E-ISSN : 2988-585X (Online)

**Journal of Elektronik Sistem InformasI (JESII)**

*Volume 2 No 1 June 2024*

*DOI : 10.31848/jesii.v2i1.3414*

# PARSING DOM Method Implementation for Web Scraping to Identify Pornographic Websites

**Muhammad Insan Al-Amin**

Department of Engineering Informatics, Monash University Australia

**Article Info**

***Article history:***

Received May 30, 24
Revised Jun 23, 24
Accepted Jun 25, 24

***Keywords:***

Parsing Dom Method
Parental Control
Pornographic
Web Scraping

**ABSTRACT**

This work explores the implementation of the DOM Parsing Method for web scraping in the context of identifying pornographic websites. The approach leverages the Document Object Model (DOM) structure of websites to extract key features that might indicate pornographic content. However, the abstract should also acknowledge limitations such as potential ethical concerns around scraping and the need for accuracy beyond just DOM parsing techniques. When utilized improperly, internet technology can be a source of illegal information. One such instance is when pornographic content is published online. One of the numerous solutions that have been developed for DNS, or what is also referred to as DNS blocking, is that it is still subject to modification and re-access. The issue can be resolved by utilizing web scraping techniques to identify keywords based on text content, extract information from the website, and retrieve information using the DOM Parsing method. This allows text content to be gathered as a source of information related to the site's content, and the TF-IDF method can be used to calculate each website's weight known as the pornographic website. This study develops a reliable technique for locating pornographic content on the internet, which may be used with parental controls or content filtering systems.

***Corresponding Author:***
Muhammad Insan Al-Amin
Monash University, Australia
Wellington Rd, Clayton VIC 3800, Australia
Email: muhammad.insanalamin@uinsgd.ac.id

## 1. INTRODUCTION

The widespread availability of internet access has opened doors to a vast amount of information. However, this accessibility also presents challenges, particularly regarding inappropriate content such as pornography and graphic violence [1]. The process of obtaining data from a website is called web scraping, or web data retrieval [2]. To view a web page and simulate a web browser, a computer software is used. A database or spreadsheet are two examples of usable formats in which the extracted data can be kept [3][4]. Exposure to such content can have negative consequences, especially for children and adolescents [5]. Although the Internet connects computer networks, it should primarily be viewed as a source of information. Information, which can be thought of as a database or a vast and comprehensive multimedia library, makes up the content of the Internet [6][7]. Because virtually every facet of real-world life—including commerce, entertainment, sports, politics, and so forth—can be found online, the Internet is even referred to as the world in a different form, or maya. Content filtering and parental

control mechanisms play a crucial role in mitigating these issues [8][9]. These systems aim to identify and restrict access to harmful websites[10]. Existing solutions often rely on website blacklists, which require constant maintenance and may not be entirely effective. There is an amazing amount of knowledge available in the virtual world. But occasionally, we require a lot of data that is dispersed among several web pages [11][10]. Manually gathering this data is undoubtedly time-consuming and exhausting. This is when the method of web scraping comes in handy. An automated method of obtaining data from a website is called web scraping[1][3]. A web scraper uses the page's source code, which is often HTML, to extract the needed data, much like how your web browser displays a webpage[12][7].

Cyberporn is one type of hacking crime, therefore one way to combat pornographic websites is by technological prevention[13][14]. Regarding the filtering technologies that have been developed, Nawala's DNS Filtering is one of them [15]. This is the most straightforward method of preventing access to a DNS-based website using the Nawala DNS Filter; any URL that are on Nawala's blacklist render the requested page unavailable. The system that uses DNS-based site blocking has the drawback of allowing a website that loads negatively to be re-accessed by simply altering its DNS[11]. Users have a lot of intriguing options when it comes to choosing content because to the website's growing range of content. Negative content has regularly had an impact on a variety of media platforms, including the uploading of pornographic photos, gambling, fraud, harassment, damage to one's reputation, and fake news.

