GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SEARCH ENGINE OPTIMIZATION

A Project Report of Capstone Project - 2

Submitted by:

AYUSH MISHRA

(1613101212/ 16SCSE101386)

in partial fulfillment for the award of the degree

of

Bachelor of Technology

IN

Computer Science and Engineering

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

DR.D GANESH GOPAL

Professor

MAY-2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report “ **SEARCH ENGINE OPTIMIZATION”** is the bonafide work of “**AYUSH MISHRA(1613101212)**” who carried out the project work under my supervision.

SIGNATURE OF HEAD
Dr.MUNISH SHABARWAL,
PhD(Management),PhD(CS)
Professor & Dean,
School of Computing Science &
Engineering

SIGNATURE OF SUPERVISOR
Dr.D GANESH GOPAL
Ph.D
Professor
School of Computing Science &
Engineering

ABSTRACT

Search engine optimization influence the presence of a website in the first page of a search engine. The visibility of a website can be paid or unpaid, but generally if a website wants to appear in the top position after a live search then search engine optimization is the most important strategic tool to use. The complete internet marketing strategy circles around the SEO. The search optimization processes try to follow the working pattern of all the search engines and more specifically try to consider the algorithm used in search engines. Online presence of an organisation is not only an easy way to reach among the target users but it may be profitable too if optimization is done keeping in view of the target users as of the reason that most of the time users search out with the keywords of their use rather than searching the organisation name, and if the page link comes in the top positions then the page may be profitable.

Table of contents

LIST OF TABLE

LIST OF FIGURES

1.INTRODUCTION

1.1 Overview & Motivation:

Search engine optimization (SEO) is the process of affecting the visibility of a website or a web page in a web search engine's unpaid result. It is the process of getting traffic from the “free,” “organic,” “editorial” or “natural” search results on search engines.

Search engine optimization is a strategic technique to take a web document in top search results of a search engine. Online presence of an organisation is not only an easy way to reach among the target users but it may be profitable too if optimization is done keeping in view of the target users as of the reason that most of the time users search out with the keywords of their use (Say; PhD in web technology) rather than searching the organisation name, and if the page link comes in the top positions then the page turns out to be profitable. This work describes the tweaks of taking the page on top position in Google by increasing the Page rank which may result in the improved visibility and profitable deal for an organisation. Google is most user-friendly search engine proved for the Indian users, which gives user-oriented results. In addition, most of other search engines use Google search patterns; so, we have concentrated on it. So, if a page is optimised in Google it is optimised for most of the search engines.

To understand what SEO really means, let's break that sentence down and look at the parts:

- Quality of traffic. You can attract all the visitors in the world, but if they're coming to your site because Google tells them you're a resource for Apple computers when really you're a farmer selling apples, that is not quality traffic. Instead you want to attract visitors who are genuinely interested in products that you offer.

- Quantity of traffic. Once you have the right people clicking through from those search engine results pages (SERPs), more traffic is better.
- Organic results. Ads make up a significant portion of many SERPs. Organic traffic is any traffic that you don't have to pay for.
- The O part of SEO—optimization—is where the people who write all that content and put it on their sites are gussying that content and those sites up so search engines will be able to understand what they're seeing, and the users who arrive via search will like what they see.
- Optimization can take many forms. It's everything from making sure the title tags and meta descriptions are both informative and the right length to pointing internal links at pages you're proud of.

1.2 Objective

The purpose of the project is to optimize the website using various search engine optimization techniques and get it listed among the top pages of Google and other search engines.

SEO, or search engine optimization, is the **process** by which people take to make sure that the websites are appealing to the search engines. Even though techniques can vary depending on your industry, the main idea is to make sure the basic white hat tactics are followed on a regular basis.

The goals of the project are:

1. To get the website listed in the top pages of the search engine.

2. To build an interactive, user friendly website.

3. To make the website rich in graphics to enhance its target audience.

4. To decrease the competition of the organization behind the website.

5. Understanding both the abilities and limitations of search engines.

1.3 Scope

A search engine optimization company plays a huge role in a site's marketing endeavours. Search engine optimization entails hiking up the chances of the user logging in the particular sites when they search with the help of the related keywords on any of the search engines.

The project will include making of a user-friendly and interactive website based on experiences and reviews' sharing and providing the users to get an idea about the same in a very unique and interesting way. This process would also include buying of a domain name as well as web hosting for it. This will be preceded by an exhaustive keyword research on our part. The website would include original content by all the reviewers and this will be ensured by the site admin. Once the website goes live various SEO techniques (off-page) would be applied to constantly improve and/or maintain the ranking of the website in the SERP. Apart from this we will be using Google web analytics to constantly monitor the site speed, site content, search engine traffic, bounce rate, clicks per page, etc.

2.LITERATURE SURVEY

2.1 Technical Feasibility

Technical Feasibility assesses the current resources and technology, which are required to accomplish the goals within the allocated time and for this, the team ascertains whether the current resources and technology can be upgraded or added to accomplish specified requirements.

Technical feasibility of the product has been studied under following heads:

- **(i)** **Hardware availability -** This project requires any device with internet connection. (Here we will be using laptops for testing purpose).
- **(ii)** **Platform Independence –** This project will be platform independent since its website can be accessed with the help of internet on any platform.
- **(iii)** **User friendly -** As the website’s interface will be built on WordPress which has rich user interface tools and for its SEO, Google analytics which is a free service provided by the Google will be used, this system will definitely be user friendly. As hardware availability and user-friendly nature of this project is possible to achieve, this system is technically feasible.

2.2 Economic Feasibility This project requires a domain name and web hosting for constructing the website and further deploying it.

(i) Hardware cost – There is no hardware cost associated with this. However, a domain name and web hosting is needed to host the website. So it will result into some minimal charges. (ii) Software cost - Software needed for development of this website are open source software i.e. wordpress. Hence, there is no cost associated with them. As hardware cost are very low and as software are freely available, this system is economic feasible.

2.3 Operational Feasibility

Operational feasibility is a measure of how well a proposed system solves the problems. We know without SEO where a website will lie and after SEO how much its usability will increase Thus the system is operational feasible.

2.3.1 Initial Phase Operational Techniques

Techniques	Importance
Page Title	Page Title is the first thing that Google search for. The title tag tells search engines, what the page is all about. So, it is important to always use the keyword which suits your webpage and which is in demand of users. The title tag will always appear in the web browser tag.
Meta Description	Meta Description Tag provides the short description of the page. This description may appear in the **Google** search engine result page. This description is more helpful for the user to understand about your webpage or your website. It is extremely important part of search marketing.
Meta Title	Meta Keywords are comma separated words that describe the contents of a website. This may or may not appear the search engine result page. This is less important as compare to Meta description.
Heading Tag	Heading Tags serve to divide the page into sections. Heading tags provide structure to your webpage and the structure webpage is easy to follow and can be rank better as compare to unstructured pages.
Image Alt Attribute	There are number of users that may not view the images on their browser, like- users with slow internet connectivity, users that disabled the images on their web browser, etc. So, HTML provides us the facility of image alt to describe the image in readable text mode. [11] Search Engine use the alt attribute to determine the best image to return for a query and improve the ranking of the website.

