



Enhancement in Crawling and Searching

(Using Extended Weighted Page Rank Algorithm based on VOL)

Isha Mahajan¹, Ms. Harjinder Kaur², Dr. Darshan Kumar³

¹M.Tech Student, ²Head of Department, ³Director

Department of Computer Science & Engineering

SSIET, Dinanagar - 143531, Distt. Gurdaspur, Punjab (India)

¹ishamahajan90@gmail.com, ²harrysaini988@gmail.com, ³darshanjind@gmail.com

Keywords: As the World Wide Web is becoming gigantic day by day, the number of web pages is increasing into billions around the world. To make searching much easier for users, search engines came into existence. Search engines are used to find specific information on the WWW. Without search engines, it would be almost impossible for us to locate anything on the Web unless or until we know a specific URL address. Every search engine maintains a central repository or databases of HTML documents in indexed form. Whenever a user query comes, searching is performed within that database of indexed web pages. The size of a repository of every search engine cannot keep each page available on the WWW. So it is desired that only the most relevant and important pages be stored in the database to increase the efficiency of search engines.

This search engine database is maintained by special software called "Crawler." A Crawler is a software that traverses the web and downloads web pages. Web Crawlers are also known as "Web Spiders," "Robots," "Internet Bots," "Agents" and automatic Indexers" etc. Broad search engines, as well as many more specialized search tools, rely on web crawlers to acquire large collections of pages for indexing and analysis. Since the Web is a distributed, dynamic and rapidly growing information resource, a crawler cannot download all pages. It is almost impossible for crawlers to crawl the whole web pages from World Wide Web. Crawlers crawl the only fraction of web pages from World Wide Web. So a crawler should observe that the fraction of pages crawled must be most relevant and the most important ones, not just random pages. The crawler is an important module of a search engine. The quality of a crawler directly affects the searching quality of search engines. In our Work, we propose to improve the crawling of a web crawler, to crawl only relevant and important pages from WWW, which will lead to reduced server overheads. With our proposed architecture we will also be optimizing the crawled data by removing least used or never browsed pages. The crawler needs a huge memory space or database for storing page content etc, by not storing irrelevant and unimportant pages and never removing accessed pages, we will be saving a lot of memory space that will eventually speed up the queries to the database. In our approach, we propose to use Extended Weighted page rank based on visits of links algorithm to sort the search results, which will reduce the search space for users, by providing mostly visited pages and most time devoted pages by the user on the top of search results list. Hence reducing search space for the user.

Keywords— Web Crawler, Extended Weighted Page Rank based on Visits of links, Weighted Page Rank, Page Rank, Page Rank based on visit of links, Search Engine, Crawling, bot, Information Retrieval Engine, Page Reading Time, User Attention Time, World Wide Web, Inlinks, Outlines, Web informational retrieval, online search.

1. Introduction

According to Internet World Stats survey as on as on March 25, 2017 - 3,731,973,423 (app. 373 billion) people use the internet [1]. Among them, 50.2% internet users are from Asia. According to Pew Research Center is the Internet and American Life Project Survey Report, 59% adults use web search engines on a daily

basis [2]. On the average 75% traffic to a website is generated by the Search Engine [3].

World Wide Web is growing rapidly day by day, the number of web pages is increasing into millions and billions around the world. According to worldwidewebsite.com on 27 May 2017 - World Wide Web contains at least 47.2 billion pages [4]. To make

searching for information much easier for users, web search engines came into existence. Web Search engines are used to find specific information on the World Wide Web. Without search engines, it would be almost impossible for us to locate anything on the internet unless or until we know a specific URL address. Every search engine maintains a central repository or databases of HTML documents in indexed form. Whenever a user comes and asks the query, searching is performed within the search engine database of indexed web pages.

The database or central repository of Search Engine is maintained by special software called "Web Crawlers." A Web Crawler is a program that visits Web sites and reads their pages and other information to create entries in a search engine index. A Web crawler may also be called a Web Spider, an Ant, an Automatic Indexer, an Agent, a Worm, a Wanderer, a Harvester. The crawler is a major component of a web search engine. The quality of a crawler directly affects the searching quality of search engines.

The biggest problem to deal with is the size of the Web, which currently is in the order of billions of pages [5]. Since the Web is a distributed, dynamic and rapidly growing information resource, a crawler cannot download all pages. It is almost impossible for crawlers to crawl whole web pages from World Wide Web. So, Crawlers crawls the only fraction of web pages from WWW. Therefore the size of a repository of every search engine cannot accommodate each page available on the WWW. This large size of web induces a low coverage problem, with no search engine indexing more than one-third of the publicly available Web [6]. So crawler should crawl only important and relevant pages.

2. Web Crawler

A Web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. The major search engines on the Web like **Google, Bing, Yahoo, Ask, DuckDuckGo** and **Baidu** all have such a program, which is also known as a "spider" or a "bot." Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly [7].

Crawlers are typically programmed to visit sites that have been submitted by their owners as new or updated. Entire sites or specific pages can be selectively visited and indexed. Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read [8].

A. Web Crawling - A Recursive Process

In the simplest form, a crawler starts from a seed page and then uses the external links or outlines within it to crawl other pages. The process repeats with the new pages offering more external links to follow until a

sufficient number of pages are identified, or some higher level objective is reached [9]. So Web crawling is a recursive process. Here is the process that a web crawler follows [10]:

1. Start from one preselected page. We call this starting page the "**Seed Page.**"
2. Extract all the links on that page.
3. Follow each of those links to find new pages.
4. Again extract all the links from all of the new pages found at step 3.
5. Now again follow each of those links found at step 4, to find new pages and extract all the links from all of those new pages found.
6. Similarly, this crawling process goes on and on until a sufficient number of pages are identified, or some higher level objective is reached.

B. Architecture of Web Crawler

Following are the components of a Web Crawler:

- **Multi-threaded Downloader:** It downloads documents in parallel by various parallel running threads.
- **Parser:** Job of Parser is to extract the links to follow from downloaded web pages. It will find the new URLs to follow for a crawler. The parser also finds application specific data from a web page.
- **Scheduler:** Selects the next URL to be downloaded.
- **Crawl Frontier:** A queue is having all URLs of pages or queue having URLs of pages to be downloaded. Sometimes also called Priority Queue or Processing Queue.
- **World Wide Web:** Collection of interlinked documents.
- **Storage or Database Repository (DB):** To save the information extracted from downloaded documents such as body text and Metadata or any application specific information.

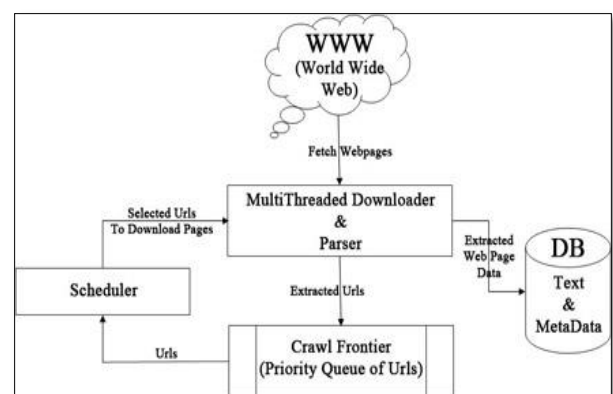


Fig. 1. Architecture of Web Crawler

C. Working of Web Crawler

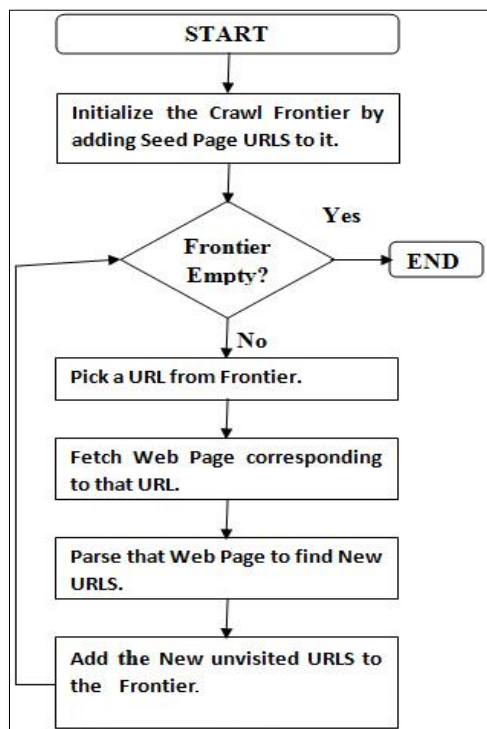


Fig. 2. Working of Web Crawler

The basic working of a web-crawler can be discussed as follows [10]:

1. Select a starting seed Url or URLs.
2. Add it to the frontier.
3. Now pick the URL from the border.
4. Fetch the web page corresponding to that URL.
5. Parse that web page to find new URL links.
6. Add all the newly found URLs into the frontier.
7. Go to step 2 and repeat while the Frontier is not empty.

D. Crawling Policy

The behavior of a Web crawler is the outcome of a combination of following policies. A Standard Crawlers follow these well define policies to crawl the web pages. Below are the standard policies [10]:

- A **Selection policy** that states which page to download. Depending on page selection criteria, it will select the pages for crawling.
- A **Re-visit policy** that states when to check for changes to the pages.
- A **Politeness policy** that states how to avoid overloading websites. Crawler should not send successive requests to a single web server at the same time. There should be some interval (like 10, 20 or 30 seconds) for sending another page request for the same crawler. If a Crawler is performing multiple requests per second and downloading large files, a server would have a

hard time keeping up with requests from multiple crawlers. Also at same time user asking for a page to the web server, user's response time will also get affected.

- A **Parallelization policy** that states how to coordinate distributed web crawlers. It takes care of same pages not being crawled by multiple instances running of the single crawler on different machines.

E. Controlling web crawlers to see the website content

The first thing a crawler is supposed to do when it visits your site is looking for a file called "robots.txt." This file contains instructions for the spider on which parts of the website to index, and which parts to ignore. The only way to control what a spider sees on your site is by using a robots.txt file. All spiders are supposed to follow some rules, and the major search engines do perform these rules for the most part. Fortunately, the major search engines like Google or Bing are finally working together on standards. An ethical crawler will consider "robots.txt," but spamming crawlers simply ignore robots.txt, so spamming crawlers can be blocked with .htaccess file programming [10].

3. Literature Review

F. Data Mining

Data mining can be defined as the process of extracting useful information from a large amount of data. The application of data mining to extract relevant information from the web is called as web mining [11]. So the Web mining is a data mining technique used to extract information from World Wide Web. It plays a vital role in search engines for ranking of web pages and can be divided into three categories [12]:

- **Web Content Mining (WCM):** It is the process of extracting useful information from the contents of web documents. This mining technique is used on the web documents and results page that are obtained from a search engine. WCM is to mine the content of web pages.
- **Web Structure Mining (WSM):** It is the processes of discovering link structure of the hyperlinks in inter documents level from the internet. It is used in many application areas. Page Rank Algorithm, Weighted Page Rank Algorithm and Weighted Page Rank based on Visits of Links Algorithm uses WSM to compute the web page score or rank.
- **Web Usage Mining (WUM):** It is the process to discover interesting usage patterns from web data to understand and better serve the needs of web-based applications. Also, WUM to extract information from the web server logs.

G. Page Rank Algorithm

Page Rank was developed at Stanford University by Larry Page and Sergey Brin (Ph.D. Research Scholars,

also founders of Google) in 1996. Page Rank uses the hyper link structure of The Web [13]. It is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. It considers the back link in deciding the rank score. If the addition of the ranks of all the back links of the page is significant then the page has large rank. A simplified version of Page Rank is given below [13]:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (1)$$

Notations used are:

- u and v represent the web pages.
- B(u) is the set of pages that point to u.
- PR(u) and PR(v) are rank scores of page u and v respectively.
- N_v indicates the number of outgoing links of page v.
- C is a factor applied for Normalization.

In Page Rank, the rank of page P is evenly divided among its outgoing links. Later Page Rank was modified observing that not all users follow the direct links on WWW. Therefore, it provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of pages that are linking it [12]. If a backlink comes from an important page, then that backlink is given a higher weight than those backlinks comes from non-important pages. Thus, the modified version of Page Rank is given as:

$$PR(u) = (1 - d) + d \left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \dots + \frac{PR(Tn)}{C(Tn)} \right) \quad (2)$$

Where d is a damping factor which set its value to 0.85. d can be thought of as the probability of users following the links and could regard (1 - d) as the page rank distribution from non-directly linked pages. We assume several web pages T1 ... Tn which point to u web page. T1 is the incoming link page to page u and C(T1) are the outgoing links from page T1. Simplified formula is:

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v}$$

Page Rank algorithm is used by the famous search engine "Google." Page Rank algorithm is the most frequently used algorithm for ranking billions of web pages. During the processing of a query, Google's search algorithm combines pre-computed Page Rank scores with text matching scores to obtain an overall ranking score for each web page [13].

Page Rank algorithm uses Web Structure Mining (WSM) for finding links or backlinks of the web page.

H. Weighted Page Rank Algorithm

In 2004, Wenpu Xing et. al. [14] discussed a new approach known as weighted page rank algorithm (WPR). This algorithm is an extension of Page Rank algorithm. WPR takes into account the importance of both the links and the outlines of the pages and distributes rank scores based on the popularity of the pages.

WPR performs better than the conventional Page Rank algorithm regarding returning larger number of relevant pages to a given query. According to author the more popular web pages are the more linkages that other web pages tend to have to them or are linked to by them. A Weighted Page Rank Algorithm – assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outline pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$.

$W_{in}(v,u)$ given in the following equation is the weight of link(v, u) calculated based on the number of links of page u and the number of links of all reference pages of page v.

$$W_{(v,u)}^{in} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (4)$$

Where I_u and I_p represent the number of links of page u and page p, respectively. R(v) denotes the reference page list of page v. $W_{out}(v,u)$ given in the following equation is the weight of link(v, u) calculated based on the number of outlines of page u and the number of outlinks of all reference pages of page v.

$$W_{(v,u)}^{out} = \frac{O_u}{\sum_{p \in R(v)} O_p} \quad (5)$$

Where O_u and O_p represent the number of outlinks of page u and page p, respectively. R(v) denotes the reference page list of page v. Considering the importance of pages; the original Page Rank formula is modified in the following equation:

$$WPR(u) = d + (1 - d) \sum_{v \in B(u)} WPR(v) W_{(v,u)}^{in} W_{(v,u)}^{out} \quad (6)$$

Notations used are:

- u and v represent the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- WPR(u) and WPR(v) are rank scores of page u and v respectively.