This study offers a novel way for online scraping to find websites with improper content by using Parsing DOM (Document Object Model) techniques. This approach focuses on content analysis rather than just targeting websites with a reputation for hosting pornography. Compared to conventional blacklisting techniques, this research attempts to offer a more reliable and flexible way for recognizing inappropriate content. The system may be able to identify a greater variety of hazardous content by concentrating on content analysis, increasing the efficacy of content filtering systems. It is significant to highlight that specific terms associated with pornography are not used in this study. The emphasis is instead on recognizing a wider variety of unsuitable items. This method guarantees a more morally sound and comprehensive answer for content filtering. DOM Parsing could still be part of an approach to identifying porn sites, but recent research suggests other more effective and accurate techniques. A combined approach that combines DOM parsing with other techniques such as deep learning may be at the forefront in the future.

## 2. METHOD

The software engineering technique known as RUP (Rational Unified Process) was created by compiling the finest software development industry practices. With an emphasis on model building using the Unified Model Language (UML), RUP employs an object-oriented notion. RUP, or the Rational Unified Process, is an agile technique used in software development. The project life cycle is divided into four phases by RUP. All six of the fundamental development disciplines—business modeling, requirements, analysis and design, implementation, testing, and deployment—occur during each phase.

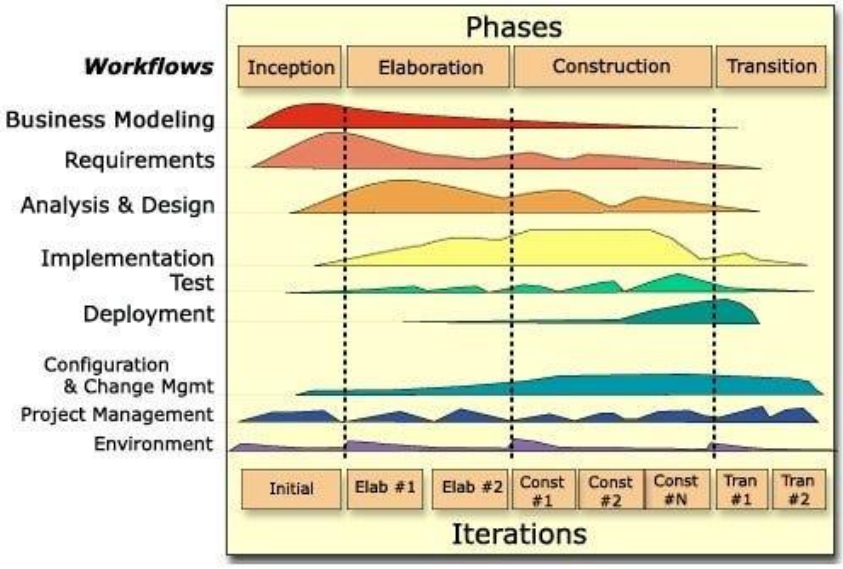


Figure 1. RUP Process Methodology

RUP (Rational Unified Process) is an iterative software development framework created by Rational Software Corporation, now the division of IBM. RUP is not a single process with rigid rules.

2.1 Analyze System

Most sources of information can be found on the Internet, and information, both positive and negative, can be easily found in the Internet. Internet users can easily search for information, so they have the freedom to choose whatever content they want. In this case, many Internet service providers have used a domain name filter system to block pornographic websites. However, this method is ineffective in preventing access to pornography content because a blocked domain name server can be replaced with a new domain name Server, so that a website that has a blocked domain name servers can be accessed only by changing the domain name of the server. Analysis of data processing from randomlytaken URL links. This data will be processed using DOM Parsing and weighed using TF-IDF.

