

## Chief Editor

Dr. A. Singaraj, M.A., M.Phil., Ph.D.

## Editor

Mrs.M.Josephin Immaculate Ruba

## EDITORIAL ADVISORS

1. Prof. Dr.Said I.Shalaby, MD,Ph.D.  
Professor & Vice President  
Tropical Medicine,  
Hepatology & Gastroenterology, NRC,  
Academy of Scientific Research and Technology,  
Cairo, Egypt.
2. Dr. Mussie T. Tessema,  
Associate Professor,  
Department of Business Administration,  
Winona State University, MN,  
United States of America,
3. Dr. Mengsteab Tesfayohannes,  
Associate Professor,  
Department of Management,  
Sigmund Weis School of Business,  
Susquehanna University,  
Selinsgrove, PENN,  
United States of America,
4. Dr. Ahmed Sebihi  
Associate Professor  
Islamic Culture and Social Sciences (ICSS),  
Department of General Education (DGE),  
Gulf Medical University (GMU),  
UAE.
5. Dr. Anne Maduka,  
Assistant Professor,  
Department of Economics,  
Anambra State University,  
Igbariam Campus,  
Nigeria.
6. Dr. D.K. Awasthi, M.Sc., Ph.D.  
Associate Professor  
Department of Chemistry,  
Sri J.N.P.G. College,  
Charbagh, Lucknow,  
Uttar Pradesh. India
7. Dr. Tirtharaj Bhoi, M.A, Ph.D,  
Assistant Professor,  
School of Social Science,  
University of Jammu,  
Jammu, Jammu & Kashmir, India.
8. Dr. Pradeep Kumar Choudhury,  
Assistant Professor,  
Institute for Studies in Industrial Development,  
An ICSSR Research Institute,  
New Delhi- 110070, India.
9. Dr. Gyanendra Awasthi, M.Sc., Ph.D., NET  
Associate Professor & HOD  
Department of Biochemistry,  
Dolphin (PG) Institute of Biomedical & Natural  
Sciences,  
Dehradun, Uttarakhand, India.
10. Dr. C. Satapathy,  
Director,  
Amity Humanity Foundation,  
Amity Business School, Bhubaneswar,  
Orissa, India.



ISSN (Online): 2455-7838

SJIF Impact Factor (2016): 4.144

EPRA International Journal of

# Research & Development (IJRD)

Monthly Peer Reviewed & Indexed  
International Online Journal

Volume:2, Issue:3, March 2017



Published By :  
EPRA Journals

CC License





## WEB SKELTONIZER SERVICE

**Ms. S. Selvakanmani<sup>1</sup>**

<sup>1</sup> Assistant Professor, Department of CSE, Velammal Institute of Technology,  
Panchetti, Thiruvallur District, T.N, India

**Mr. D. Robin Reni<sup>2</sup>**

<sup>2</sup> Second Year Student, Department of CSE, Velammal Institute of Technology,  
Panchetti, Thiruvallur District, T.N, India

**Mr. S. S. Yogeshwaraa<sup>3</sup>**

<sup>3</sup> Second Year Student, Department of CSE, Velammal Institute of Technology,  
Panchetti, Thiruvallur District, T.N, India

### ABSTRACT

*This paper deals with Web Skeletonizer service software that can clone any simple website which are static in their nature. Through this simple application, we are able to get their pages along with its entire background coding and edit as per our choice to present in front of the users. This application will not only save time and development effort but also saves you from additional investment in designing section. This system will also able to show the data consumption in MB while using internet.*

**INDEX TERMS:** *Web Skeletonizer, Cloning, Web Pages, Information Crawling, URL, Images*

### I. INTRODUCTION

Web pages are typically designed to facilitate visual interaction with the human readers. Web page designers normally organize the web page information into different units or functional types, which are arranged in coherent visual segments in the page, such as header, footer, navigation menu, major content, etc[1]. For example, on the news site these elements are used to differentiate various kinds of news items. However, these visual segments are not explicitly declared in the source code. What we can find from the source code are the list items, paragraphs, instead of clear visual segmentation or pattern. But automatically identifying different segments from web pages can be very useful for different fields. One application is information crawling, web skeletonizer algorithm automatically analyzes the

structure of web pages and estimates different regions of web page and then only important parts are saved during the web crawling. What's more, information extraction, information retrieval, and Web page classification can obtain better performance based on the web segmentation structure.[1]

Web skeletonizer service software is responsible for cloning any simple website which are static in their nature. Sometimes you need to get some special appearance and design for your new website and you do not able to get free templates of such websites. Through this simple application, you will able to get their pages along with its entire background coding and edit as per your choice to present in front of the users. This system will also able to show the website history like Data of Modification, content – length etc.