- $W_{(v,u)}^{in}$ is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages (i.e. outlinks) of page v.
- $W_{(v,u)}^{out}$ is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages (i.e. outlinks) of page v.

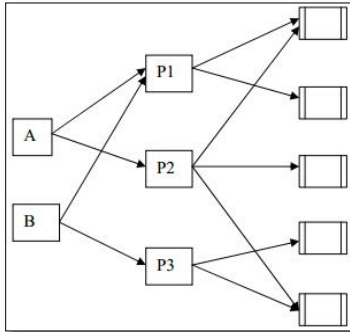


Fig. 3. Web Interlinked Structure

Weighted Page Rank algorithm uses Web Structure Mining (WSM) for finding links and outlinks of the webpage.

I. Page Rank based on Visits of Links Algorithm

In 2011, Gyanendra Kumar et. al. [15] proposed a new algorithm in which they considered user's browsing behavior. As most of the ranking algorithms proposed earlier are either a link or content oriented in which consideration of user usage trends are not available. They propose in their paper, a page ranking mechanism called Page Ranking based on Visits of Links (PR_{VOL}) is being devised for search engines, i.e. Page Rank and takes a number of visits of inbound links of web pages into account. This concept is very useful to display most valuable pages on the top of the result list by user browsing behavior, which reduces the search space to a large scale. In this paper as the author describe that in the original Page Rank algorithm, the rank score of page p, is evenly divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page p. So, he proposed an improved Page Rank algorithm. In this algorithm, we assign more rank value to the outgoing links which are most visited by users. In this manner, a page rank value is calculated based on visits of inbound links.

The modified version of Page Rank based on VOL is given in the following equation:

$$PR_{vol}(u) = d + (1 - d) \sum_{v \in B(u)} \frac{PR_{vol}(v)L_u}{TL(v)} \quad (7)$$

Notations used are:

- u and v represent the web pages.

- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- $PR_{vol}(u)$ and $PR_{vol}(v)$ are rank scores of page u and v respectively.
- L_u is the number of visits of the link which is pointing page u from v.
- TL(v) denotes a total number of visits of all links present on v.

Page Rank based on VOL algorithm uses Web Structure Mining (WSM) for finding inlinks of webpage and Web Usage Mining (WUM) to find visits of links.

J. Weighted Page Rank based on Visits of Links Algorithm

In 2012, Neelam Tyagi et. al. [16] discussed that the original Weighted PageRank algorithm assigns larger rank values to more important (popular) pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as recorded as $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$, respectively. Here we proposed an improved Weighted Page Rank algorithm. In this algorithm, we assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks. We do not consider here the popularity of outlinks which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links.

The modified version based on WPR_{VOL} is given in following equation:

$$WPR_{vol}(u) = d + (1 - d) \sum_{v \in B(u)} \frac{WPR_{vol}(v)W_{(v,u)}^{in}L_u}{TL(v)} \quad (8)$$

Notations used are:

- u and v represents the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- $WPR_{vol}(u)$ and $WPR_{vol}(v)$ are rank scores of page u and v respectively.
- $W_{(v,u)}^{in}$ is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages (i.e. outlinks) of page v.

Weighted Page Rank based on VOL algorithm uses Web Structure Mining (WSM) for finding inlinks of webpage and Web Usage Mining (WUM) to find visits of links.

K. Extended Weighted Page Rank based on Visits of Links Algorithm

In 2017, Isha Mahajan et. al. [28] proposes to enhance the quality of search and to display the most target oriented pages at the top of the search list. The new approach focuses on the user query preference, where consideration is done on the most useful or important pages. To determine the usefulness of pages, they take time spent on webpage i.e. User Activities Time (UAT) and Page Reading Time (PRT) as an essential factor along with Visits of Links (VOL) on that webpage. These all will decide the importance of a page. User Activities Time is the actual time spent by the user to read the webpage, which we suppose reflects the usefulness of information in the page as conceived by the user. This proposed approach will compute the rank according to visits of links of inbound links as well as user attention given to the web page. This algorithm behaves completely different from other page ranking algorithms because it takes users usage trends or user browsing behavior in its working.

So the Modified version of WPR_{VOL} is given in the following equation:

$$EWPR_{volT}(u) = (1 - d) + d \left[\frac{UAT(u)}{PRT(u)} \left(\sum_{v \in B(u)} \frac{EWPR_{volT}(v) W_{(v,u)}^{in} L_u}{TL(v)} \right) \right] \quad (9)$$

Notations used are:

- u and v represent the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- UAT(u) is the User Activities Time i.e. the total time the user spends on that web page by doing activities like cursor movement with the mouse, KeyPress, Touch, etc.
- PRT(u) is the Page Reading Time i.e. the time page has been actively opened in a browser tab. In more technical words, we can say when Focus is on that page.
- $EWPR_{volT}(u)$ and $EWPR_{volT}(v)$ are rank scores of page u and v respectively.
- $W_{(v,u)}^{in}$ is the weight of link(v, u) calculated based on the number of links of page u and the number of inlinks of all reference pages (i.e. outlinks) of page v.
- L_u is the number of visits of the link which is pointing page u from v.
- TL(v) denotes a total number of visits of all links present on v.

Weighted Page Rank based on VOL algorithm uses Web Structure Mining (WSM) for finding inlinks of webpage and Web Usage Mining (WUM) to find visits of links.

4. Related Work

Animesh Tripathy et al. [17] describe the design of a web crawler that uses Page Rank algorithm for crawling and distributed searches. He presents web mining architecture of the system and describes efficient techniques for achieving high performance. As WWW is growing rapidly and data in the present day scenario is stored in a distributed manner. The need to develop a search engine based architectural model for people to search through the Web. Broad web search engines, as well as many more specialized search tools, rely on web crawlers to acquire large collections of pages for indexing and analysis. In this paper, They describe the design of a web crawler that uses Page Rank algorithm for distributed searches and can be run on a network of workstations. The crawler scales to several hundred pages per second is resilient against system crashes and other events and can be adapted to various crawling applications.

Carlos Castillo et. al. [5] presents a comparative study of strategies for Web crawling. We show that a combination of breadth first ordering with the largest sites first is a practical alternative since it is fast, simple to implement, and able to retrieve the best-ranked pages at a rate that is closer to the optimal than other alternatives. Our study was performed on a large sample of the Chilean Web which was crawled by using simulators so that all strategies were compared under the same conditions and actually crawls to validate our conclusions. We also explored the effects of large-scale parallelism in the page retrieval task and multiple-page requests in a single connection for effective amortization of latency times.

Lay-Ki Soon et. al. [18] proposed URL signature to be implemented in web crawling, aiming to avoid processing duplicated web pages for further web crawling. The experimental result indicates that URL signature can reduce the processing of duplicated web pages significantly for further web crawling at a negligible cost compared to the one without URL signature.

Farha R. Qureshi et. al. [19] proposed since standard URL normalization was proposed to avoid processing of duplicate web pages, but it fails to do so when syntactically different URLs lead to similar web pages. It proves that considering only URL is not enough and needs some additional information. There comes the URL signature which considered not only URL but also body text. This works perfectly in reducing processing of duplicate pages.

Saurabh Pakhidde et. al. [20] proposed new crawler architecture. Web crawler rejects the page whose url does not contain the search keyword while searching information on World Wide Web. They proposed the architecture which scan the web pages and parse them check for their relevancy by assigning each page a page weight and arrange them in terms of most relevant page first and then the second page and so on according to the page weight, so that we may gain more relevant information or site addresses at top of result.

Thus we will get more accuracy while searching some information on network.

Prashant Dahiwale et. al. [21] proposed a crawler that follows a completely new crawling strategy to compute the relevance of the page. It analyses the content of the page based on the information contained in various tags within the HTML source code and then computes the total weight of the page. The page with the highest weight thus has the maximum content and highest relevance. This crawler is best suitable where true analysis of data is needed such as business analysis.

Sachin Gupta et. al. [27] proposed the crawler architecture to improve the crawling process using very technologies like Hashing Algorithm and Standard Normalization to remove the crawling of duplicate pages. Also Weighted Page Rank algorithm used to stop the crawling of irrelevant and unimportant pages.

5. Problem Formulation

From the earlier work done, we have formulated these problem statements:

- i. As World Wide Web is becoming giant day by day and Search Engines crawling Module (Web Crawler) in today's dynamic world crawls only fraction of web pages from World Wide Web. So a Crawler should observe that the fraction of pages crawled must be most relevant and the most important ones, not just random pages.
- ii. Crawling of unimportant data from web pages, leads to server overhead and network load. Therefore critically affecting Server performance.
- iii. Crawling of redundant data from web pages, again leads to network load, server overhead and wasting a lot of memory space.
- iv. General purpose search engines retrieve and download unwanted information. Making Search Space for the user too large to search. Therefore search space for user needs to be reduced, or user deserves the most valuable pages on the top of the results list, as it will save the user's network bandwidth and user's most valuable time.
- v. The execution time of search engine (i.e. the time in which records are retrieved by users) needs to be reduced.
- vi. The crawler needs a very large memory space of database for storing page content etc, by removing irrelevant/redundant pages, we will be saving a lot of memory space, that will eventually speed up the searches (queries) from the database.

6. Proposed Work

Our new approach is a hybrid of various technologies used earlier as well as new algorithms.

1. Our crawler like other crawlers starts with a seed URL or multiple seeds.
2. It fetches the page corresponding to that seed URL and creates an object of that page and stores it into a pages objects queue.
3. After that outlines of the seed page from that object are fetched, and their objects are stored in the queue.
4. Similarly, outlines of the pages from their page objects, found at step 3 are fetched, and their objects are stored in the queue.
5. So, this is a recursive process goes on and on, until maximum no of pages have been fetched or up to a threshold depth level or until the required higher level objective be achieved.

Note: Queue in our approach will store that objects of fetched pages. We can also say this queue is a pages objects array.

L. Structure of an Object of Page

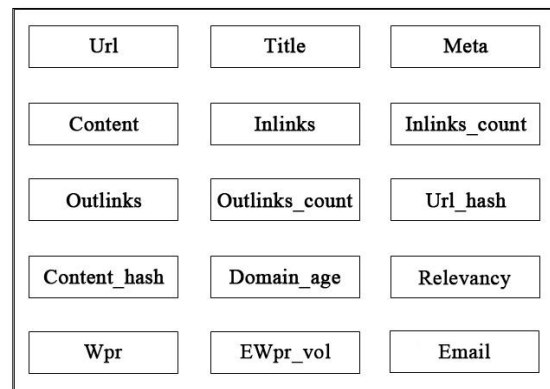


Fig. 4. Page Object Structure

Page object stores the URL, title, meta, content, links, inlinks_count, outlinks, outlinks_count, url_hash, content_hash, domain_age, relevancy, wpr, ewpr_vol, email fields.

- **Url** field will store the unique address for a page that is accessible on the Internet.
- **Title** field will store the title of the page.
- **Meta** field will store the content of meta elements of the page.
- **The content** field will store the content of body element (<body>) of the page.
- **Inlinks** field will store URL of web pages that point towards this page.
- **Inlinks_count** field will store the value of total no. of links of this page.
- **Outlines** field will store url of web pages that point to other web pages from this page.
- **Outlinks_count** field will store the value of total no. of outlinks of this page.

- **Url_hash** field will store the signature/hash of url of the page, generated with a md5 hashing algorithm.
- **Content_hash** field will store the signature of body content of the page, generated with a md5 hashing algorithm.
- **Domain_Age** field will store the age of domain name of the page.
- **Relevancy** field will store the relevancy weight or score of a page.
- **Wpr** field will store the score or rank of the page, computed using Weighted Page Rank (WPR) Algorithm.
- **EWpr_vol** field will store the score or rank of the page, computed using Extended Weighted Page Rank based on Visits of links (EWPR_{vol}) Algorithm.

- **Email** field will store the single email address or multiple email addresses if available on that page.

Pages will be fetched to crawl in the breath first approach as displayed in fig. 5 , where A is the Seed Url and pages B, C, D are its outlinks. After A, its outlinks B then C and then D will be fetched. Similarly E, F and G are the outlinks of pages B. After D, outlinks of B (E, F, G) will be fetched to process and so on.

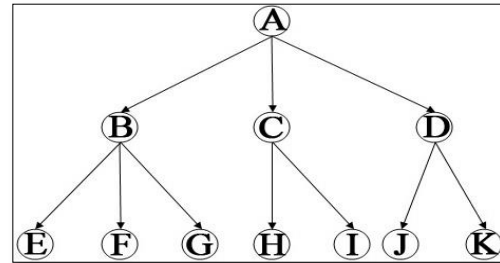


Fig. 5. Pages Tree Structure

discarded. If URL signature of that object does not exist in the database, then the signature of body content from that page object will be compared with the earlier stored body content signatures of pages in the database. If that page content signature exists in the database, then that page (page object) will be discarded. Otherwise, that page object will be stored in the new queue, which is an optimized version of the earlier queue. We call that queue as an optimized queue.

In the next step, we will calculate the Domain Age of URLs from those page objects in the queue and compare domain age of page with a threshold value. We have taken threshold value for domain is 365 days. If domain age is less than threshold we assumed, then we will check the relevancy of the page, if relevancy is less than the threshold value (which is 3, Dahiwal et. al. assumed in [21]), we will discard that page object. Otherwise, if relevancy found greater than the threshold, store that page details from its object directly to the database.

M. Proposed Architecture of Web Crawler

Once we have an array of pages objects in computer memory (Queue), next we will remove the redundancy by removing duplicate content, duplicate URLs and near duplicate URLs using Standard Normalization (STN) of URLs and Hashing Algorithm. We will use a md5 hashing algorithm to find the signature of URL and signature of body content of each page. STN of URLs will remove redundancy by finding syntactically same pages. Syntactically same URLs are considered as equivalent URLs. After Url Normalization, We have a new page objects queue with canonical or normalized URLs that will be supplied for Signature comparison.