2.2 Process Dom Parsing Method

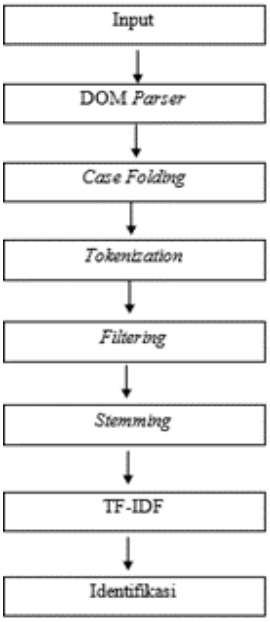


Figure 1 Data Processing

is a process to extract a website from the URL input. Data taken from the HTML tag is converted to plaintext and saved as a document. The following example:

a. Input: balabala.html
b. Simple text: sex, pornography, and Miyabi

Case folding is changing all the big letters to small letters. The following are :

Table 1 Case Folding Process

| No | Documen | Case Folding Result |
|---|---|---|
| 1 | Sex, porn and Miyabi. | sex, porn and miyabi. |

Tokenizing is the separation of sentence text from a word called a token, are :

Table 2 Tokenizing Process

| No | *Case Folding Result* | *Tokenizing Result* |
|---|---|---|
| 1 | sex, porn and miyabi. | sex<br>, |

Filtering is the removal of words and signs that have no meaning, such as special characters, i.e.:

Table 3 Filtering Process

| No | Result *Tokenizing* | Result *Filtering* |
|---|---|---|
| 1 | sex,<br>porn and<br>miyabi. | sex<br>porn<br>and<br>miyabi |

Stemming is taking the root of a word that has a suffix and a prefix and turning it into a base word, such as running, which has the base word run.

*Table 4* Stemming Process

| No | Result *Filtering* | Result *Stemming* |
|---|---|---|
| 1 | swimming<br>running | swim<br>run |

The process of matching keywords from the keyword table results in a true value when the term is found in the keyword table and a false value when it is not.

*Table 5* Matching Process

| No | DB Keyword | Hasil *Case Folding* | Result |
|---|---|---|---|
| 1 | sex porn<br>-<br>miyabi | sex porn<br>and miyabi | True True<br>False True |

The TF-IDF method explains the weight count of each document by counting the number of terms that appear on each document, then searching for the idf value using the log(n/df) formula, where n is the total number of documents tested and divided by the df value. Furthermore, the weight value, represented by the W variable, is found by multiplying the tf (term) and idf frequency values. (invers frekuensi documen).

Table 6 TF-IDF Process

| Term | tf | | | | df | idf | Wij = tf.idf | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | Log( n/df) | 1 | 2 | 3 | 4 |
| sex | 2 | | 1 | | 2 | 0,301 | 0,602 | 0 | 0,301 | 0 |
| porn | 1 | 2 | 1 | 1 | 4 | 0 | 0 | 0 | 0 | 0 |
| miyabi | 1 | | | | 1 | 0,602 | 0,602 | 0 | 0 | 0 |

## 3. RESULT AND DISCUSSION

The goal of the study is to create a DOM-based technique for locating adult content on the internet. The suggested technique looks for unsuitable content by analyzing text and images.

### *3.1. Implementation*

System implementation is the activity of creating a system or application using the help of software or hardware according to the analysis and design to produce a system that works.

Table 7 Psedecode Implementation

| |
|---|
| START Sub tfidf()<br><br>input : data clean document := document if document = null Print "Empty Document" end If document ≠ null number := count data clean key := keyword Do for i = number to i++{ casefolding := number[i] } Do for I = number of i++ { token := amount[i) } Do For i = amount to i ++{ filtering := sum[i].<br><br>} Do for i = sum to i++{ if key = sum [i] stemming := sum[i] end if }<br><br>for each data clean (tf, idf) do weight := tf x idf w := weight return w<br><br>end if END |

### 3.2. System Architecture

System architecture defines specific components structurally, in ficture :