Websites are the module for world wide communication. Its not easy to build a website at once. We need lots of talent, experience and tools to implement a website. In our project ,we introduce an methodology of cloning an complete website and customize to the user need. [2]

Its smart categorization method will able to keep all the captured materials by making its default type folders in the particular location. Before starting this application, you will have the set the memory area where all these cloned materials is to be saved. During cloning process, it will make differentiate between CSS, HTML, JQUERY, Images and icons and saved in particular folder, so that it will be easy for you to use as per your development work. Users will only have to place the website URL either by copying in the text field or by writing themselves and press on start button to begin cloning.

## II. EXISTING SYSTEM

Cloning of website and all its contents is not an easy task. Even categorization of downloaded materials were not possible by the previous system by which users comes with many difficulties about knowing the types of file and the existing system was only able to download from html file. It was not able to download the CSS and images which are available on the website, thus getting only the text part and no clue provided for the coding section through which users can edit the code and use for their own website. When running the simple html file, it was not able to produce output as available on the server due to lack of other attached files or links with particular html file.

There are different approaches to segment web pages described in a survey[1], such as DOM-Based approach[5][6], Visual- Based approach[7][8], Text-Based approach[3] and Hybrid approach[4][5]. And web page segmentation has been proposed to address problems in different fields including mobile web, duplicate detection, information retrieval, web page clustering, etc.

In a DOM-based approach, one simply looks at the DOM tree for cues on how to segment a page. It is very intuitive since one only needs to parse the HTML rather than rendering the page. However, Web pages are getting more complex than ever, there are many different ways to build a HTML document structure for the same content, style and layout information. Cai et al.[6] propose VIPS algorithm to extract semantic content structure of a web page by utilizing heuristic rules based on the DOM representation as well as visual features. It detects web content structure following the automatic top-down and tag-tree independent principles and obtains a better segmentation of a web page at semantic level.

The text based approach retrieves segments of web pages based on the properties of text such as paragraph similarity, clustering, among others. Kohlsch et al.[5] utilize the notion of text-density as a measure to identify the individual text segments of a web page, reducing the problem to a 1D partitioning task. The distribution of segment-level text density seems to follow a negative hypergeometric distribution. The presented Block Fusion algorithm identifies segments using the text density metric as a valuable heuristic.

The hybrid approach guarantees some semantic coherence of blocks. It takes into account the spatial location, visual properties of the page (e.g. by using CSS) and relationships among blocks. BoM[8] is a good example that uses the geometric aspects of page and categorization of DOM elements to perform the segmentation. The segmentation is processed to build three structures: content, geometric and logic structure. Each one represents a different perspective of the segmentation and the logic structure is the final result.

## III. MOTIVATION

The major problem is we cannot clone the entire website, so that many of the main content of the website will be missed out . The motto of our project is to the user should reproduce all the content of the required website without any hassles, which is cost efficient. Also there should not be any data loss. ficient

## IV. PROPOSED SYSTEM

This new web skeletonizer application will able to identify the downloaded pages and files, and if found escape in order to copy new files. It will automatically able to create different folders as per the type and saved such documents and files in that particular folder which will help in easy searching process. If you will enter the same URL and its documents have been saved previously, an alert message will be provided to the user to begin the same task. Some websites requires authentication, so to overcome from this, user can enter their id and password before cloning while using this application.

## V. MODULES

- ▶ Target the Website Based on its URL
- ▶ Load its Content
- ▶ Separate their Body Component
- ▶ Save module
- ▶ Analyze Module

## VI. OUTPUT



Fig. 1 Target the website

Fig.2 Analyze Module

## VII. CONCLUSION

Web Skeletonizer Service will be an great tool to clone an website and get the complete content of it without any hassles. This new web skeletonizer application will able to identify the downloaded pages and files, and if found escape in order to copy new files. Users will only have to place the website URL either by copying in the text field or by writing themselves and press on start button to begin cloning. Its smart categorization method will able to keep all the captured materials by making its default type folders in the particular location

## REFERENCES

1. N. C. Campus, "Web page segmentation: A review," 2011.
2. Hanyang Feng, Wenzhe Zhan, "Web Page Segmentation and Its Application for Web Information Crawling" (Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference).
3. D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Vips: a vision-based page segmentation algorithm," Microsoft technical report, MSR-TR-2003-79, Tech. Rep., 2003.
4. A. Sanoja and S. Gancarski, "Block-o-matic: A web page segmentation framework," in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014, pp. 595–600.
5. C. Kohlschütter and W. Nejdl, "A densitometric approach to web page segmentation," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 1173–1182.
6. D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Extracting content structure for web pages based on visual representation," in *Asia-Pacific Web Conference*. Springer, 2003, pp. 406–417.
7. X. Yin and W. S. Lee, "Understanding the function of web elements for mobile content delivery using random walk models," in *Special interest tracks and posters of the 14th International Conference on World Wide Web*. ACM, 2005, pp. 1150–1151.
8. M. Sarkis, C. Concolato, and J.-C. Dufourd, "Msos: A multi-screen-oriented web page segmentation approach," in *Proceedings of the 2015 ACM Symposium on Document Engineering*. ACM, 2015, pp. 85–88.