In the next step, first, we compare the signature of page URL from a page object with the signature of all the pages URLs earlier stored in the database if that particular signature exists in database then the page (page object) having that URL will be ignored or

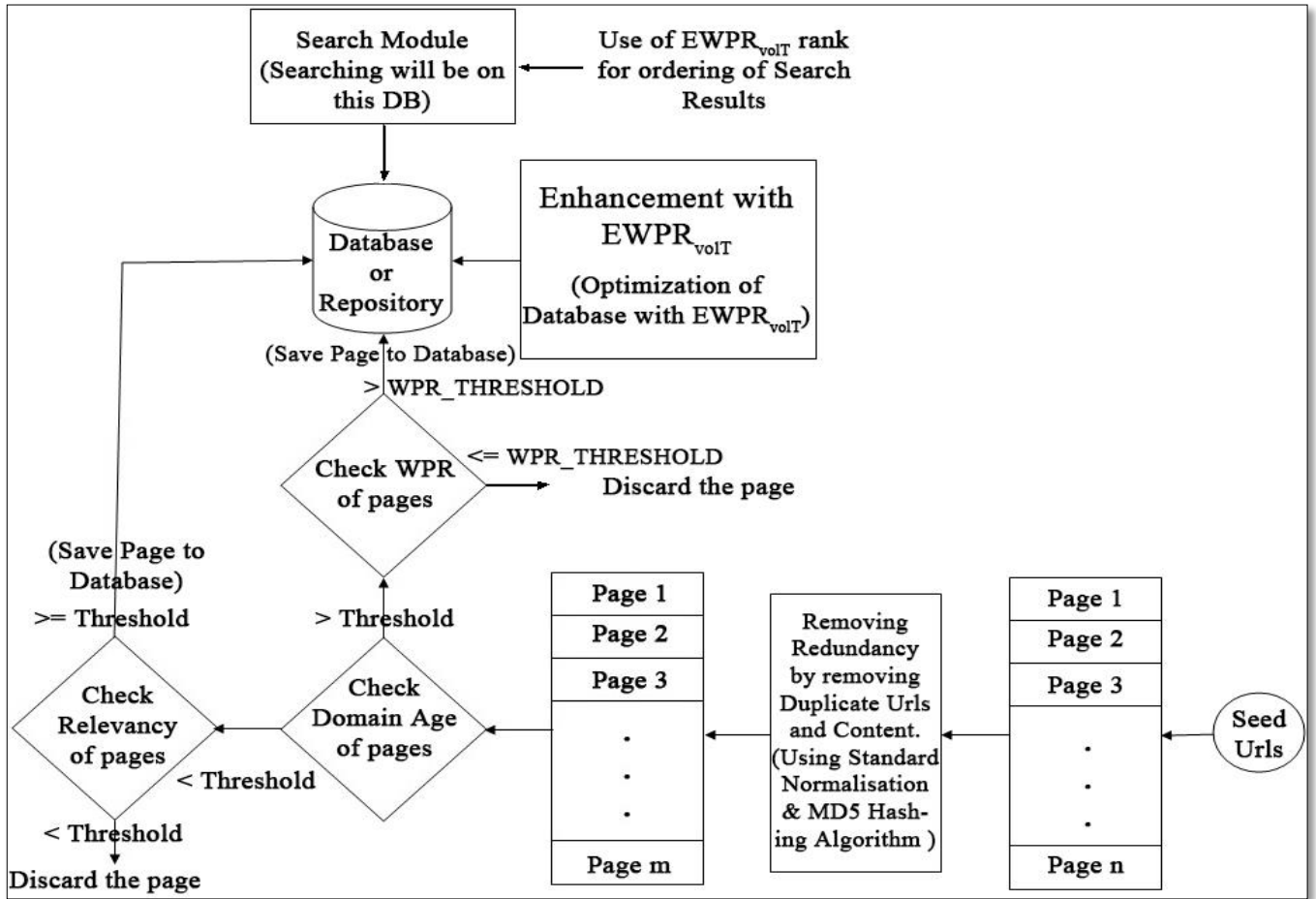


Fig. 6. Proposed Architecture of Web Crawler

In the next step, we will calculate the Domain Age of URLs from those page objects in the queue and compare domain age of page with a threshold value. We have taken threshold value for domain is 365 days. If domain age is less than threshold we assumed, then we will check the relevancy of the page, if relevancy is less than the threshold value (which is 3, Dahiwale et. al. assumed in [21]), we will discard that page object. Otherwise, if relevancy found greater than the threshold, store that page details from its object directly to the database.

Relevancy of page is calculated using the approach of **Dahiwale et. al.** used in predicting relevancy of pages in his rank crawler [21].

If Domain age of page is more than the threshold, then another score/rank of the page will be computed using Weighted Page Rank (WPR) Algorithm [14], if score falls less than WPR_THRESHOLD (threshold value of WPR), we will discard that page object. If the score is more than the threshold value, that page will be added to the repository.

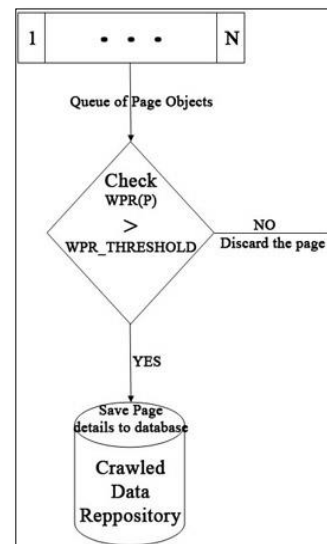


Fig. 7. Block Diagram of WPR Check Module for page

0.15 is a magic number or minimum threshold, we have set for weighted page rank comparison of pages. 0.15 is minimum WPR value that a page has if it has zero inlink or zero Outline.

While storing the pages in the database, this proposed system will send an email containing our Extended Vol-Analytics code (a client-side script) to

Website Administrator. The purpose of sending email is to add that script to his/her website. Once Web Admin will add our script to their websites, our search engine will start tracking their website visits of links activities, Page Reading Time and User Attention Time.

This Extended Vol-Analytics Script is developed using AJAX, runs on background. This Script sends details of webpage like its address, its caller address, its User Attention Time, its Page reading Time, how many no. of times page got called from the particular resource and its website id to our server where those got tracked and stored in our database.

Minified Version of Our Extended VOL-Analytics Code :

```
<script type="text/javascript">
/* Extended VOL with Time Analytics Code */
<!--
var
timeoutID=0,start,end,pagofocustime=0,opentime,midtime=0,exactime=0,inactivetime=1e4,idletime=0,window_focus=!0;document.addEventListener("DOMContentLoaded",function(e){function t(){this.addEventListener("mousemove",i,!1),this.addEventListener("mousedown",i,!1),this.addEventListener("keypress",i,!1),this.addEventListener("DOMMouseScroll",i,!1),this.addEventListener("mousewheel",i,!1),this.addEventListener("touchmove",i,!1),this.addEventListener("MSPointerMove",i,!1),n()}function n(){timeoutID=window.setTimeout(o,1e4)}function i(e){window.clearTimeout(timeoutID),d()}function o(){idletime+=inactivetime,window_focus&&n()}function d(){n()}start=performance.now(),midtime=start,window.onblur=function(){window_focus=!1,end=performance.now(),pagofocustime+=end-midtime,exactime+=end-midtime},window.onfocus=function(){window_focus=!0,midtime=performance.now(),window_focus&&t(),window.onbeforeunload=function(){end=performance.now(),pagofocustime+=end-midtime,opentime=end-start,exactime=pagofocustime-idletime,window.XMLHttpRequest.prototype=new XMLHttpRequest.prototype.onreadystatechange=function(){4===xmlhttp.readyState&&200===xmlhttp.status&&console.log(xmlhttp.responseText)},""!=document.referrer&&(xmlhttp.open("GET","http://www.crawlsearch.xyz/tracking/update_visit.php?url="+location.href+"&caller_url="+document.referrer+"&webid=3&page_focus_time="+pagofocustime+"&exact_time="+exactime,!0),xmlhttp.send())}});
-->
</script>
```

To hide the business logic of our Extended VOL Script, Javascript code minification can be used. Minification refers to the process of removing unnecessary or redundant data without affecting how the resource is processed by the browser - e.g. code

comments and formatting, removing unused code, using shorter variable, function names and so on [29].

With Minification it becomes difficult for web masters or any other try to temper our extended vol analytics code and malicious attempts towards our server (where all tracking of visits of links, User Attention Time and Page Reading Time will be done and store in database) can be reduced. This Code Minification can be done using the javascript code minification API [30].

This code uses HTTP Referer for its working. HTTP referer (originally a misspelling of referrer) is an HTTP header field that identifies the address of the webpage that linked to the resource being requested. By checking the referer, the new webpage can see where the request originated [22]. Our code runs at the background, sends website id (web id), user attention time, page reading time, page URL and caller URL to our server, where we track those and store those in our database, to have visits of links of various web pages from various resources.

At the start, we have harvested the email address from pages. Email address is stored in page Object. Web Administrator will get only one mail for the behalf of a website or single domain.

After 45 days, A cron job on our server will execute one more script which will calculate the rank of web pages using Extended Weighted Page Rank based on Visits of Links (EWPR_{volT}) Algorithm [26]. This algorithm uses web usage mining (WUM) and can only work if we have Visited of links data available, user attention time and page reading time.

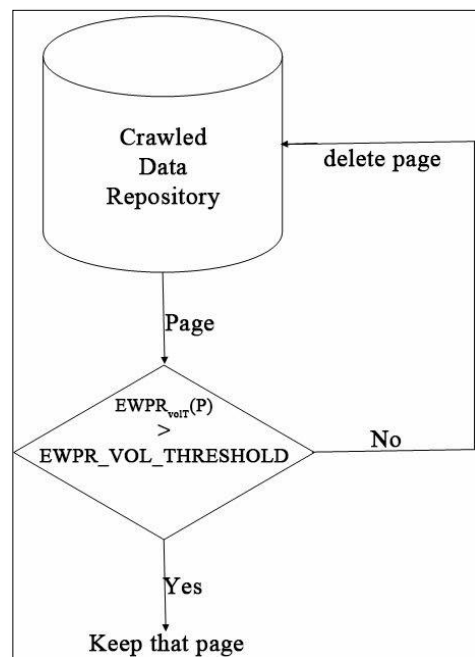


Fig. 8. Block Diagram of WPR_VOL Check Module for page

If the rank of some pages is less than EWPR_VOL_THRESHOLD (Extended WPRVOL Threshold value), we will optimize our database by deleting those pages. This Algorithm keeps those pages in the database, which are being actively used or

accessed by users. This script deletes never or less actively browsed (accessed) pages from the database. The ewprvolt algorithm considers the user browsing behavior of web pages over the web, removes pages decided by our threshold.

0.15 is the minimum EWPR_VOL_THRESHOLD value we kept that a page has if it has zero links or zero Outline or Zero visit (never accessed page).

Cron Jobs are used for scheduling tasks to run on the server. They are most commonly used for automating system maintenance or administration. However, they are also relevant to web application development. There are many situations when a web application may need certain tasks to run periodically [31].

N. Url Normalization or Standard Normalization of URLs

URL normalization (or URL canonicalization) is the process by which URLs are modified and standardized consistently. The goal of the normalization process is to transform a URL into a normalized or canonical URL so it is possible to determine if two syntactically different URLs may be equivalent. Search engines employ URL normalization to reduce indexing of duplicate pages. Web crawlers perform URL normalization to avoid crawling the same resource more than once [23].

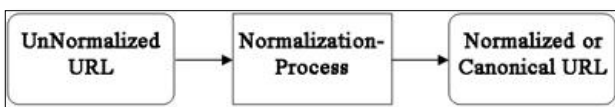


Fig. 9. Process of Url Normalization

The standard URL normalization (STN) is one of the tasks performed by web crawler during the process of web crawling. There is a set of predefined activities to be done for converting URLs into a canonical format. After the normalization, URLs which are syntactically identical are considered as equivalent, thus reducing the crawling of redundant web pages [19].

We will use Standard Normalization on URLs to remove optional parameters in URL and finding syntactically same URLs. Finding Syntactically Equivalent or identical URLs means removal of redundancy by not crawling again the redundant web pages. These duplicate URLs are removed at the time when Url Signature comparison is made with earlier stored page URLs signatures in the database. After URL Normalization, We have a new page objects queue with canonical or normalized URLs that will be supplied for Signature comparison. URL Normalization helps us to find near duplicate URLs to remove redundant web pages crawling.

Following are some examples of Standard Normalization of URLs:

- HTTP://WWW.EXAMPLE.COM is transformed into http://www.example.com.
- http://www.example.com:80 is turned into http://www.example.com.

- http://www.example.com/index.html is transformed into http://www.example.com.
- http://208.77.188.166 is turned into http://www.example.com (i.e. replacing it with domain name).
- https://www.example.com is transformed into http://www.example.com.
- http://www.example.com/bar.html#section1 is transformed into http://www.example.com/bar.html.
- http://www.example.com/alice is transformed into http://www.example.com/alice/.
- http://www.example.com/display?lang=en&article=fred is transformed into http://www.example.com/display?article=fred&lang=en.
- http://www.example.com/display?id=123&fakefoo=fakebar is transformed into http://www.example.com/display?id=123 (i.e. removing unnecessary query variables).
- https://www.example.com/display? is transformed into http://www.ex
- ample.com/display (i.e. removing empty query).

O. Removing Redundancy with MD5 Hashing Algorithm

Standard Normalization of URLs is successful to avoid processing of duplicate pages, when they are syntactically identical or equivalent but fails to do so for syntactically different URLs, which lead to crawling if similar web pages. For e.g. these websites in the table have syntactically different URLs, which leads to crawling of similar web pages.

TABLE I. SYNTACTICALLY DIFFERENT URLS

| Website 1 | Website 2 (Alias of Website 1) |
|---|---|
| http://www.dehreereihealing.com/ | http://www.reikichandigarh.com/ |
| http://www.chandasbestos.com/ | http://nonasbestosmillboard.com/ |
| https://www.justdial.com/ | https://www.justdial.in/ |
| http://www.prettyvilla.com/ | http://saidassandsons.com/ |
| http://www.himachaltouristguide.com/index.php/kangra/dharamsala | http://www.himachaltouristguide.com/index.php/districts-of-himachal/kangra/dharamsala |
| http://www.infopathankot.com | http://www.pathankotpunjab.com |
| http://www.himachaltouristguide.com/index.php/kullu/manali | http://www.himachaltouristguide.com/index.php/districts-of-himachal/kullu/manali |
| http://srijan.net/ | http://srijantechnologies.com/ |
| http://www.hotmail.com | http://www.live.com |
| http://www.epicwebsol.com | http://www.epicwebsolutions.in |

| | |
|-----------------------------|----------------------------|
| http://www.pathankothub.com | http://www.pathankothub.in |
|-----------------------------|----------------------------|

It proves that considering only Url Normalization to remove redundancy while crawling is not enough and needs other duplicity removal techniques as well.

There comes hashing or signature generation technique, which is very helpful for removing duplicity. To remove crawling of similar pages whose URL are different, we will compare the signatures generated by the md5 hashing algorithm.

