Gottlob, G. and C. Koch, 2004. Logic-based web information extraction. ACM SIGMOD Record, 33: 87-94. DOI: 10.1145/1024694.1024711

Gupta, S., G. Kaiser, D. Neistadt and P. Grimm, 2003. DOM-based content extraction of HTML documents. Proceedings of the 12th International Conference on World Wide Web, (WWW’ 03), ACM New York, NY, USA, pp: 207-214. DOI: 10.1145/775152.775182

Hammer, J., H. Garcia-Molina, J. Cho, R. Aranha and A. Crespo, 1997. Extracting semistructured information from the web. Proceedings of the Workshop on Management of Semistructured Data, (MSD' 97), Tucson, Arizona, pp: 18-25.

Hammouda, K.M. and M.S. Kamel, 2004. Efficient phrase-based document indexing for web document clustering. IEEE Trans. Know. Data Eng., 16: 1279-1296. DOI: 10.1109/TKDE.2004.58

Laender, A.H., B.A. Ribeiro-Neto, A.S. da Silva and J.S. Teixeira, 2002. A brief survey of web data extraction tools. ACM Sigmod Record, 31: 84-93. DOI: 10.1145/565117.565137

Lei, F., M. Yao and Y. Hao, 2009. Improve the performance of the webpage content extraction using webpage segmentation algorithm. Proceedings of the International Forum on Computer Science-Technology and Applications, Dec. 25-27, IEEE Xplore Press, Chongqing, pp: 323-325. DOI: 10.1109/IFCSTA.2009.84

Liu, W., X. Meng and W. Meng, 2006. Vision-based web data records extraction. Proceedings of the 9th International Workshop on the Web and Databases, (WWD' 06), Chicago, Illinois, pp: 20-25

Sharma, A., 2010. Accessing the deep web using ontology. Proceedings of the 3rd International Conference on Emerging Trends in Engineering and Technology, Nov. 19-21, IEEE Xplore Press, Goa, pp: 565-568. DOI: 10.1109/ICETET.2010.35

Wang, J. and F.H. Lochovsky, 2003. Data extraction and label assignment for web databases. Proceedings of the 12th International Conference on World Wide Web, (WWW' 03), ACM New York, NY, USA, pp: 187-196. DOI: 10.1145/775152.775179

Wang, J., 2003. Information discovery, extraction and integration for the hidden web. University of Science and Technology Clear Water Bay.

Xian, X., P. Zhao, W. Fang, J. Xin and Z. Cui, 2009. Data source selection for large-scale deep web data integration. Proceedings of the ND Pacific-Asia Conference on Web Mining and Web-based Application, Jun. 6-7, IEEE Xplore Press, Wuhan, pp: 178-182. DOI: 10.1109/WMWA.2009.25

Zhai, Y. and B. Liu, 2005. Web data extraction based on partial tree alignment. Proceedings of the 14th International Conference on World Wide Web, (WWW' 05), ACM New York, NY, USA, pp: 76-85. DOI: 10.1145/1060745.1060761

Zhang, L., M. Li, N. Dong and Y. Wang, 2011. An improved DOM-based algorithm for web information extraction. J. Inform. Comput. Sci., 8: 1113-1121.