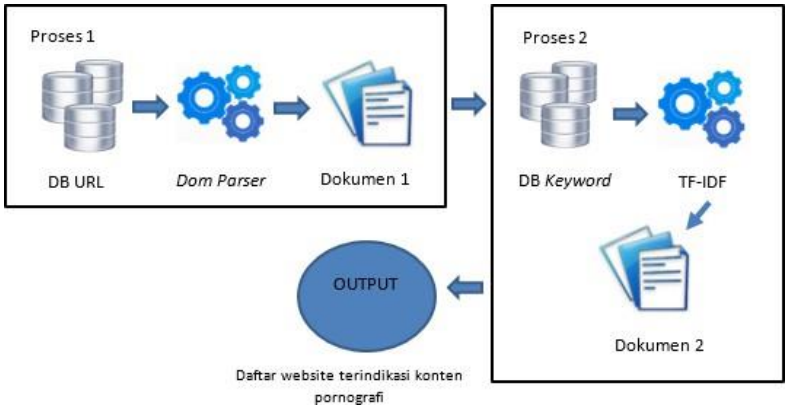


Figure 2 System Architecture

### 3.3. Application Architecture

Application architecture design defines the flow of information that takes place when an application is running, in the ficture :

Figure 3 Architecture Application

### 3.4. System Implementation

The index page is the initial page that serves to generate an object from the class tf-idf during program execution, enabling the execution of weight calculation procedures utilizing the tf - idf method, internet scraping, and text preparation.


Figure 4 System Implementation

Scraping is the method of meticulously getting a text object ready for processing, including tokenization, filtering, stemming, and case folding, so that the processing can start.
In order to lower the weight value on each document, the tf-idf page is the one that meticulously builds objects classes for the tf - idf when the page is begun.


*Figure 5* Scrapping Implementation

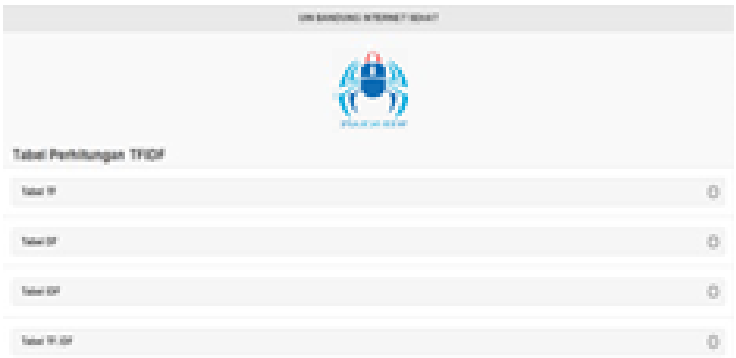

Figure 6 TF-IDF Implementation

An identification page is a page that runs weight calculations using the tf-idf method; the output is a list table containing a list of websites that have been found to have pornographic content.

*Figure 7* Identification Web Scrapping

## 4. CONCLUSION

The built-in system is capable of identifying pornographic websites without opening the website in advance so that users do not have to check by opening each website whether it contains pornography or not. The DOM Parsing-based method proposed in this study shows promising results in identifying adult content on the web. The combination of text analysis and image analysis offers a more adaptive approach than traditional filtering methods. Further research is needed to improve accuracy and ensure the ethical and legal aspects of its implementation.

## ACKNOWLEDGEMENTS

Thank you to all parties who have supported this research so that it can be completed properly. Especially, for the parents who have always supported and guided me in this research. I`m very grateful.

## REFERENCES

[1] E. Prayitno, T. Suprawoto, and ..., "Optimasi Hasil Pencarian Pada Web Scrapping Menggunakan Pembobotan Kata Tf-Idf," *J. Innov. Res. Knowl.*, vol. 1, no. 7, pp. 241–246, 2021, [Online]. Available: https://bajangjournal.com/index.php/JIRK/article/view/822

[2] N. Yona Sidratul Munti and D. Asril Syaifuddin, "Analisa Dampak Perkembangan Teknologi Informasi Dan Komunikasi Dalam Bidang Pendidikan," *J. Pendidik. Tambusai*, vol. 4, no. 2, pp. 1799–1805, 2020.