The MD5 message-digest algorithm is a widely used cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32 digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications and is also commonly used to verify data integrity [23].

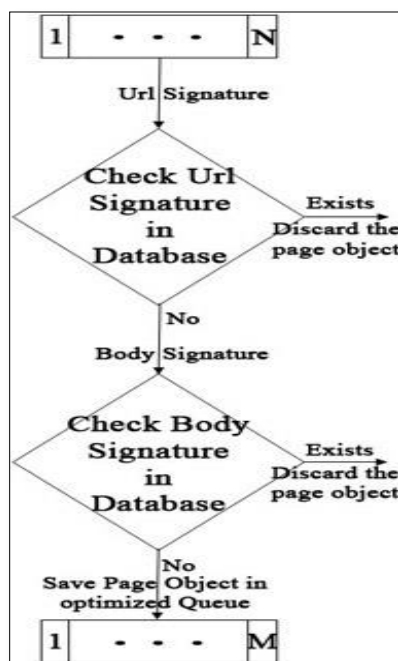


Fig. 10. Block Diagram of Removing Redundancy with Hash

In our approach, we will first compare the Url signature of every page from page object queue with earlier URLs signatures stored in the database, so that not to crawl same pages again. If Url signature exists in the database, we will discard that page. Otherwise if Url signature does not exist in the database, we will next compare the body content signature of same page object with body content signatures earlier stored in the database. If body content signature of the page already exists in the database, discard that page. Otherwise store that page object in the new queue, which is optimized one from the earlier queue. In this way with body text signature generation, similar or duplicate pages which are syntactically different will be discarded from recrawling. This works perfectly in reducing processing of duplicate pages.

P. How to Calculate Domain Age

Domain age is the age of a website on the Internet. To calculate domain age, WHOIS Protocol is used.

WHOIS (pronounced as the phrase “who is”) is a query and response protocol that is widely used for querying databases that store the registered users or assignees of an Internet resource, such as a domain name, an IP address block, or an autonomous system, but is also used for a wider range of other information. The protocol stores and delivers database content in a human-readable format [25].

To find domain age of a domain, we have used domain age checker script provided by Sunny Verma on GitHub developers community [32].

Our Domain Age Calculation Module gives us the age of website in no. of days (e.g. 2740 days).

We are using domain age concept in our proposed work because we are using Weighted Page Rank Algorithm and Extended Weighted page rank based on visits of links algorithm in our approach. These algorithms use links and outline of webpage for calculation of rank of webpage. Inlinks are possible for older aged domain/website only. For new born websites it is hardly possible to have links. However, Crawler crawls older age as well as new age websites both. So, before applying WPR Algorithm we will find the age of web page, if it is older aged (as per our domain age threshold criteria) WPR algorithm will be applied on it and depending upon the computed WPR Rank comparison with WPR_THRESHOLD, webpage will be stored in the database. Otherwise, relevancy of new age domain will be calculated and depending computed relevancy comparison with relevancy threshold; webpage will be stored in the database. Domain threshold we are assuming in this research is one year (365 days).

Older domains can get a little favor in edge in search engine ranking as well.

Q. How to calculate Relevancy of newly age webpage

We use Relevancy Parameter to decide, whether to store the new age page in database or not. This Relevancy Parameter will decide, the web page is beneficial to keep in a database or not. To decide relevancy, we are using the approach of Dahiwale et. al. [21] mentioned in their relevancy prediction research paper.

To check relevancy, we need a search string to be searched within the page. In our approach, search string will be the last string after the host name in Url.

For Example:

- History will be searching string in this Url http://www.infopathankot.com/history.htm
- Dharamsala will be searching string in this Url http://www.himachaltouristguide.com/index.php/kangra/dharamsala
- Web_crawler will be searching string in this URL http://en.wikipedia.org/wiki/Web_crawler
- Wheat Oats Upma Recipe will be searching string in this Url

http://www.chatpatirecipes.com/kids-recipes/wheat-oats-upma-recipe.php

- Computer Application Courses will be searching string in this Url
- http://www.ssgidinanagar.org/courses-offered/computer-application-courses/

For the cases where there is no string after host name or domain name, name before TLD (Top Level Domain like .com, .net, .org, .in etc) will be the search string.

For Example:

- Pathankot Punjab will be searching string in Url http://www.pathankotpunjab.com
- So did Technologies will be searching string in Url http://www.sodiztechnologies.com
- Educational Era will be searching string in this Url http://www.educationalera.com
- Himachal Tourist Guide will be searching string in Url http://www.himachaltouristguide.com
- Manimahesh Yatra will be searching string in Url http://www.manimaheshyatra.co.in
- Hosting Acres will be searching string in Url http://www.hostingacres.com

We will create and use data dictionary of words, with the help of which, our system will generate search string Pathankot Punjab from PathankotPunjab string. Himachal Tourist Guide will be generated from Himachal tourist guide string with use of Data Dictionary.

Web Crawler works using the source code of the page i.e. the downloader gets the content from a page, the parser uses the source code to analyze the tags and contents of the page so as to get the page weight and calculate the degree of relevancy of a page. The analysis part of the source code is as follows:

Let the Total weight of the page be 't' units

$$t = (N_t * T) + (N_m * M) + (N_h * H) + (N_b * B)$$

Where T = 4 units, M = 3 units, H = 2 units, B = 1 units

These values are assumed by considering search engines (like google, bing, etc) ratings to these tags.

Consider the source code associated with the web page of the URL: http://www.sodiztechnologies.com/web-development.php as given below:

```
<!DOCTYPE html>
<html>
  <head>
    <meta charset="UTF-8" />
    <meta name="keywords"
content="web development, web design"/>
```

```
<meta name="description"
content="Web Development and Web Design
tutorials"/>
<title>Web Development
Tutorials</title>
</head>
<body>
  <p>This Page contains the links of
Web development and design tutorials.</p>
  <h1 align="Center">Web
Development Tutorials</h1>
  <li><a
href="http://www.w3schools.com">W3Schools</a></li
>
  <li><a
href="http://www.w3schools.com">W3Schools</a></li
>
  <li><a
href="http://www.w3schools.com">W3Schools</a></li
>
</body>
</html>
```

Search String for this page by our approach will be web development and relevancy score or weight of the page is:

$$t = (1*4) + (2*3) + (1*2) + (2*1)$$

$$t = 4 + 6 + 2 + 2 = 14$$

Which is more than threshold value 3 (i.e. $t > 3$), so we can assume page is relevant and we can store it.

R. How to calculate Weighted Page Rank (WPR)

To calculate Weighted Page Rank of the Web page, following formula is suggested by Wen Xing et. al. [14].

$$WPR(u) = d + (1 - d) \sum_{v \in B(u)} WPR(v) W_{(v,u)}^{in} W_{(v,u)}^{out}$$

Where, the following are notations used in the formula:

- U and v represent the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- WPR(u) and WPR(v) are rank scores of page u and v respectively.
- $W_{(v,u)}^{in}$ is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages (i.e. outlinks) of page v.
- $W_{(v,u)}^{out}$ is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages (i.e. outlinks) of page v.

$$\text{Also } W_{(v,u)}^{in} = \frac{I_u}{\sum_{p \in R(v)} I_p}, W_{(v,u)}^{out} = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

Where I_u and I_p represent the number of inlinks of page u and page p , respectively and O_u and O_p represent the number of outlinks of page u and page p , respectively. $R(v)$ denotes the reference page list of page v .

This algorithm works by taking into account the importance of both the inlinks and outlinks of the web pages and distributes rank scores based on the popularity of the pages. The popularity from the number of inlinks and outlinks is recorded as $W_{(v,u)}^{in}$ and $W_{(v,u)}^{out}$.

S. How to calculate Extended Weighted Page Rank based on VOL

$EWPR_{volT}$ stands for Extended Weighted Page Rank based on Visits of Links. To calculate $EWPR_{volT}$, following formula is proposed [28]:

$$EWPR_{volT}(u) = (1 - d) + d \left[\frac{UAT(u)}{PRT(u)} \left(\sum_{v \in B(u)} \frac{EWPR_{volT}(v) W_{(v,u)}^{in} L_u}{TL(v)} \right) \right]$$

Following are the new notations used in above formula:

- $UAT(u)$ is the User Activities Time i.e. total time user actually spends on that webpage by doing activities like cursor movement with mouse, Key Press, Touch etc.
- $PRT(u)$ is the Page Reading Time i.e. the time page has been actively opened in browser tab. In more technical words, we can say when Focus is on that page.
- $EWPR_{volT}(u)$ and $EWPR_{volT}(v)$ are rank scores of page u and v respectively.
- $WPR_{vol}(u)$ and $WPR_{vol}(v)$ are rank scores of page u and v respectively.
- $W_{in}(v,u)$ is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages (i.e. outlinks) of page v .

$$\text{Also } W_{(v,u)}^{in} = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

Where I_u and I_p represent the number of inlinks of page u and page p , respectively. $R(v)$ denotes the reference page list of page v .

This algorithm considers the user's usage trends. The user's browsing behaviour can be calculated by number of hits (visits) of links [28].

T. Improved Searching with EWPRvolT

Extended Weighted Page Rank based on Visits of Links can be used at the time of preparing query results for user by search engine module. As $EWPR_{volT}$ works on the basis of user browsing trends. So, this concept is very useful to display most valuable pages on the top of the result list, which reduce the search space for user to a large extent. To support the users to navigate in the result list, our search module will use $EWPRVOL$ (SELECT column_name,column_name FROM table_name ORDER BY ewpr_vol DESC) and displays the mostly visited and

Note: For the successful execution of our proposed crawler approach, initially we need to have adequate data in our crawler database, so that crawler can get inlinks for new webpage from that database. One method to fill crawler database at the initial stage is to crawl a directory like <http://dmoztools.net>. Dmoz is a multilingual open-content directory of World Wide Web links. Alternatively, the other Method can be used to set our threshold value ($WPR_THRESHOLD$) equal to 0.15, rather than greater than 0.15.

7. Experiment and Results

The main objective of the experimental work reported in this file is to evaluate the crawling performance of a web crawler with techniques used like MD5 for a Signature generation, Relevancy Computation, Weighted Page Rank Algorithm and Extended Weighted Page Rank based on VOL Algorithm. We have implemented our research work on VPS Server with machine name crawl search.xyz and IP address (69.10.35.137), which consists of Shared Processor – Genuine Intel (Intel® Xeon® CPU

actively accessed pages on the top of result list making searching for users more enhanced in terms of quality. With our approach mentioned earlier, by not storing unimportant and irrelevant pages, also by deleting never accesses data, we will be saving a lot of memory space in database that will eventually speed up the queries (searches) from the database. Execution Time of search engine will be reduced to provide fast search results to users.

E3-1230 V5@3.40GHZ, CPU Cores : 4), RAM – 3GB, Hard Disk – 75 GB, Operating System – Linux (Centos release 7.3.1611 Core), Data Transfer Limit - 3TB, Port or Link – 100Mbps. We have implemented our crawler using technologies PHP 5.5.38 (cli) Version, MYSQL 5.6.35 Version, AJAX and APACHE Server 2.4.25 Version. We have tested our work on live websites.

U. Crawling Results for Seed <http://www.vvguptaandco.com>

1. Maximum Pages set to be crawled are 500.
2. Total Pages fetched are 500.
3. No. of Unique Pages found are 59.
4. No. of Rejected Pages on behalf of Standard Normalization of URLs are 2.
5. No. of Rejected Pages on behalf of MD5 Hashing Algorithm are 15
6. Total No. of Saved Pages on WPR Criteria is 39.

```
crawlsearchxyz@vps:~/public_html/isha
[crawlsearchxyz@vps isha]$ php -f downloader.php "http://www.vvguptaandco.com"
Could not open input file: downloader.php
[crawlsearchxyz@vps isha]$ clear
[crawlsearchxyz@vps isha]$ php -f downloader.php "http://www.vvguptaandco.com"

-----
CRAWLER STARTED AT 06/07/2017 06:52:26 am
-----

(+) Saving seed 'http://www.vvguptaandco.com' for reference
-----

Getting links for seed: http://www.vvguptaandco.com
-----

Fetching [1]: http://www.vvguptaandco.com
Fetching [2]: http://www.vvguptaandco.com/index.html
Fetching [3]: http://www.vvguptaandco.com/services.html
Fetching [4]: http://www.vvguptaandco.com/taxation.html
Fetching [5]: http://www.vvguptaandco.com/accounting.html
Fetching [6]: http://www.vvguptaandco.com/auditing.html
Fetching [7]: http://www.vvguptaandco.com/company-law-matters.html
Fetching [8]: http://www.vvguptaandco.com/project-financing.html
Fetching [9]: http://www.vvguptaandco.com/allied-services.html
Fetching [10]: http://www.vvguptaandco.com/
Fetching [11]: http://www.vvguptaandco.com/index.html
Fetching [12]: http://www.vvguptaandco.com/services.html
Fetching [13]: http://www.vvguptaandco.com/taxation.html
Fetching [14]: http://www.vvguptaandco.com/accounting.html
Fetching [15]: http://www.vvguptaandco.com/auditing.html
Fetching [16]: http://www.vvguptaandco.com/company-law-matters.html
Fetching [17]: http://www.vvguptaandco.com/projct-financing.html
Fetching [18]: http://www.vvguptaandco.com/allied-services.html
Fetching [19]: http://www.vvguptaandco.com/
Fetching [20]: http://www.mca.gov.in/MCA21/Download_eForm_choose.html
Fetching [21]: http://www.vvguptaandco.com/it-form-assessment-year-2017-18.html
Fetching [22]: http://www.vvguptaandco.com/it-form-assessment-year-2016-17.html
Fetching [23]: http://www.vvguptaandco.com/it-form-assessment-year-2015-16.html
Fetching [24]: http://www.vvguptaandco.com/it-form-assessment-year-2014-15.html
Fetching [25]: http://www.vvguptaandco.com/it-form-assessment-year-2013-14.html
Fetching [26]: http://www.vvguptaandco.com/it-form-assessment-year-2012-13.html
Fetching [27]: http://www.vvguptaandco.com/it-form-assessment-year-2011-12.html
Fetching [28]: http://aces.gov.in/download.jsp
Fetching [29]: http://www.pextax.com
```