2.3.2 Promotional Phase Techniques

Techniques	Importance
Site Map	Sitemap offer the opportunity to inform search engines immediately about any changes on your website. The changes in a website will be indexed faster. Sitemap also help in classifying your website content.
Join Groups	When you join groups, you have more chance to advertise your website. You can update your website link in the group, by which more number of users can search for your website and may increase the ranking also.
Social Networking Sites	Social Networking sites are also an effective way to improve the ranking of a website. Social Networking sites like facebook, Twitter are more famous now days. You can advertise your website there also. Advertising through these social networking sites are increasing day by day.
Link Building	Link building is a process of building or creating the link in order to improve the ranking of your website. You can build your link free or by paid some amount also. Like-**Google** Adwords, where you need to pay some amount to advertise your website.
Blogging	Blogging may be defined as discussion forum or information site. **Google** also providing the blog facility. You can create a blog and upload the link of your website for any queries to the user. They will contact to your website and may help to improve the searching of your website.

SYSTEM REQUIREMENTS ANALYSIS

2.4 Platform specification

This project will be platform independent since its website can be accessed with the help of internet on any platform. But to deliver high levels of reliability, availability and serviceability the configuration needed is as follows:

(i) Hardware Specification: Any device with internet connectivity is required.

(ii) Software Specification: The above-mentioned device must contain a well updated web browser to run the website and implement the SEO techniques and software.

2.5 Functional requirements

Project shall have a website on travel blog which appears among the top searches of a search engine whenever a person searches something related to this stuff.

1. Website shall have only original content from the authenticated users and reviewers.

2. The website once deployed shall maintain its ranking in the SERP with the help of SEO techniques.

3. Website shall have its content as per the Google algorithms so that web crawlers can display this website on the necessary searches.

2.6 Business end requirements 1. Website should recover its ranking in case of any failure. 2. Website shall have an option of including new techniques so as to maintain its ranking.

2.7 Non-functional requirements

This section specifies the required system quality factors that are not related to the specific functional requirements. These requirements are always required to be fulfilled.

(i)Safety and Security: The admin will keep a check on the original content being posted by the reviewers and also has to ensure that this does not get copied. There will a method to conform to the copywriting issues of the content.

(ii)Performance: This subsection checks the fact that our interface must perform in a way user expects. It is also associated with the speed with which the system shall function.

(iii)Reliability: The target audience must rely on the matter displayed on the site as the content won't be plagiarized and will the reviewer will rely on the website so as not to disclose their identity on any other platform without the it's consent.

(iv)Reusability: Since the content posted on the website is original in nature; each time anybody visits the site would get the idea about the particular experience. Hence the content never gets outdated. Also in case of any changes to the fact mentioned in the post the admin will make sure to have as little a difference in the description and reality as possible.

TOOLS REQUIRED :

Hardware and Software Requirements-

Software Requirements :

- **Operating System :** Windows XP Professional
- **Front End :** Java (J2EE)

Hardware Requirements:

- **HardDisk :** 40Gb
- **RAM :** 512Mb
- **Processor :** Pentium IV
- **Monitor :** 17”Color Monitor