[3] A. Purnomo, "Impementasi Web Scraping Pada OJS Dengan Metode CSS Selector," *RESOLUSI Rekayasa Tek. Inform. dan Inf.* , vol. 3, no. 2, pp. 37–42, 2022, [Online]. Available: https://djournals.com/resolusi

[4] Y. Saputra, N. I. Putri, E. S. Nurpajriah, D. Jaelani, and A. Hamdani, "Perencanaan Strategis Sistem Informasi untuk Mendukung Keputusan Organisasi dengan Ward dan Peppard," vol. 6, no. 2, pp. 137–145, 2023.

[5] V. Tasril, "Penggunaan Add Ons Dalam Perlindungan Untuk Cyberporn," *Penerbit Tahta Media*, 2023, [Online]. Available: http://tahtamedia.co.id/index.php/issj/article/view/85%0Ahttp://tahtamedia.co.id/index.php/issj/article/download/85/84

[6] D. Montanesa and Y. Karneli, "Pemahaman Remaja Tentang Internet Sehat Di Era Globalisasi," *Edukatif J. Ilmu Pendidik.*, vol. 3, no. 3, pp. 1059–1066, 2021, [Online]. Available: https://edukatif.org/index.php/edukatif/article/view/509

[7] K. D. Tembo, "A sketch of two parallaxes of porn and its use: revelation and regulation," *Porn Stud.*, vol. 8, no. 4, pp. 439–463, 2021, doi: 10.1080/23268743.2021.1879667.

[8] A. D. R. Salsabilah, I. Zulfa, and M. Saputra, "Parsing Data Log Hasil Pemblokiran Situs Negatif Di Satuan Kerja Perangkat Aceh," *J. Tek. Inform. dan Elektro*, vol. 6, no. 1, pp. 21–36, 2024, doi:

10.55542/jurtie.v6i1.965.
[9] A. Morichetta, M. Trevisan, L. Vassio, and J. Krickl, "Understanding web pornography usage from traffic analysis," *Comput. Networks*, vol. 189, no. July 2020, p. 107909, 2021, doi: 10.1016/j.comnet.2021.107909.
[10] Muhammad Hasim S and A. Muh Alryan Pangeran Syamsibah, "Web Proxy Pada Jaringan LAN di Lab Jaringan JTIK," *J. Renew. Energy Smart Device*, vol. 1, no. 1, pp. 37–39, 2023, doi: 10.61220/joresd.v1i1.237.
[11] D. I. Mulyana, F. Ardiyansyah, N. Hidayat, and A. Zulfikar, "Optimasi Keamanan Jaringan Wifi dari Situs Judi Online dan Pornografi dengan DNS Filtering dan Orangepi," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 647–655, 2024, doi: 10.57152/malcom.v4i2.1274.
[12] H. S. Wirawan, "Perancangan Keamanan Akses Internet Berbasis Text Filtering Pada Universitas Atma Jaya Makassar," *Temat. J. Penelit. Tek. Inform. dan Sist. Inf.*, pp. 71–82, 2022, doi: 10.56963/tematika.v9i2.131.
[13] K. Harahap, "Penerapan Algoritma Exact String Matching Dalam Pembuatan Program Blokir Situs Porno," *J. Pelita Inform.*, vol. 8, no. 1, pp. 47–50, 2019, [Online]. Available: www.worldsex.com
[14] A. C. DWINATA, "Identifikasi Konten Negatif Pada Twitter Dengan Deep Learning." 2023. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/42556%0Ahttps://dspace.uii.ac.id/bitstream/handle/123456789/42556/17523030.pdf?sequence=1&isAllowed=y
[15] A. Hunaepi, M. Makhsun, and S. Sarwani, "Deteksi Situs Pornografi Berdasarkan Gambar Menggunakan," *J. Tek. Inform.*, vol. 12, no. 2, 2019.