Fig. 11. Fetching Pages for vvguptaandco.com

```
crawlsearchxyz@vps:~/public_html/isha
Fetching [497]: http://www.vvguptaandco.com/contact.html#query
Fetching [498]: http://www.sodiztechnologies.com
Fetching [499]: http://www.vvguptaandco.com/index.html
Fetching [500]: http://www.vvguptaandco.com/services.html

=> Saving raw pages for comparision in future...
=> Removing duplicate links...
=> Checking domain age...

Here is the Domain vvguptaandco.com
1) http://www.vvguptaandco.com, age=1854 days, relevancy=6
Here is the Domain vvguptaandco.com
2) http://www.vvguptaandco.com/services.html, age=1854 days, relevancy=52
Here is the Domain vvguptaandco.com
3) http://www.vvguptaandco.com/taxation.html, age=1854 days, relevancy=27
Here is the Domain vvguptaandco.com
4) http://www.vvguptaandco.com/accounting.html, age=1854 days, relevancy=34
Here is the Domain vvguptaandco.com
5) http://www.vvguptaandco.com/auditing.html, age=1854 days, relevancy=42
Here is the Domain vvguptaandco.com
6) http://www.vvguptaandco.com/company-law-matters.html, age=1854 days, relevancy=21
Here is the Domain vvguptaandco.com
7) http://www.vvguptaandco.com/project-financing.html, age=1854 days, relevancy=15
Here is the Domain vvguptaandco.com
8) http://www.vvguptaandco.com/allied-services.html, age=1854 days, relevancy=18
Here is the Domain mca.gov.in
9) http://www.mca.gov.in/MCA21/Download_eForm_choose.html, age=4338 days, relevancy=0
Here is the Domain vvguptaandco.com
10) http://www.vvguptaandco.com/it-form-assessment-year-2017-18.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
11) http://www.vvguptaandco.com/it-form-assessment-year-2016-17.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
12) http://www.vvguptaandco.com/it-form-assessment-year-2015-16.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
13) http://www.vvguptaandco.com/it-form-assessment-year-2014-15.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
14) http://www.vvguptaandco.com/it-form-assessment-year-2013-14.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
15) http://www.vvguptaandco.com/it-form-assessment-year-2012-13.html, age=1854 days, relevancy=0
Here is the Domain vvguptaandco.com
16) http://www.vvguptaandco.com/it-form-assessment-year-2011-12.html, age=1854 days, relevancy=0
```