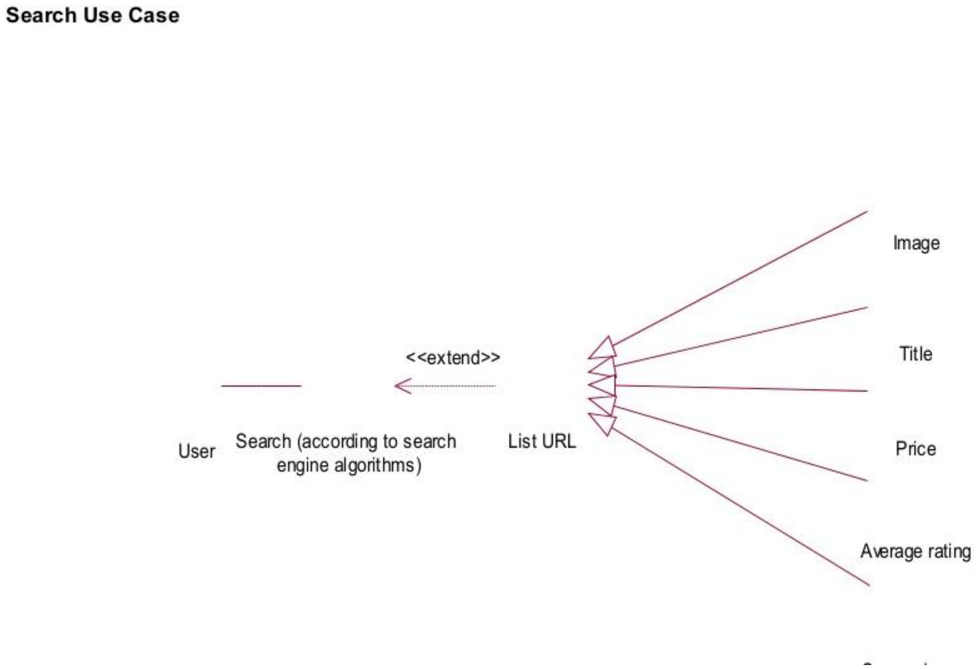


Fig.1 URL generation Use-Case

User Use Case

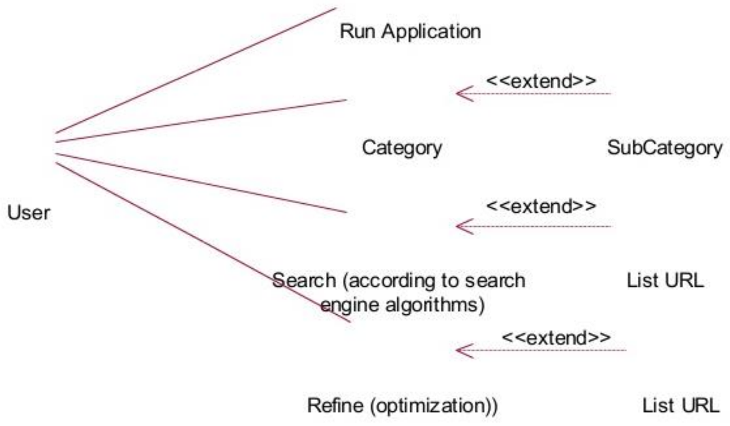


Fig.2 User side Use-Case

Class Diagrams

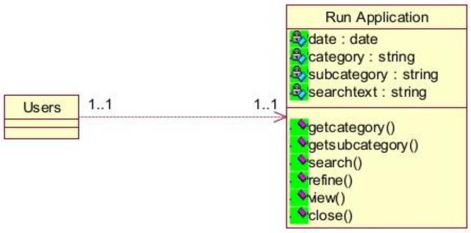


Fig.3 User side Class Diagram

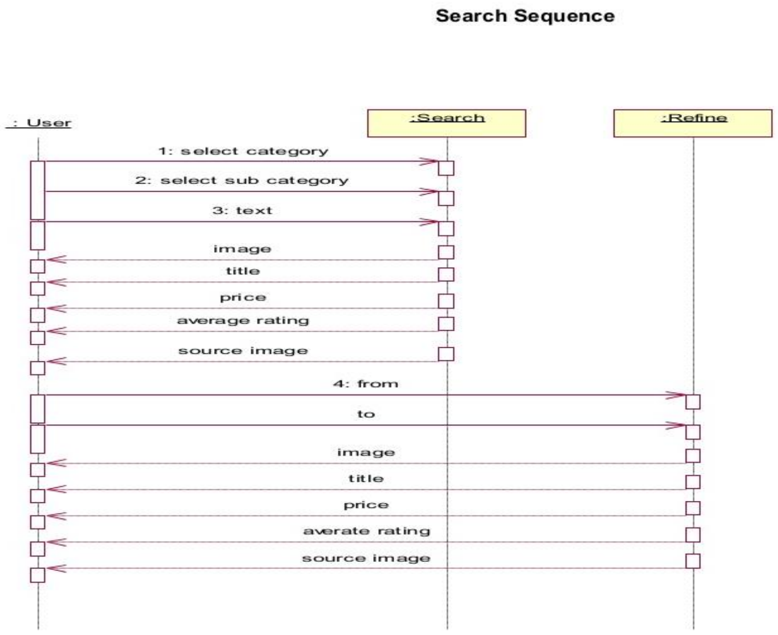


Fig.4 Structural view

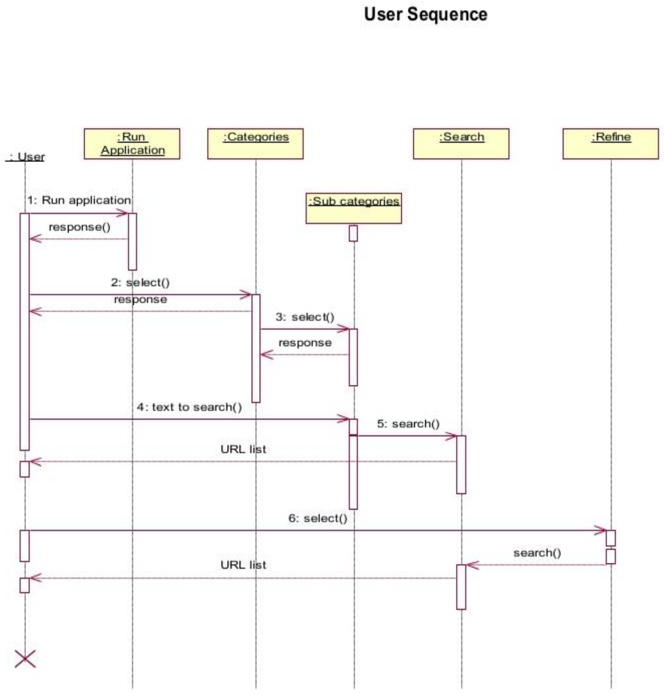


Fig.5 Run Application Structure

IMPLEMENTATION

Index.php

```
<html><head>
<title>Sitemap Generator</title>
<style>
input[type=text] {
    width:40%;
          margin:5px 10px auto auto;
          height:45px;
    box-sizing: border-box;
    border: 2px solid #ccc;
    border-radius: 4px;
    font-size: 16px;
    background-color: white;
    background-image: url('searchicon.png');
    background-position: 10px 10px;
    background-repeat: no-repeat;
    padding: 12px 20px 12px 40px;
    -webkit-transition: width 0.4s ease-in-out;
    transition: width 0.4s ease-in-out;

}

input[type=submit] {
  width: 10%;
  background-color: #4CAF50;
  color: white;
  padding: 14px 20px;
  margin: 8px 0;
  border: none;
  border-radius: 4px;
  cursor: pointer;
}

input[type=submit]:hover {
  background-color: #45a049;

}

.div1
{
margin:10% auto 20% 20% ;
}



</style>
</head>
<body>
<?php
include'menu.php';
?>
```

```
<div class="div1">
<form action="sitemap.php" method="post">
   <label for="fname">Site Url:</label>
   <input type="text" id="site" name="site" placeholder="Enter website url here..">
<input type="submit" value="Submit">
  </form>
</div>
</body>
</html>
```

Menu.php

```
<head>
   <!-- Chrome, Firefox OS and Opera -->
<meta name="theme-color" content="rgb(201, 21, 2)"> <!-- Windows Phone -->
<meta name="msapplication-navbutton-color" content="rgb(201, 21, 2);">
<!-- iOS Safari -->
<meta name="apple-mobile-web-app-status-bar-style" content="rgb(201, 21, 2);">

   <meta name="viewport" content="width=device-width, initial-scale=1" />
```

```
<link rel="icon" href="logo1.ico" type="image/ico" sizes="16x16" />
<style>


input[type=text]:focus {
   width: 45%;
}

@media all and (max-width : 768px) {

input[type=text] {
   width: 100%;
        margin:6px 0px auto auto;
        height:45px;
   box-sizing: border-box;
   border: 2px solid #ccc;
   border-radius: 10px;
   font-size: 16px;
   background-color: white;
   background-image: url('searchicon.png');
   background-position: 10px 10px;
   background-repeat: no-repeat;
   padding: 12px 20px 12px 40px;
   -webkit-transition: width 0.4s ease-in-out;
   transition: width 0.4s ease-in-out;
}
input[type=text]:focus {
   width: 100%;
}

}
.bar1, .bar2, .bar3 {
   width: 35px;
   height: 4px;
   background-color:white;
   margin: 6px 0;
   transition: 0.4s;
}

.change .bar1 {
   -webkit-transform: rotate(-45deg) translate(-9px, 6px);
   transform: rotate(-45deg) translate(-9px, 6px);
}
.change .bar2 {opacity: 0;}
.change .bar3 {
   -webkit-transform: rotate(45deg) translate(-8px, -8px);
   transform: rotate(45deg) translate(-8px, -8px);
}
.logo1
{
left:100px;
width:140px;
height:50px;
top:30px;
```

```
position:relative;
}
@media all and (max-width : 768px) {
    .logo1
    {
 display:none;
    }
}

</style>
<script>
function myFunction(x) {
    x.classList.toggle("change");
}
</script>
<link rel="stylesheet" href="style/homepage.css" />
</head>
<div style=' z-index:+1;'>

<div  style='padding-bottom:60px;'>
 <nav>

</label>
        <input type="checkbox" id="drop" />
            <ul class="menu">
                <li><a href="/index.php">Home</a></li>
                <li><a href="sitemapgen.php">Sitemap</a></li>
                <li><a href="/contact.php">Contact</a></li>
                <li><a href="/about.php">About</a></li>
                                        <li> </li>
            </ul>
</nav>
</div>

</div>
```