Fig. 12. Calculating Domain age and relevancy for vvguptaandco.com

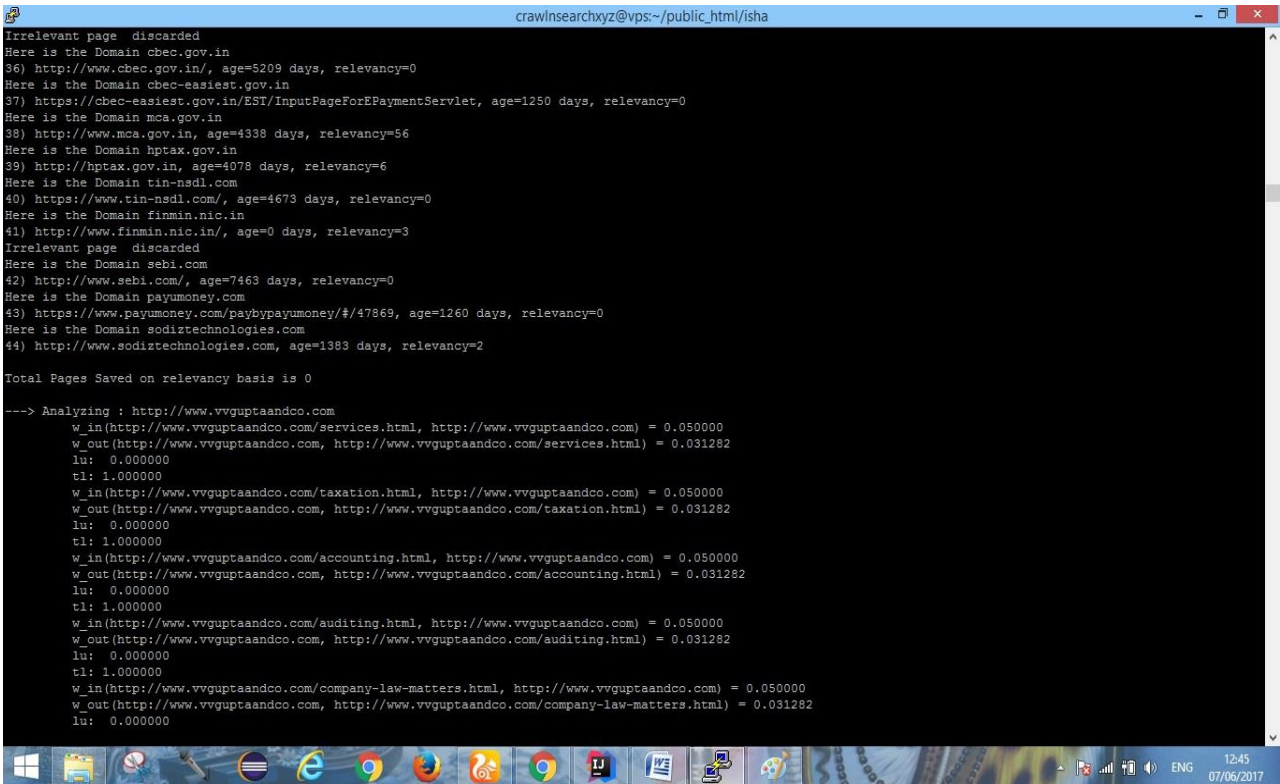


Fig. 13. Saves Relevant Pages to Database for vvguptaandco.com

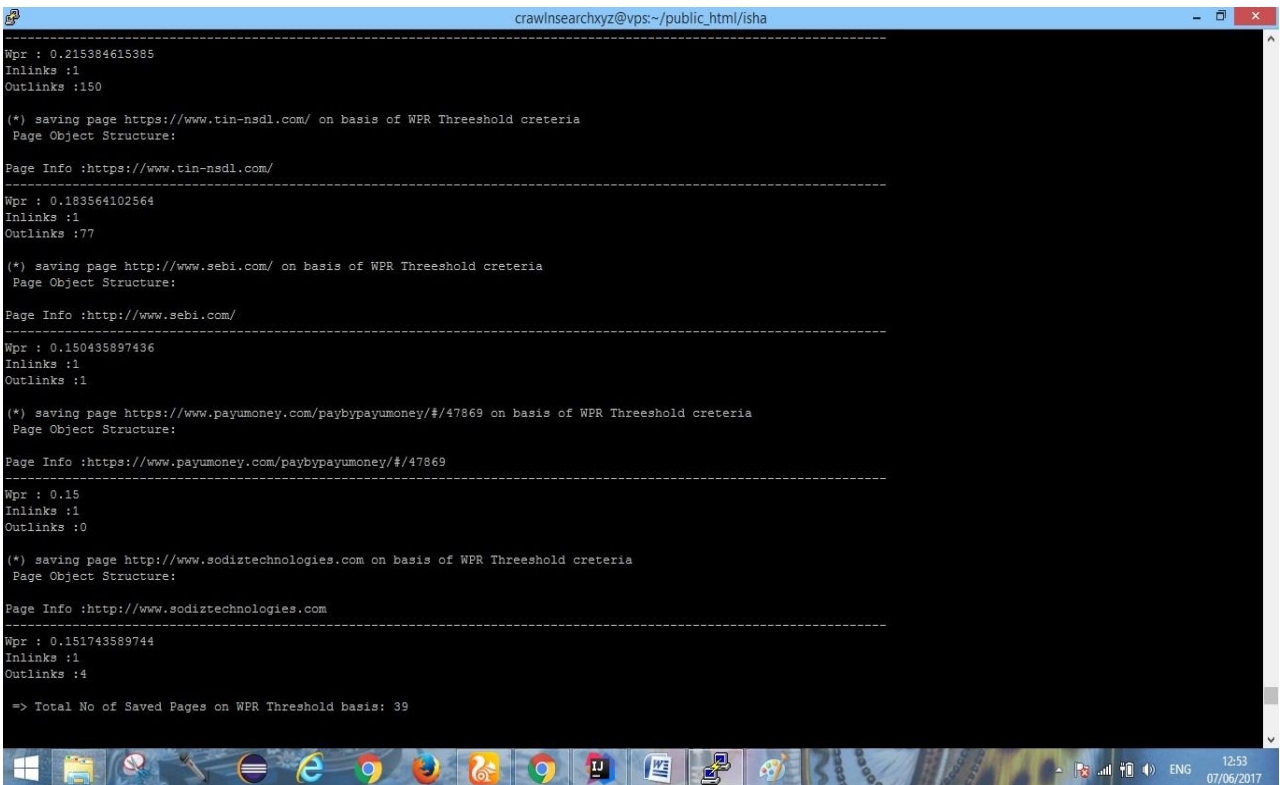


Fig. 14. Saving Pages on the basis of WPR Criteria for vvguptaandco.com

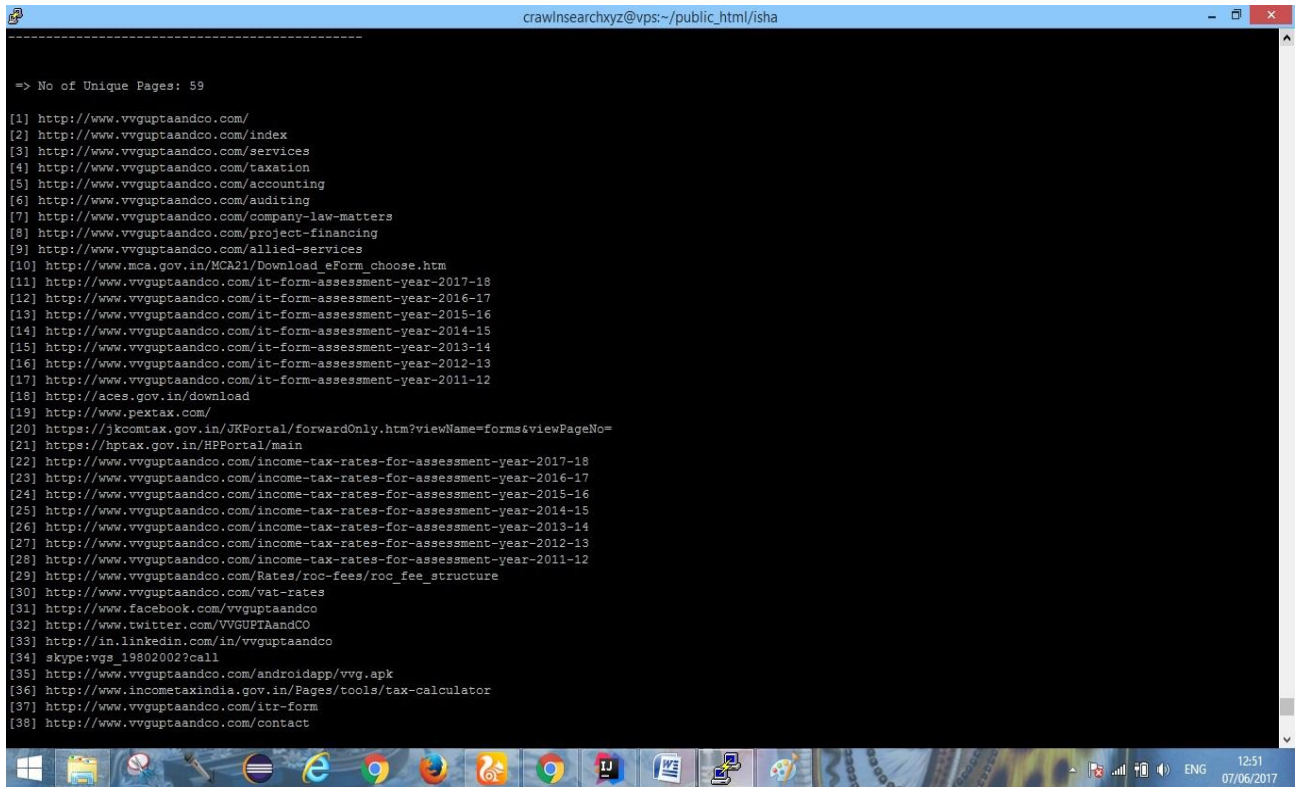


Fig. 15. No. of unique Pages found for vvguptaandco.com

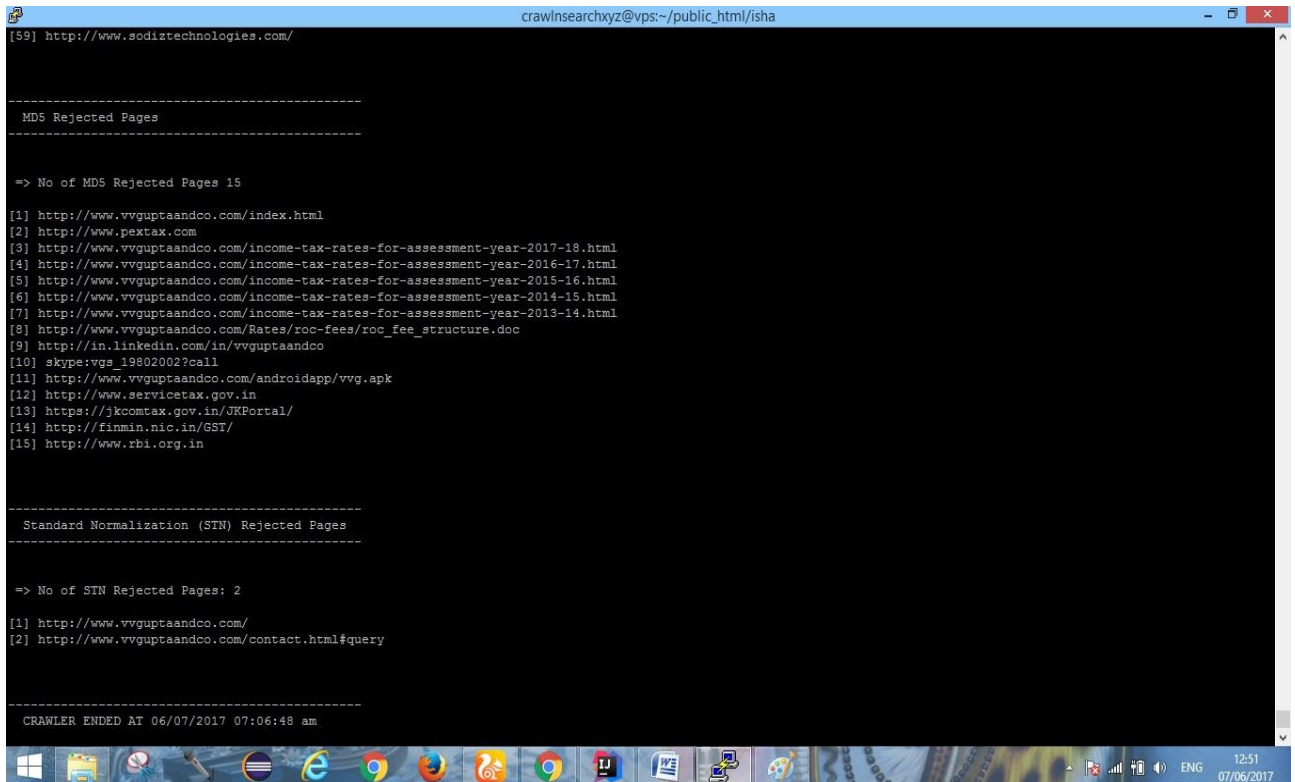
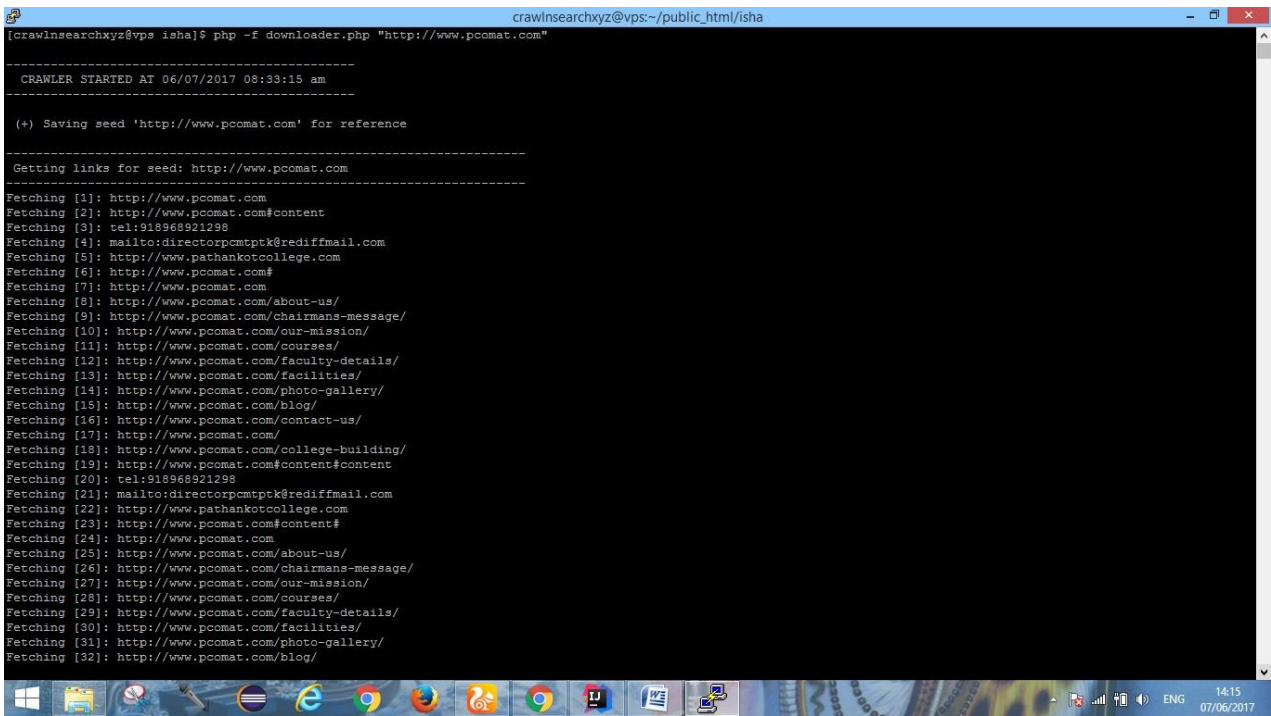


Fig. 16. No. of MD5 and STN rejected Pages for vvguptaandco.com

V. Crawling Results for Seed <http://www.pcomat.com>

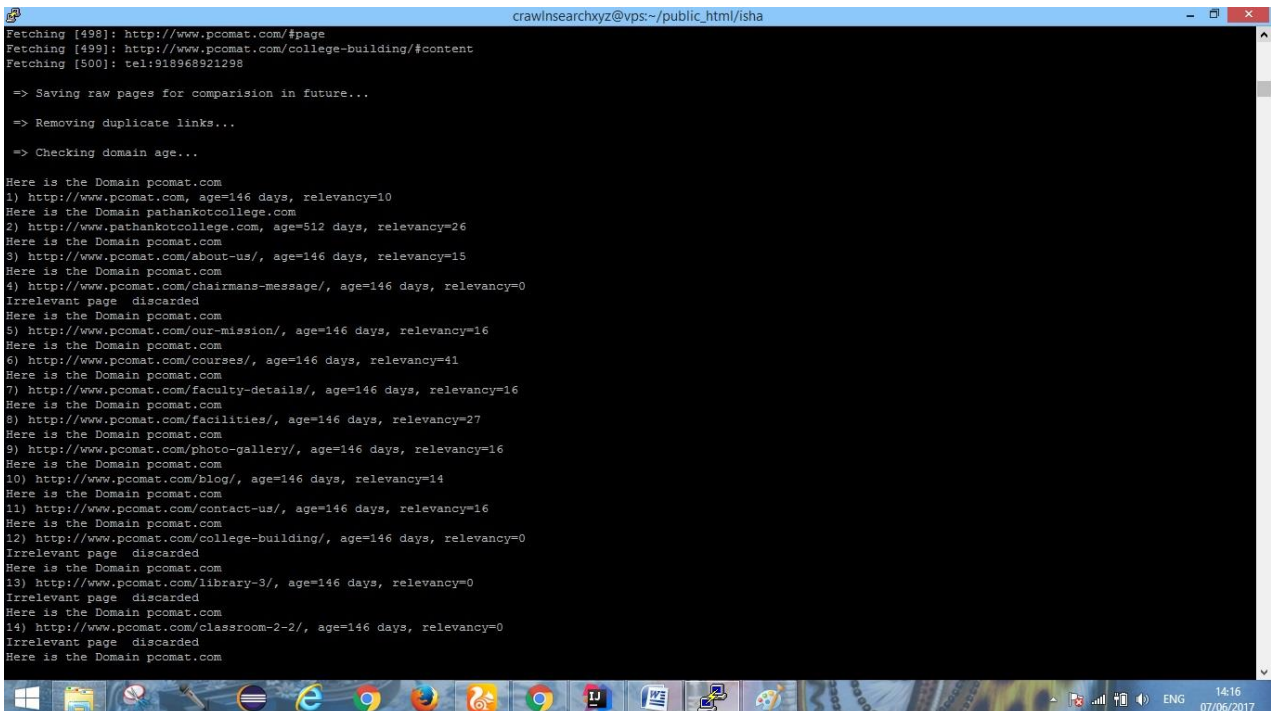
1. Maximum Pages set to be crawled are 500.
2. Total Pages Fetched are 500.
3. No. of Unique Pages found are 108.

4. No. of Rejected Pages on behalf of Standard Normalization of URLs are 20.
5. No. of Rejected Pages on behalf of MD5 Hashing Algorithm are 33.
6. Total No. of Saved Pages on Relevancy basis is 10.
7. Total No. of Saved Pages on WPR Threshold Criteria is 46.



```
crawlsearchxyz@vps:~/public_html/isha
[crawlsearchxyz@vps isha]$ php -f downloader.php "http://www.pcomat.com"
-----
CRAWLER STARTED AT 06/07/2017 08:33:15 am
-----
(+ Saving seed 'http://www.pcomat.com' for reference
-----
Getting links for seed: http://www.pcomat.com
-----
Fetching [1]: http://www.pcomat.com
Fetching [2]: http://www.pcomat.com#content
Fetching [3]: tel:918968921298
Fetching [4]: mailto:directorpcmtptk@rediffmail.com
Fetching [5]: http://www.pathankotcollege.com
Fetching [6]: http://www.pcomat.com#
Fetching [7]: http://www.pcomat.com
Fetching [8]: http://www.pcomat.com/about-us/
Fetching [9]: http://www.pcomat.com/chairmans-message/
Fetching [10]: http://www.pcomat.com/our-mission/
Fetching [11]: http://www.pcomat.com/courses/
Fetching [12]: http://www.pcomat.com/faculty-details/
Fetching [13]: http://www.pcomat.com/facilities/
Fetching [14]: http://www.pcomat.com/photo-gallery/
Fetching [15]: http://www.pcomat.com/blog/
Fetching [16]: http://www.pcomat.com/contact-us/
Fetching [17]: http://www.pcomat.com/
Fetching [18]: http://www.pcomat.com/college-building/
Fetching [19]: http://www.pcomat.com#content#content
Fetching [20]: tel:918968921298
Fetching [21]: mailto:directorpcmtptk@rediffmail.com
Fetching [22]: http://www.pcomat.com/college-building/#content
Fetching [23]: http://www.pcomat.com#content#
Fetching [24]: http://www.pcomat.com
Fetching [25]: http://www.pcomat.com/about-us/
Fetching [26]: http://www.pcomat.com/chairmans-message/
Fetching [27]: http://www.pcomat.com/our-mission/
Fetching [28]: http://www.pcomat.com/courses/
Fetching [29]: http://www.pcomat.com/faculty-details/
Fetching [30]: http://www.pcomat.com/facilities/
Fetching [31]: http://www.pcomat.com/photo-gallery/
Fetching [32]: http://www.pcomat.com/blog/
```

Fig. 17. Fetching Pages for pcomat.com



```
crawlsearchxyz@vps:~/public_html/isha
Fetching [498]: http://www.pcomat.com/#page
Fetching [499]: http://www.pcomat.com/college-building/#content
Fetching [500]: tel:918968921298
=> Saving raw pages for comparision in future...
=> Removing duplicate links...
=> Checking domain age...
Here is the Domain pcomat.com
1) http://www.pcomat.com, age=146 days, relevancy=10
Here is the Domain pathankotcollege.com
2) http://www.pathankotcollege.com, age=512 days, relevancy=26
Here is the Domain pcomat.com
3) http://www.pcomat.com/about-us/, age=146 days, relevancy=15
Here is the Domain pcomat.com
4) http://www.pcomat.com/chairmans-message/, age=146 days, relevancy=0
Irrelevant page discarded
Here is the Domain pcomat.com
5) http://www.pcomat.com/our-mission/, age=146 days, relevancy=16
Here is the Domain pcomat.com
6) http://www.pcomat.com/courses/, age=146 days, relevancy=41
Here is the Domain pcomat.com
7) http://www.pcomat.com/faculty-details/, age=146 days, relevancy=16
Here is the Domain pcomat.com
8) http://www.pcomat.com/facilities/, age=146 days, relevancy=27
Here is the Domain pcomat.com
9) http://www.pcomat.com/photo-gallery/, age=146 days, relevancy=16
Here is the Domain pcomat.com
10) http://www.pcomat.com/blog/, age=146 days, relevancy=14
Here is the Domain pcomat.com
11) http://www.pcomat.com/contact-us/, age=146 days, relevancy=16
Here is the Domain pcomat.com
12) http://www.pcomat.com/college-building/, age=146 days, relevancy=0
Irrelevant page discarded
Here is the Domain pcomat.com
13) http://www.pcomat.com/library-3/, age=146 days, relevancy=0
Irrelevant page discarded
Here is the Domain pcomat.com
14) http://www.pcomat.com/classroom-2-2/, age=146 days, relevancy=0
Irrelevant page discarded
Here is the Domain pcomat.com
```

Fig. 18. Calculating Domain age and relevancy for pcomat.com

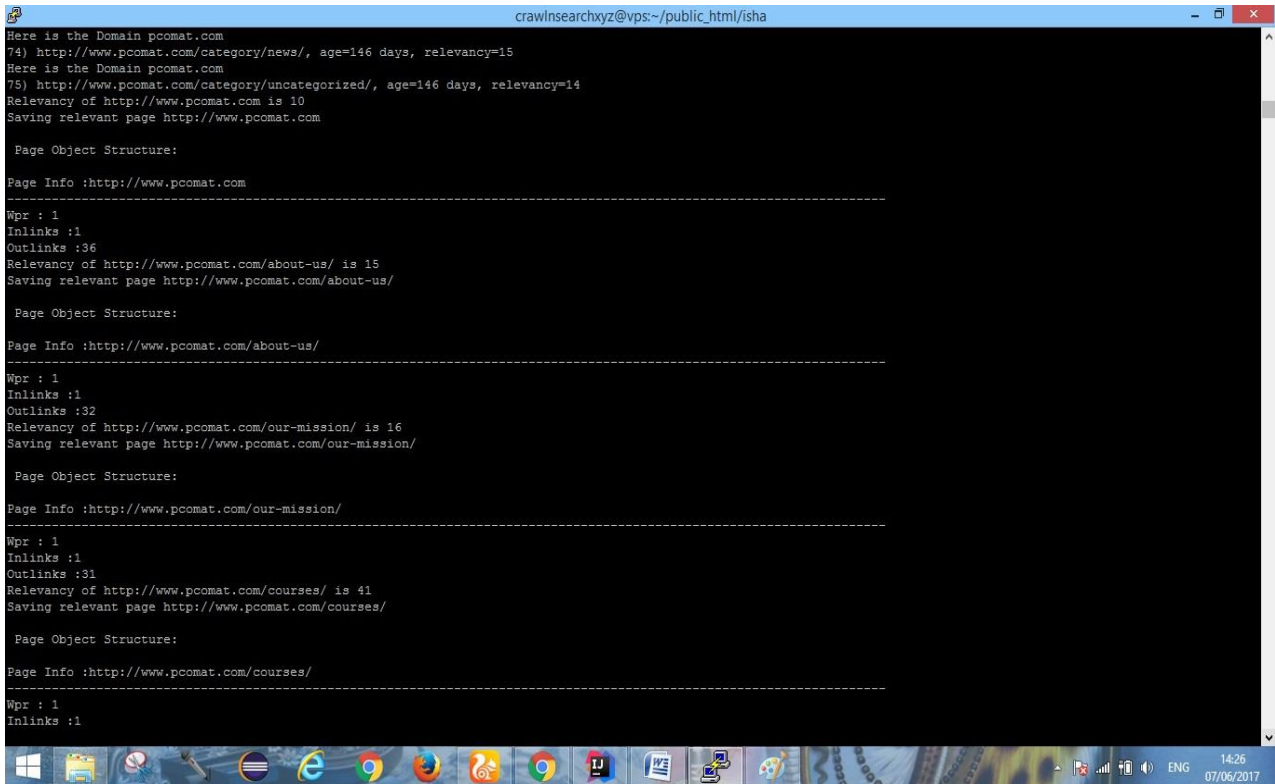


Fig. 19. Saves Relevant Pages to Database for pcomat.com

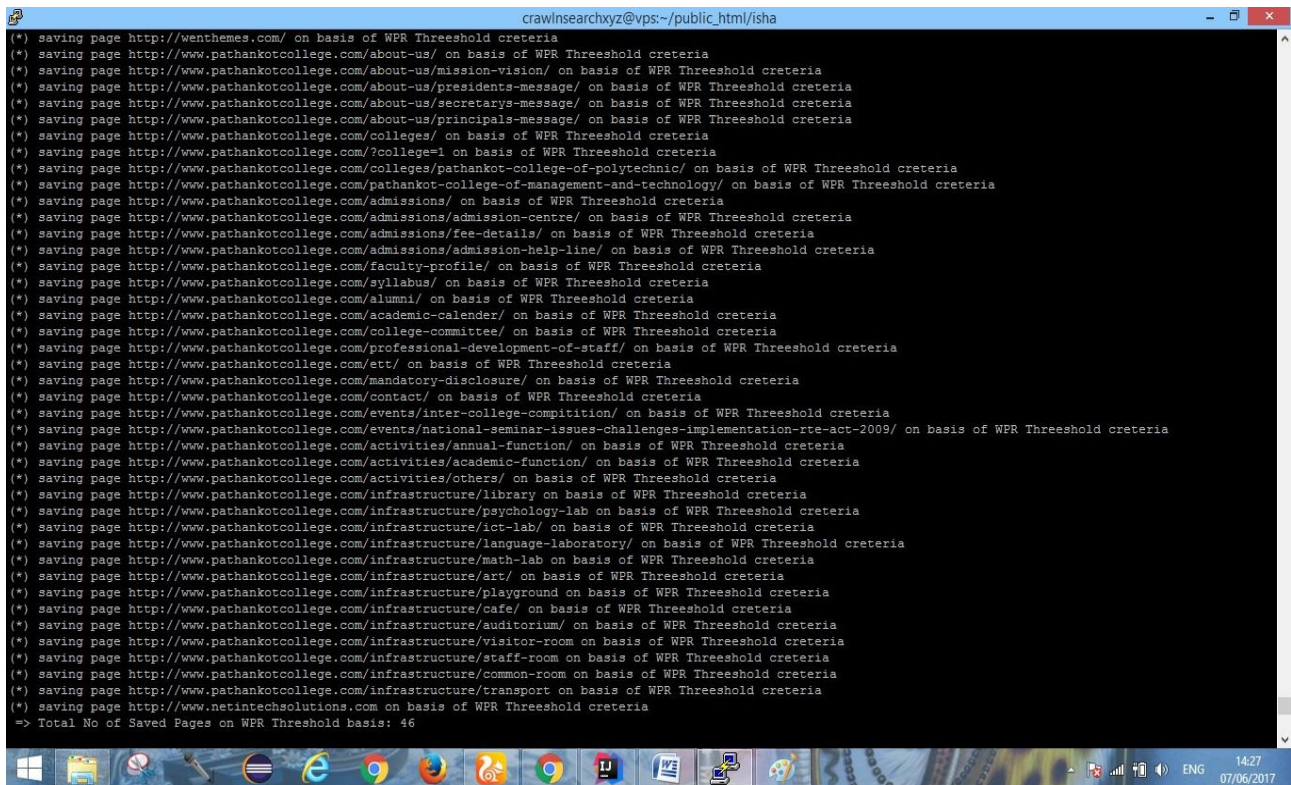


Fig. 20. Saving Pages on the basis of WPR Criteria for pcomat.com


```
-----  
crawInsearchxyz@vps:~/public_html/isha  
-----  
Unique Pages found in this crawl  
-----  
  
=> No of Unique Pages: 108  
  
[1] http://www.pcomat.com/  
[2] http://www.pcomat.com/#content  
[3] tel:918968921298  
[4] mailto:directorpcmtptk@rediffmail.com  
[5] http://www.pathankotcollege.com/  
[6] http://www.pcomat.com/about-us/  
[7] http://www.pcomat.com/chairmans-message/  
[8] http://www.pcomat.com/our-mission/  
[9] http://www.pcomat.com/courses/  
[10] http://www.pcomat.com/faculty-details/  
[11] http://www.pcomat.com/facilities/  
[12] http://www.pcomat.com/photo-gallery/  
[13] http://www.pcomat.com/blog/  
[14] http://www.pcomat.com/contact-us/  
[15] http://www.pcomat.com/college-building/  
[16] http://www.pcomat.com/#content%23content  
[17] http://www.pcomat.com/#content%23  
[18] http://www.pcomat.com/library-3/  
[19] http://www.pcomat.com/classroom-2-2/  
[20] http://www.pcomat.com/physics-and-chemistry-laboratory/  
[21] http://www.pcomat.com/indian-republic-day-celebration-held-on-26-january-2017/  
[22] http://www.pcomat.com/voter-day-celebration-at-pathankot-college-of-management-and-technology-25-january-2017/  
[23] http://www.pcomat.com/mst-will-be-held-on-05-february-2017/  
[24] http://www.pcomat.com/inter-college-competition-on-10-feb-2017/  
[25] http://www.infopathankot.com/  
[26] http://www.facebook.com/  
[27] http://www.twitter.com/  
[28] http://www.linkedin.com/  
[29] https://  
[30] http://www.youtube.com/  
[31] http://www.pinterest.com/  
[32] http://www.pcomat.com/visit-our-other-website/  
[33] https://maps.google.com/maps?z=16&#038%3Bq%3Dpathankot%2Bcollege%2Bof%2Bmanagement%2Band%2Btechnology%2Bopposite%2Bcanada%2Bpalace%2Bjalandhar%2B-%2Bdalhousie%2Bby  
pass%2C%2Bmanun%2Bpathankot%2B-%2B145001%2Bpunjab%2BIndia  
[34] https://wordpress.org/  
[35] http://wenthemes.com/
```

Fig. 21. No. of unique Pages found for pcomat.com

```
-----  
crawInsearchxyz@vps:~/public_html/isha  
-----  
[107] http://www.pcomat.com/contact-us/#page  
[108] http://www.pcomat.com/college-building/#content  
-----  
MDS Rejected Pages  
-----  
  
=> No of MDS Rejected Pages 33  
  
[1] http://www.pcomat.com#content  
[2] tel:918968921298  
[3] mailto:directorpcmtptk@rediffmail.com  
[4] http://www.pcomat.com#content#content  
[5] http://www.pcomat.com#content#  
[6] http://www.pcomat.com#content#page  
[7] http://www.pathankotcollege.com/about-us/about-society/  
[8] http://www.pathankotcollege.com/wp-content/uploads/2016/08/B-ED-TWO-YEARS.pdf  
[9] http://www.pcomat.com#content  
[10] http://www.pcomat.com##  
[11] http://www.pcomat.com##page  
[12] http://www.pcomat.com#page  
[13] http://www.pcomat.com/about-us/#content  
[14] http://www.pcomat.com/about-us/#page  
[15] http://www.pcomat.com/chairmans-message/#content  
[16] http://www.pcomat.com/chairmans-message/#page  
[17] http://www.pcomat.com/our-mission/#content  
[18] http://www.pcomat.com/our-mission/#page  
[19] http://www.pcomat.com/courses/#content  
[20] http://www.pcomat.com/courses/#page  
[21] http://www.pcomat.com/faculty-details/#content  
[22] http://www.pcomat.com/faculty-details/#page  
[23] http://www.pcomat.com/facilities/#content  
[24] http://www.pcomat.com/facilities/#page  
[25] http://www.pcomat.com/photo-gallery/#content  
[26] https://i0.wp.com/www.pcomat.com/wp-content/uploads/2017/01/classroom.jpg?fit=900%2C600  
[27] https://i1.wp.com/www.pcomat.com/wp-content/uploads/2017/01/labwithstudents.jpg?fit=900%2C600  
[28] http://www.pcomat.com/photo-gallery/#page  
[29] http://www.pcomat.com/blog/#content  
[30] http://www.pcomat.com/blog/#page  
[31] http://www.pcomat.com/contact-us/#content
```

Fig. 22. No. of MD5 rejected Pages for pcomat.com

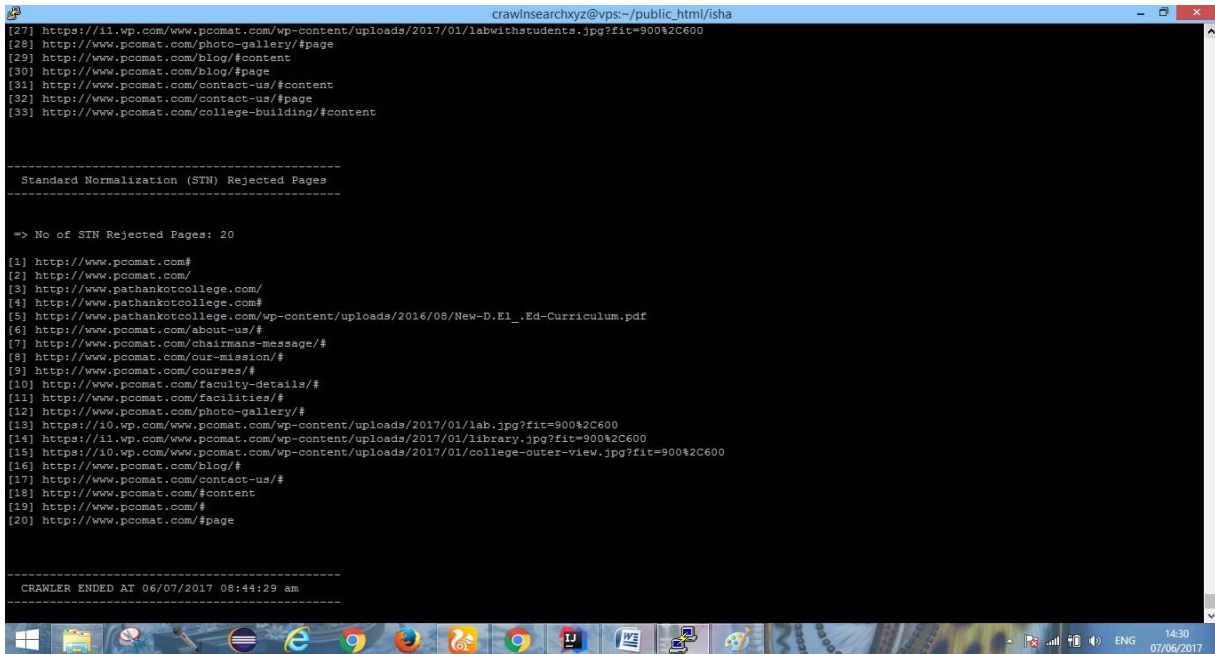


Fig. 23. No. of STN rejected Pages for pcomat.com

W. Crawling Results for Seed <http://www.infopathankot.com>

1. Maximum Pages set to be crawled are 500.
2. Total Pages Fetched are 500.
3. No. of Unique Pages found are 65.
4. No. of Rejected Pages on behalf of Standard Normalization of URLs are 11.
5. No. of Rejected Pages on behalf of MD5 are 8.
6. Total No. of Saved Pages on Relevancy basis is 1.
7. Total No. of Saved Pages on WPR Threshold Criteria is 56.

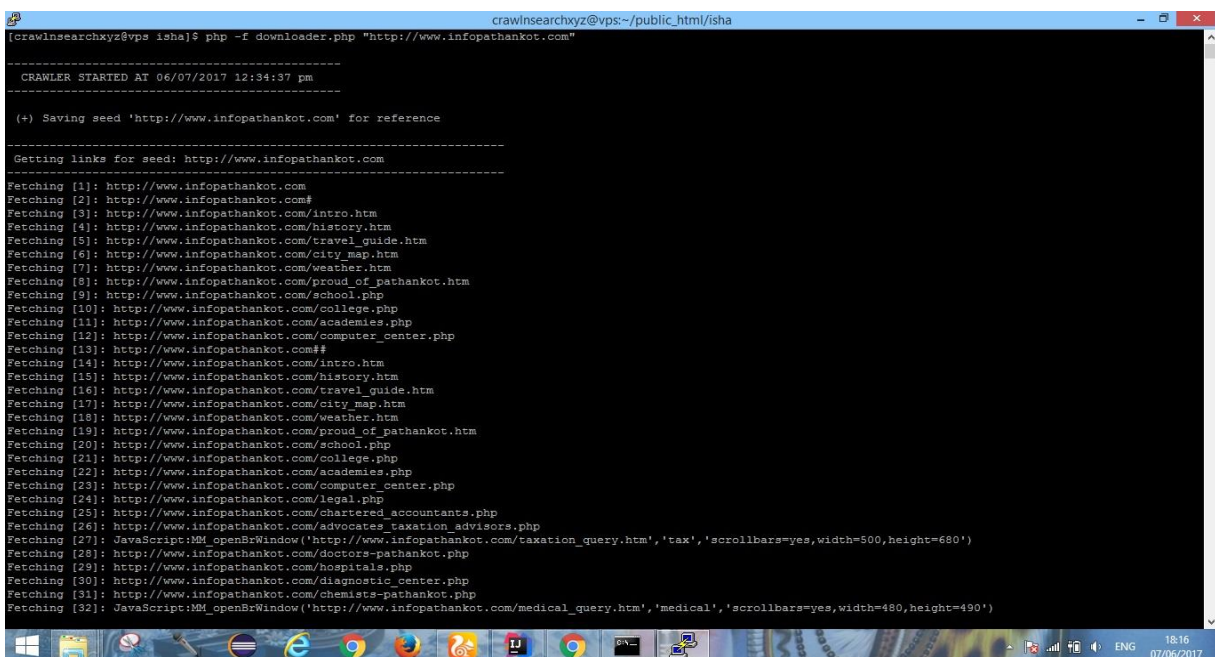


Fig. 24. Fetching Pages for infopathankot.com

```
crawlsearchxyz@vps:~/public_html/isha
Fetching [495]: http://www.infopathankot.com/administrative.htm
Fetching [496]: http://www.infopathankot.com/miscellaneous.htm
Fetching [497]: http://www.infopathankot.com/property_dealers.php
Fetching [498]: http://www.infopathankot.com/tour_travels.php
Fetching [499]: http://www.pincode.in/search.php
Fetching [500]: http://www.infopathankot.com/photo_gallery.htm

=> Saving raw pages for comparision in future...

=> Removing duplicate links...

=> Checking domain age...

Here is the Domain infopathankot.com
1) http://www.infopathankot.com, age=3850 days, relevancy=0
Here is the Domain infopathankot.com
2) http://www.infopathankot.com/intro.htm, age=3850 days, relevancy=39
Here is the Domain infopathankot.com
3) http://www.infopathankot.com/history.htm, age=3850 days, relevancy=21
Here is the Domain infopathankot.com
4) http://www.infopathankot.com/travel_guide.htm, age=3850 days, relevancy=17
Here is the Domain infopathankot.com
5) http://www.infopathankot.com/city_map.htm, age=3850 days, relevancy=8
Here is the Domain infopathankot.com
6) http://www.infopathankot.com/weather.htm, age=3850 days, relevancy=41
Here is the Domain infopathankot.com
7) http://www.infopathankot.com/school.php, age=3850 days, relevancy=175
Here is the Domain infopathankot.com
8) http://www.infopathankot.com/college.php, age=3850 days, relevancy=64
Here is the Domain infopathankot.com
9) http://www.infopathankot.com/academies.php, age=3850 days, relevancy=21
Here is the Domain infopathankot.com
10) http://www.infopathankot.com/computer_center.php, age=3850 days, relevancy=17
Here is the Domain infopathankot.com
11) http://www.infopathankot.com/legal.php, age=3850 days, relevancy=24
Here is the Domain infopathankot.com
12) http://www.infopathankot.com/chartered_accountants.php, age=3850 days, relevancy=14
Here is the Domain infopathankot.com
13) http://www.infopathankot.com/advocates_taxation_advisors.php, age=3850 days, relevancy=0
Here is the Domain infopathankot.com
14) http://www.infopathankot.com/doctors-pathankot.php, age=3850 days, relevancy=0
Here is the Domain infopathankot.com
15) http://www.infopathankot.com/hospitals.php, age=3850 days, relevancy=18
```