Sitemap.config

```
<?php
/*
Sitemap Generator by Slava Knyazev. Further acknowledgements in the README.md file.

Website: https://www.knyz.org/
I also live on GitHub: https://github.com/knyzorg
Contact me: Slava@KNYZ.org
*/

//Make sure to use the latest revision by downloading from github: https://github.com/knyzorg/Sitemap-Generator-Crawler

/* Usage
Usage is pretty strait forward:
- Configure the crawler by editing this file.
- Select the file to which the sitemap will be saved
- Select URL to crawl
- Configure blacklists, accepts the use of wildcards (example: http://example.com/private/* and *.jpg)
- Generate sitemap
- Either send a GET request to this script or run it from the command line (refer to README file)
- Submit to Google
- Setup a CRON Job execute this script every so often

It is recommended you don't remove the above for future reference.
*/

// Default site to crawl
$site = "https://www.extramovies.online";

// Default sitemap filename
$file = "sitemap.xml";
$permissions = 0644;

// Depth of the crawl, 0 is unlimited
$max_depth = 0;

// Show changefreq
$enable_frequency = false;

// Show priority
```

```
$enable_priority = false;

// Default values for changefreq and priority
$freq = "daily";
$priority = "1";

// Add lastmod based on server response. Unreliable and disabled by default.
$enable_modified = false;

// Disable this for misconfigured, but tolerable SSL server.
$curl_validate_certificate = true;

// The pages will be excluded from crawl and sitemap.
// Use for exluding non-html files to increase performance and save bandwidth.
$blacklist = array(
    "*.jpg",
    "*/secrets/*",
    "https://www.knyz.org/supersecret"
);

// Enable this if your site do requires GET arguments to function
$ignore_arguments = false;

// Not yet implemented. See issue #19 for more information.
$index_img = false;

//Index PDFs
```

```
$index_pdf = true;

// Set the user agent for crawler
$crawler_user_agent = "Mozilla/5.0 (compatible; Sitemap Generator Crawler;
+https://github.com/knyzorg/Sitemap-Generator-Crawler)";

// Header of the sitemap.xml
$xmlheader ='<?xml version="1.0" encoding="UTF-8"?>
<urlset
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">';

// Optionally configure debug options
$debug = array(
    "add" => true,
    "reject" => false,
    "warn" => false
);



//Modify only if configuration version is broken
$version_config = 2;
```

Sitemap.function.php

```
<?php

// Abstracted function to output formatted logging
function logger($message, $type)
{
    global $debug, $color;
    if ($color) {
        switch ($type) {
            case 0:
                //add
                echo $debug["add"] ? "\033[0;32m [+] $message \033[0m\n" : "";
                break;
            case 1:
                //reject
                echo $debug["reject"] ? "\033[0;31m [-] $message \033[0m\n" : "";
                break;
            case 2:
                //manipulate
                echo $debug["warn"] ? "\033[1;33m [!] $message \033[0m\n" : "";
                break;
            case 3:
                //critical
                echo "\033[1;33m [!] $message \033[0m\n";
                break;
        }
        return;
    }
    switch ($type) {
        case 0:
            //add
            echo $debug["add"] ? "[+] $message\n" : "";
            break;
        case 1:
            //reject
            echo $debug["reject"] ? "31m [-] $message\n" : "";
            break;
        case 2:
            //manipulate
            echo $debug["warn"] ? "[!] $message\n" : "";
            break;
        case 3:
            //critical
            echo "[!] $message\n";
            break;
    }
}

function flatten_url($url)
```

```
{
    global $real_site;
    $path = explode($real_site, $url)[1];
    return $real_site . remove_dot_seg($path);
}

/**
 * Remove dot segments from a URI path according to RFC3986 Section 5.2.4
 *
 * @param $path
 * @return string
 * @link http://www.ietf.org/rfc/rfc3986.txt
 */
function remove_dot_seg($path)
{
    if (strpos($path, '.') === false) {
        return $path;
    }

    $inputBuffer = $path;
    $outputStack = [];

    /**
     * 2.  While the input buffer is not empty, loop as follows:
     */
    while ($inputBuffer != '') {
        /**
 * A.  If the input buffer begins with a prefix of "../" or "./",
 *     then remove that prefix from the input buffer; otherwise,
 */
if (strpos($inputBuffer, "./") === 0) {
    $inputBuffer = substr($inputBuffer, 2);
    continue;
}
if (strpos($inputBuffer, "../") === 0) {
    $inputBuffer = substr($inputBuffer, 3);
    continue;
}

/**
 * B.  if the input buffer begins with a prefix of "/./" or "/.",
 *     where "." is a complete path segment, then replace that
 *     prefix with "/" in the input buffer; otherwise,
 */
if ($inputBuffer === "/.") {
    $outputStack[] = '/';
    break;
}
if (substr($inputBuffer, 0, 3) === "/./") {
    $inputBuffer = substr($inputBuffer, 2);
    continue;
}

/**
```

```
 * C.  if the input buffer begins with a prefix of "/../" or "/..",
 *     where ".." is a complete path segment, then replace that
 *     prefix with "/" in the input buffer and remove the last
 *     segment and its preceding "/" (if any) from the output
 *     buffer; otherwise,
 */
if ($inputBuffer === "/..") {
    array_pop($outputStack);
    $outputStack[] = '/';
    break;
}
if (substr($inputBuffer, 0, 4) === "/../") {
    array_pop($outputStack);
    $inputBuffer = substr($inputBuffer, 3);
    continue;
}

/**
 * D.  if the input buffer consists only of "." or "..", then remove
 *     that from the input buffer; otherwise,
 */
if ($inputBuffer === '.' || $inputBuffer === '..') {
    break;
}

/**
 * E.  move the first path segment in the input buffer to the end of
```

```
         *     the output buffer, including the initial "/" character (if
         *     any) and any subsequent characters up to, but not including,
         *     the next "/" character or the end of the input buffer.
         */
        if (($slashPos = stripos($inputBuffer, '/', 1)) === false) {
            $outputStack[] = $inputBuffer;
            break;
        } else {
            $outputStack[] = substr($inputBuffer, 0, $slashPos);
            $inputBuffer = substr($inputBuffer, $slashPos);
        }
    }

    return ltrim(implode($outputStack), "/");
}

// Check if a URL has already been scanned
function is_scanned($url)
{
    global $scanned;

    if (isset($scanned[$url])) {
        return true;
    }

    //Check if in array as dir and non-dir
    $url = ends_with($url, "/") ? substr($url, 0, -1) : $url . "/";
    if (isset($scanned[$url])) {
        return true;
    }

    return false;
}

function ends_with($haystack, $needle)
{
    $length = strlen($needle);
    if ($length == 0) {
        return true;
    }
    return (substr($haystack, -$length) === $needle);
}

// Gets path for a relative linl
// https://somewebsite.com/directory/file => https://somewebsite.com/directory/
//                              https://somewebsite.com/directory/subdir/                              =>
https://somewebsite.com/directory/subdir/
function get_path($path)
{
    $path_depth = explode("/", $path);
    $len = strlen($path_depth[count($path_depth) - 1]);
    return (substr($path, 0, strlen($path) - $len));
}
```

```
//Get the root of the domain
function domain_root($href)
{
    $url_parts = explode('/', $href);
    return $url_parts[0] . '//' . $url_parts[2] . '/';
}

//The curl client is create outside of the function to avoid re-creating it for
performance reasons
$curl_client = curl_init();
function get_data($url)
{
    global $curl_validate_certificate, $curl_client, $index_pdf, $crawler_user_agent,
$enable_modified;

    //Set URL
    curl_setopt($curl_client, CURLOPT_URL, $url);
    //Follow redirects and get new url
    curl_setopt($curl_client, CURLOPT_RETURNTRANSFER, 1);
    //Get headers
    curl_setopt($curl_client, CURLOPT_HEADER, 1);
    //Optionally avoid validating SSL
    curl_setopt($curl_client, CURLOPT_SSL_VERIFYPEER,
$curl_validate_certificate);
    //Set user agent
    curl_setopt($curl_client, CURLOPT_USERAGENT, $crawler_user_agent);
```

```
//Get data
$data = curl_exec($curl_client);
$content_type = curl_getinfo($curl_client, CURLINFO_CONTENT_TYPE);
$http_code = curl_getinfo($curl_client, CURLINFO_HTTP_CODE);
$redirect_url = curl_getinfo($curl_client, CURLINFO_REDIRECT_URL);

//Scan new url, if redirect
if ($redirect_url) {
    logger("URL is a redirect.", 1);
    if (strpos($redirect_url, '?') !== false) {
        $redirect_url = explode($redirect_url, "?")[0];
    }
    unset($url, $data);

    if (!check_blacklist($redirect_url)) {
        echo logger("Redirected URL is in blacklist", 1);

    } else {
        scan_url($redirect_url);
    }
}

//If content acceptable, return it. If not, `false`
$html = ($http_code != 200 || (!stripos($content_type, "html"))) ? false : $data;

//Additional data
if ($enable_modified){
        curl_setopt($curl_client, CURLOPT_FILETIME, true);
        $timestamp = curl_getinfo($curl_client, CURLINFO_FILETIME);
        $modified = ($timestamp != -1) ? date('c', $timestamp) : null;
    }
    else $modified = null;

    if (stripos($content_type, "application/pdf") !== false && $index_pdf) {
        $html = "This is a PDF";
    }
    //Return it as an array
    return array($html, $modified, (stripos($content_type, "image/") &&
$index_img));
}

//Try to match string against blacklist
function check_blacklist($string)
{
    global $blacklist;
    if (is_array($blacklist)) {
        foreach ($blacklist as $illegal) {
            if (fnmatch($illegal, $string)) {
                return false;
            }
        }
    }
    return true;
}
```

```
//Extract array of URLs from html document inside of `href`s
function get_links($html, $parent_url, $regexp)
{
    if (preg_match_all("/$regexp/siU", $html, $matches)) {
        if ($matches[2]) {
            $found = array_map(function ($href) use (&$parent_url) {
                global $real_site, $ignore_arguments;

                logger("Checking $href", 2);

                if (strpos($href, "#") !== false) {
                    logger("Dropping pound.", 2);
                    $href = preg_replace('/\#.*/', '', $href);
                }

                //Seperate $href from $query_string
                $query_string = '';
                if (strpos($href, '?') !== false) {
                    list($href, $query_string) = explode('?', $href);

                    //Parse &amp to not break curl client. See issue #23
                    $query_string = str_replace('&amp;', '&', $query_string);
                }
                if ($ignore_arguments) {
                    $query_string = '';
                }
                if (strpos($href, '?') !== false) {
                    echo "EFEASDEFSED";
                }

                if ((substr($href, 0, 7) != "http://") && (substr($href, 0, 8) != "https://"))
{
                    // Link does not call (potentially) external page
                    if (strpos($href, ":")) {
                        logger("URL is an invalid protocol", 1);
                        return false;
                    }
                    if ($href == '/') {
                        logger("$href is domain root", 2);
                        $href = $real_site;
                    } elseif (substr($href, 0, 1) == '/') {
                        logger("$href is relative to root, convert to absolute", 2);
                        $href = domain_root($real_site) . substr($href, 1);
                    } else {
                        logger("$href is relative, convert to absolute", 2);
                        $href = get_path($parent_url) . $href;
                    }
                }
                logger("Result: $href", 2);
                if (!filter_var($href, FILTER_VALIDATE_URL)) {
                    logger("URL is not valid. Rejecting.", 1);
                    return false;
                }
```

```
                if (substr($href, 0, strlen($real_site)) != $real_site) {
                    logger("URL is not part of the target domain. Rejecting.", 1);
                    return false;
                }
                if (is_scanned($href . ($query_string ? '?' . $query_string : "))) {
                    //logger("URL has already been scanned. Rejecting.", 1);
                    return false;
                }
                if (!check_blacklist($href)) {
                    logger("URL is blacklisted. Rejecting.", 1);
                    return false;
                }
                return flatten_url($href . ($query_string ? '?' . $query_string : "));
            }, $matches[2]);
            return $found;
        }
    }
    logger("Found nothing", 2);
    return array();
}

function scan_url($url)
{
    global $scanned, $deferredLinks, $file_stream, $freq, $priority,
$enable_priority, $enable_frequency, $max_depth, $depth, $real_site, $indexed;
    $depth++;
```

```
logger("Scanning $url", 2);
if (is_scanned($url)) {
    logger("URL has already been scanned. Rejecting.", 1);
    return $depth--;
}
if (substr($url, 0, strlen($real_site)) != $real_site) {
    logger("URL is not part of the target domain. Rejecting.", 1);
    return $depth--;
}
if (!($depth <= $max_depth || $max_depth == 0)) {
    logger("Maximum depth exceeded. Rejecting.", 1);
    return $depth--;
}

//Note that URL has been scanned
$scanned[$url] = 1;

//Send cURL request
list($html, $modified, $is_image) = get_data($url);

if ($is_image) {
    //Url is an image
}

if (!$html) {
    logger("Invalid Document. Rejecting.", 1);
    return $depth--;
}

if (strpos($url, "&") && strpos($url, ";") === false) {
    $url = str_replace("&", "&amp;", $url);
}

$map_row = "<url>\n";
$map_row .= "<loc>$url</loc>\n";
if ($enable_frequency) {
    $map_row .= "<changefreq>$freq</changefreq>\n";
}
if ($enable_priority) {
    $map_row .= "<priority>$priority</priority>\n";
}
if ($modified) {
    $map_row .= "   <lastmod>$modified</lastmod>\n";
}
$map_row .= "</url>\n";
fwrite($file_stream, $map_row);
$indexed++;
logger("Added: " . $url . (($modified) ? " [Modified: " . $modified . "]" : "), 0);
unset($is_image, $map_row);

// Extract urls from <a href="??"></a>
$ahrefs = get_links($html, $url, "<a\s[^>]*href=(\"|'??)([^\" >]*?)\\1[^>]*>(.*)<\/a>");
// Extract urls from <frame src="??">
$framesrc = get_links($html, $url, "<frame\s[^>]*src=(\"|'??)([^\"
>]*?)\\1[^>]*>");

$links = array_filter(array_merge($ahrefs, $framesrc), function ($item) use
(&$deferredLinks) {
    return $item && !isset($deferredLinks[$item]);
});
unset($html, $url, $ahrefs, $framesrc);

logger("Found urls: " . join(", ", $links), 2);

//Note that URL has been deferred
foreach ($links as $href) {
    if ($href) {
        $deferredLinks[$href] = 1;
    }
}

foreach ($links as $href) {
    if ($href) {
        scan_url($href);
    }
}
$depth--;
}
```

```
// fnmatch() filler for non-POSIX systems

if (!function_exists('fnmatch')) {
    function fnmatch($pattern, $string)
    {
        return preg_match("#^" . strtr(preg_quote($pattern, '#'), array('\*' => '.*', '\?' => '.')) . "$#i", $string);
    } // end
} // end if

$version_functions = 2;
```

Sitemap.php

```
<?php



error_reporting(E_ALL);

ini_set('max_execution_time', 3000);
$site = $_POST['site'];

// Default sitemap filename
$file = "sitemap.xml";
$permissions = 0644;

// Depth of the crawl, 0 is unlimited
```

```
$max_depth = 0;

// Show changefreq
$enable_frequency = false;

// Show priority
$enable_priority = false;

// Default values for changefreq and priority
$freq = "daily";
$priority = "1";

// Add lastmod based on server response. Unreliable and disabled by default.
$enable_modified = false;

// Disable this for misconfigured, but tolerable SSL server.
$curl_validate_certificate = true;

// The pages will be excluded from crawl and sitemap.
// Use for exluding non-html files to increase performance and save bandwidth.
$blacklist = array(
    "*.jpg",
    "*/secrets/*",
    "https://www.knyz.org/supersecret"
);

// Enable this if your site do requires GET arguments to function
```

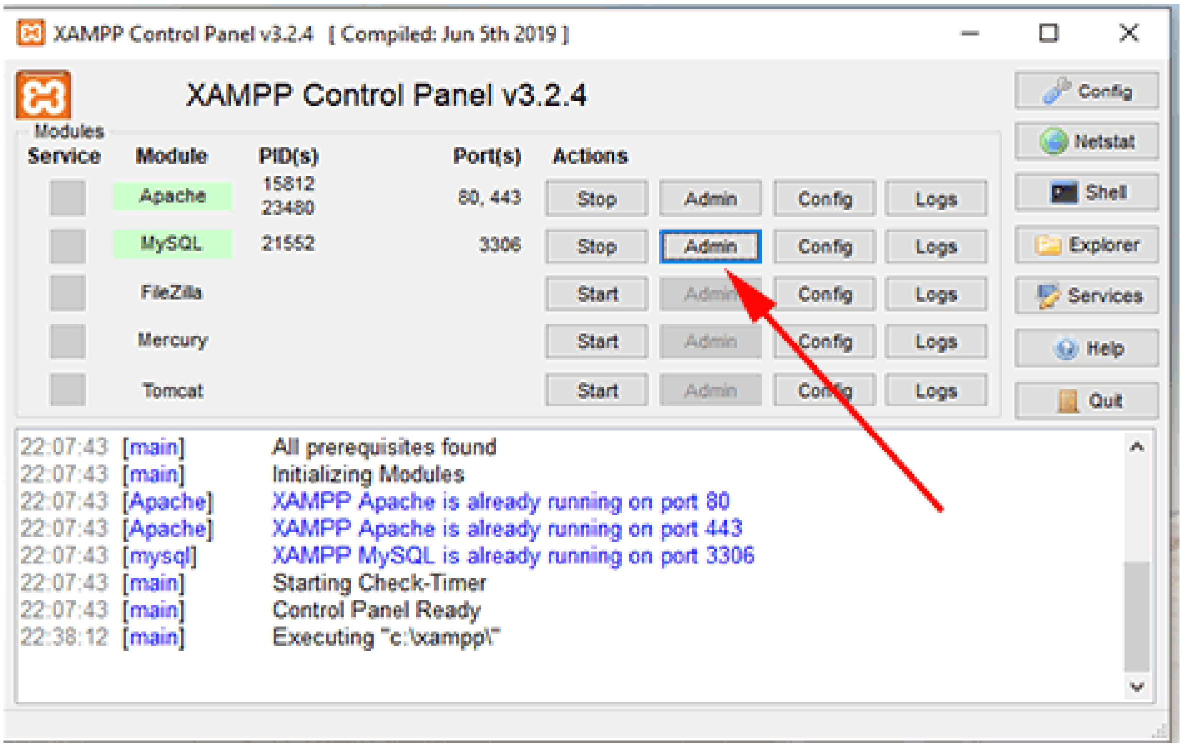
```
$ignore_arguments = false;

// Not yet implemented. See issue #19 for more information.
$index_img = false;

//Index PDFs
$index_pdf = true;

// Set the user agent for crawler
$crawler_user_agent = "Mozilla/5.0 (compatible; Sitemap Generator Crawler; +https://github.com/knyzorg/Sitemap-Generator-Crawler)";

// Header of the sitemap.xml
$xmlheader ='<?xml version="1.0" encoding="UTF-8"?>
<urlset
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">';

// Optionally configure debug options
$debug = array(
    "add" => true,
    "reject" => false,
    "warn" => false
);
```

```
//Modify only if configuration version is broken
$version_config = 2;



// Include all functions
require_once('sitemap.functions.php');

//Default html header makes browsers ignore \n
header("Content-Type: text/plain");

$color = false;

$version_script = 2;

if ($version_script != $version_functions || $version_functions !=
$version_config){
        logger("Script versions mismatch!",3);
        logger("Update necessary",3);
        logger("Version of sitemap.functions.php " .$version_functions ,3);
        logger("Version of sitemap.config.php " .$version_config ,3);
        logger("Version of sitemap.php " .$version_script ,3);
        logger("Download new files here: https://www.github.com/knyzorg/sitemap-
generator-crawler" ,3);
        die("Stopped.");
}
```

```
// Add PHP CLI support
if (php_sapi_name() === 'cli' && PHP_OS != 'WINNT') {
    parse_str(implode('&', array_slice($argv, 1)), $args);
    $color = true;
}

//Allow variable overloading with CLI
if (isset($args['file'])) {
    $file = $args['file'];
}
if (isset($args['site'])) {
    $site = $args['site'];
}
if (isset($args['max_depth'])) {
    $max_depth = $args['max_depth'];
}
if (isset($args['enable_frequency'])) {
    $enable_frequency = $args['enable_frequency'];
}
if (isset($args['enable_priority'])) {
    $enable_priority = $args['enable_priority'];
}
if (isset($args['enable_modified'])) {
    $enable_modified = $args['enable_modified'];
}
if (isset($args['freq'])) {
    $freq = $args['freq'];
}
if (isset($args['priority'])) {
    $priority = $args['priority'];
}
if (isset($args['blacklist'])) {
    $blacklist = $args['blacklist'];
}
if (isset($args['debug'])) {
    $debug = $args['debug'];
}
if (isset($args['ignore_arguments'])) {
    $ignore_arguments = !!$args['ignore_arguments'];
}
if (isset($args['pdf_index'])) {
    $pdf_index = $args['pdf_index'];
}

//Begin stopwatch for statistics
$start = microtime(true);

//Setup file stream
$tempfile = tempnam(sys_get_temp_dir(), 'sitemap.xml.');
$file_stream = fopen($tempfile, "w") or die("Error: Could not create temporary file $tempfile" . "\n");

fwrite($file_stream, $xmlheader);
```

```
// Global variable, non-user defined
$depth = 0;
$indexed = 0;
$scanned = array();
$deferredLinks = array();

// Reduce domain to root in case of monkey
$real_site = domain_root($site);

if ($real_site != $site){
    logger("Reformatted site from $site to $real_site", 2);
}

// Begin by crawling the original url
scan_url($real_site);

// Finalize sitemap
fwrite($file_stream, "</urlset>\n");
fclose($file_stream);

// Pretty-print sitemap
 if ((PHP_OS == 'WINNT') ? `where xmllint` : `which xmllint`) {
    logger("Found xmllint, pretty-printing sitemap", 0);
    $responsevalue = exec('xmllint --format ' . $tempfile . ' -o ' . $tempfile . ' 2>&1',
$discardedoutputvalue, $returnvalue);
    if ($returnvalue) {
        die("Error: " . $responsevalue . "\n");
    }
}

// Generate and print out statistics
$time_elapsed_secs = round(microtime(true) - $start, 2);
logger("Sitemap has been generated in " . $time_elapsed_secs . " second" .
(($time_elapsed_secs >= 1 ? 's' : '') . "and saved to $file"), 0);
$size = sizeof($scanned);
logger("Scanned a total of $size pages and indexed $indexed pages.", 0);

// Rename partial file to the real file name. `rename()` overwrites any existing files
rename($tempfile, $file);

// Apply permissions
chmod($file, $permissions);

// Declare that the script has finished executing and exit
logger("Operation Completed", 0);
```

SNAPSHOTS



XAMPP Control Panel v3.2.4 [Compiled: Jun 5th 2019]

XAMPP Control Panel v3.2.4

Modules

Service	Module	PID(s)	Port(s)	Actions
	Apache	15812 23480	80, 443	Stop Admin Config Logs
	MySQL	21552	3306	Stop Admin Config Logs
	FileZilla			Start Admin Config Logs
	Mercury			Start Admin Config Logs
	Tomcat			Start Admin Config Logs

```
22:07:43  [main]     All prerequisites found
22:07:43  [main]     Initializing Modules
22:07:43  [Apache]   XAMPP Apache is already running on port 80
22:07:43  [Apache]   XAMPP Apache is already running on port 443
22:07:43  [mysql]    XAMPP MySQL is already running on port 3306
22:07:43  [main]     Starting Check-Timer
22:07:43  [main]     Control Panel Ready
22:38:12  [main]     Executing "c:\xampp\"
```

Fig.6 Xampp server

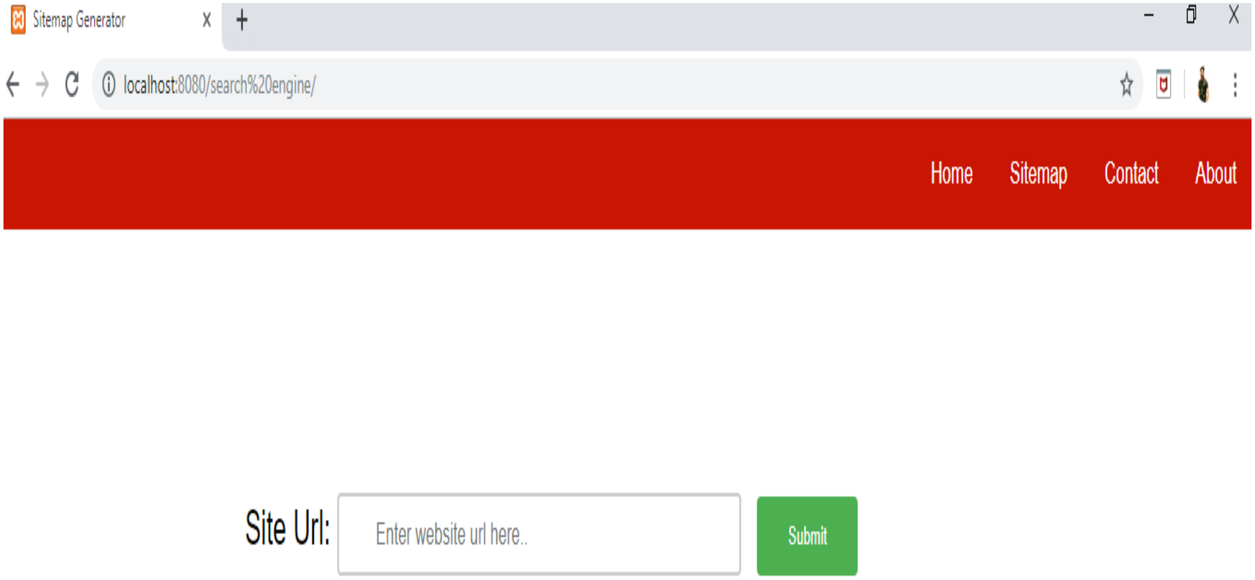
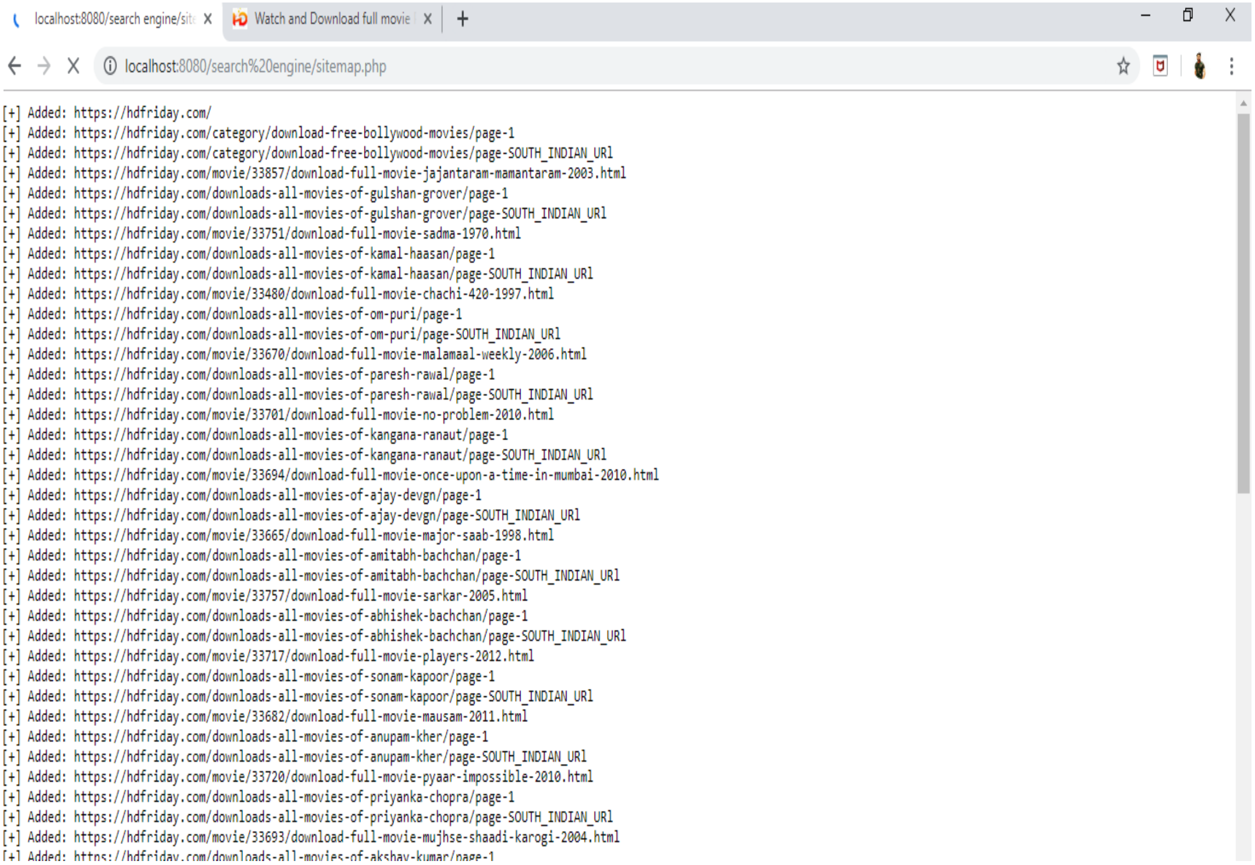


Fig.7 Implemented site view



localhost:8080/search engine/sit
Watch and Download full movie
localhost:8080/search%20engine/sitemap.php

```
[+] Added: https://hdfriday.com/
[+] Added: https://hdfriday.com/category/download-free-bollywood-movies/page-1
[+] Added: https://hdfriday.com/category/download-free-bollywood-movies/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33857/download-full-movie-jajantaram-mamantaram-2003.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-gulshan-grover/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-gulshan-grover/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33751/download-full-movie-sadma-1970.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-kamal-haasan/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-kamal-haasan/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33480/download-full-movie-chachi-420-1997.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-om-puri/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-om-puri/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33670/download-full-movie-malamaal-weekly-2006.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-paresh-rawal/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-paresh-rawal/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33701/download-full-movie-no-problem-2010.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-kangana-ranaut/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-kangana-ranaut/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33694/download-full-movie-once-upon-a-time-in-mumbai-2010.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-ajay-devgn/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-ajay-devgn/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33665/download-full-movie-major-saab-1998.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-amitabh-bachchan/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-amitabh-bachchan/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33757/download-full-movie-sarkar-2005.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-abhishek-bachchan/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-abhishek-bachchan/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33717/download-full-movie-players-2012.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-sonam-kapoor/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-sonam-kapoor/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33682/download-full-movie-mausam-2011.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-anupam-kher/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-anupam-kher/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33720/download-full-movie-pyaar-impossible-2010.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-priyanka-chopra/page-1
[+] Added: https://hdfriday.com/downloads-all-movies-of-priyanka-chopra/page-SOUTH_INDIAN_URl
[+] Added: https://hdfriday.com/movie/33693/download-full-movie-mujhse-shaadi-karogi-2004.html
[+] Added: https://hdfriday.com/downloads-all-movies-of-akshay-kumar/page-1
```

Fig.8 Directory Generation

CONCLUSION

SEO is a smart way of increasing your visibility online by saving time and cost of marketing. But before going for SEO, one must understand how it works and the affects it can have on the business. Beside this there are certain things like service/product quality, customer relationship, pre-sale service etc., which are the bases of any business. One must keep in the mind that businesses can run successfully without SEO, but they can’t run without these things.

REFERENCES

- https://github.com/search?q=search+engine+optimization
- https://www.apachefriends.org/download.html
- https://searchenginewatch.com/showpage.html?page=3628837
- Edgar Damian Ochoa: An Analysis of the Application of selected SEO techniques and their effectiveness on Google's search ranking algorithm, California state university, Northridge, May'2012 - csun-dspace.calstate.edu.
- Najam Nazar: Exploring SEO Techniques for Web 2.0Websites, Department of Computer Science and Engineering Chalmers University of Technology Göteborg, Sweden, June 2009, Publikationen registrerades 2009-08-25. Den ändrades senast 2013-04-04, Examiner: C. Carlsson.
- http://www.brightworkweb.com/images/search_engine_marketing.jpg
- http://www.webconfs.com/SEO-tutorial/introduction-to-SEO.php
- http://searchengineland.com/21-essential-SEO-tips-techniques-11580
- http://mashable.com/2012/01/09/increase-Google-rank/
- Khalil ur Rehman and Muhammad Naeem Ahmed Khan: The Foremost Guidelines for Achieving Higher Ranking in Search Results through Search Engine Optimization, International Journal of Advanced Science and Technology, Vol. 52, March, 2013.
- K.Chiranjeevi , K.Archana and J.Pradeep Kumar, "Design and Implementation of a Cost Effective Ranking Adaptation Algorithm", ISROSET-International Journal of Scientific Research in Computer Science and Engineering, Volume-01, Issue-05, Page No (24-24), Sep -Oct 2013.