Fig. 25. Calculating Domain age and Relevancy for infopathankot.com

```
crawlsearchxyz@vps:~/public_html/isha
(*) saving page http://www.infopathankot.com/diagnostic_center.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/chemists-pathankot.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/bank.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/pathankot_web_services.php on basis of WPR Threshold criteria
(*) saving page http://www.world-airport-codes.com/India/pathankot-5704.html on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/pathankot_station.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/chakki_bank.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/bus_services.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/accountants.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/marriage_places.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/cinema.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/cyber_cafe.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/gas_agencies.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/guest_houses.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/hotels.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/health_club.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/administrative.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/miscellaneous.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/property_dealers.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/tour_travels.php on basis of WPR Threshold criteria
(*) saving page http://www.pincode.in/search.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/photo_gallery.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/contact_us.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/music.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/online_games.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/jokes.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/sms.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/wallpapers.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/astrology.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/ebooks.php on basis of WPR Threshold criteria
(*) saving page http://www.himachal-touristguide.com on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/site_map.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/advertise_with_us.htm on basis of WPR Threshold criteria
(*) saving page http://www.hostingaces.com on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/how_to_reach.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/distance.php on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/taxation_query.htm on basis of WPR Threshold criteria
(*) saving page http://www.infopathankot.com/medical_query.htm on basis of WPR Threshold criteria
(*) saving page http://www.accuweather.com/us///-999/city-weather-forecast.asp?partner=accuweather&traveler=0 on basis of WPR Threshold criteria
(*) saving page http://www.accuweather.com/maps-satellite.asp on basis of WPR Threshold criteria
(*) saving page http://www.accuweather.com/index-radar.asp?partner=accuweather&traveler=0&zipcode=ASI|IN|IN028|PATHANKOT| on basis of WPR Threshold criteria
=> Total No of Saved Pages on WPR Threshold basis: 56
```

Fig. 26. Saving pages on WPR basis for infopathankot.com

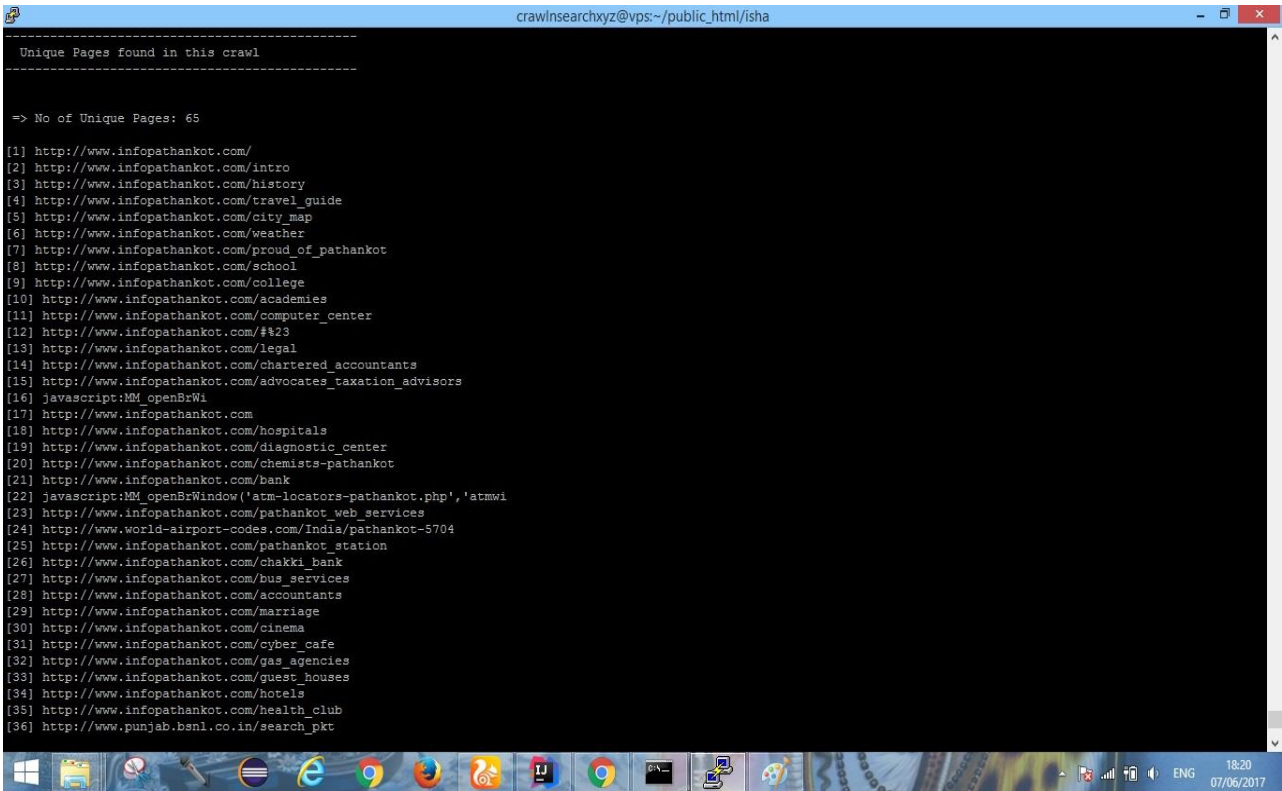


Fig. 27. Unique pages for infopathankot.com

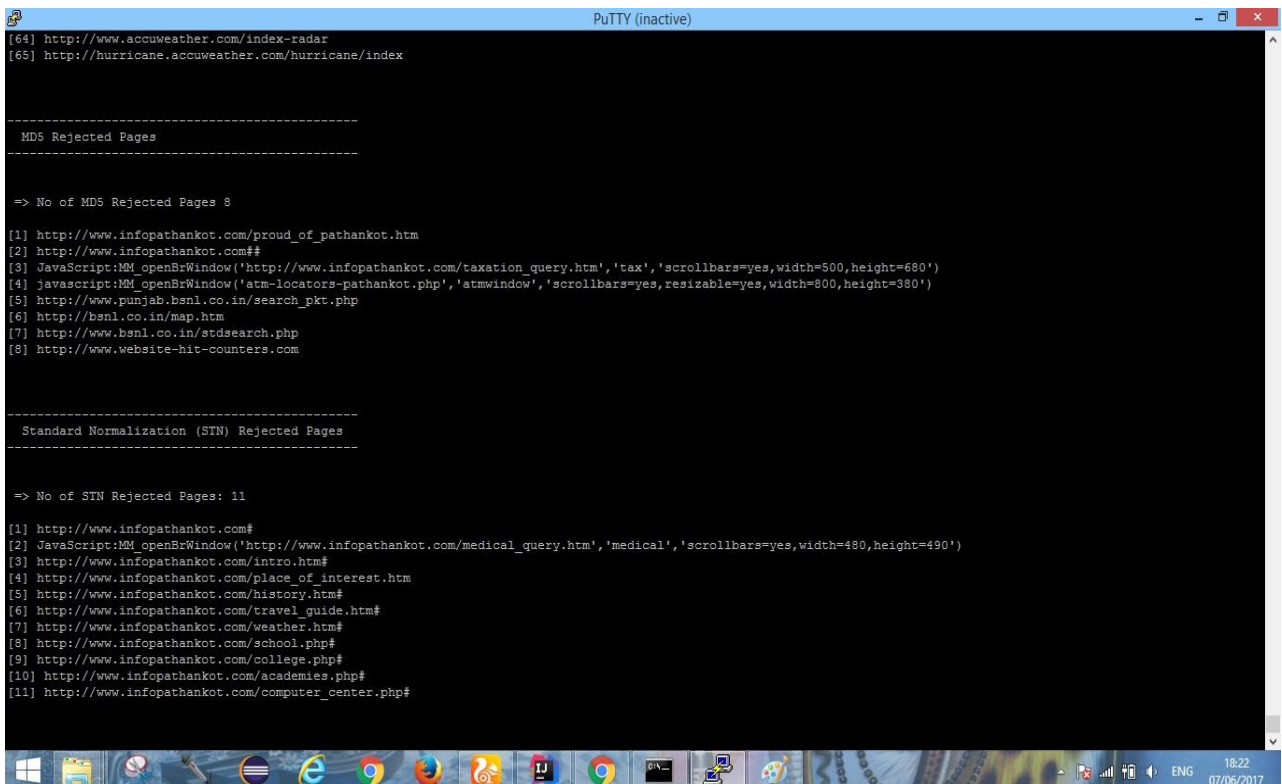


Fig. 28. MD5 and STN rejected pages for infopathankot.com

X. Snapshot of saved Visits of Links, User Attention Time and Page Reading Time

These are details stored on our server of visits of links, user attention time (UAT), page reading time (PRT) of various web pages from various web resources, who have added our Extended Vol-Analytics Script to their web pages. Our Ajax based Script sent details of webpage like its address, its caller address, its page reading time

(page_focus_time), its user attention time (exact_time) and its website id to our server where these details got tracked and stored in our database. Here is a snapshot of above-mentioned details.

| id | url | caller_url | call_count | web_id | datetime | page_focus_time | exact_time |
|-----|---|---|------------|--------|---------------------|---------------------|---------------------|
| 88 | http://www.skytravlen.com/enquiry | http://www.skytravlen.com/hotel-bookings | 5 | 7 | 2017-05-24 18:55:50 | 2829.335000000187 | 2170.664999999981 |
| 97 | http://www.vvguptaandco.com/itr-form | http://www.vvguptaandco.com/contact | 2 | 3 | 2017-05-25 16:56:22 | 54968.48999999929 | 24968.489999999292 |
| 98 | http://www.vvguptaandco.com/itr-form-assessment-ya... | http://www.vvguptaandco.com/itr-form | 7 | 3 | 2017-05-25 16:57:46 | 27112.91000000001 | 17112.91000000001 |
| 99 | http://www.vvguptaandco.com/itr-form-assessment-ya... | http://www.vvguptaandco.com/itr-form-assessment-ya... | 17 | 3 | 2017-05-25 16:58:01 | 12549.539999999999 | 12549.539999999999 |
| 100 | http://www.vvguptaandco.com/auditing | http://www.vvguptaandco.com/itr-form-assessment-ya... | 14 | 3 | 2017-05-25 17:01:12 | 7843.610000000017 | 2156.389999999983 |
| 101 | http://www.vvguptaandco.com/project-financing | http://www.vvguptaandco.com/auditing | 6 | 3 | 2017-05-25 17:01:32 | 5453.405000000001 | 5453.405000000001 |
| 102 | http://www.vvguptaandco.com/auditing | http://www.vvguptaandco.com/project-financing | 10 | 3 | 2017-05-25 17:03:10 | 13745.945 | 3745.944999999997 |
| 103 | http://www.vvguptaandco.com/accounting | http://www.vvguptaandco.com/auditing | 2 | 3 | 2017-05-25 17:05:08 | 14926.420000000007 | 5073.579999999993 |
| 104 | http://www.vvguptaandco.com/taxation | http://www.vvguptaandco.com/accounting | 9 | 3 | 2017-05-25 17:09:31 | 29435.735 | 29435.735 |
| 105 | http://www.vvguptaandco.com/accounting | http://www.vvguptaandco.com/taxation | 2 | 3 | 2017-05-25 17:11:31 | 30454.04500000005 | 20464.04500000005 |
| 106 | http://www.vvguptaandco.com/company-law-matters | http://www.vvguptaandco.com/accounting | 18 | 3 | 2017-05-25 17:11:36 | 2971.850000000004 | 2971.850000000004 |
| 107 | http://www.vvguptaandco.com/taxation | http://www.vvguptaandco.com/company-law-matters | 6 | 3 | 2017-05-25 17:11:48 | 10533.585 | 10533.585 |
| 109 | http://www.skytravlen.com/flight-bookings | http://www.skytravlen.com/sky-travlen-company | 8 | 7 | 2017-05-25 17:17:40 | 6366.910000000001 | 6366.910000000001 |
| 110 | http://www.skytravlen.com/hotel-bookings | http://www.skytravlen.com/flight-bookings | 13 | 7 | 2017-05-25 17:17:45 | 4249.605 | 4249.605 |
| 111 | http://chatpatrecipes.com/ | https://www.google.co.in/ | 9 | 4 | 2017-05-25 17:28:14 | 15733.040000000068 | 5733.040000000068 |
| 112 | http://www.chatpatrecipes.com/vegetable-recipes/fi... | http://chatpatrecipes.com/ | 12 | 4 | 2017-05-25 17:28:27 | 14918.695000000003 | 14918.695000000003 |
| 113 | http://www.drmoehnderdentalclinic.com/ | https://www.google.co.in/ | 4 | 5 | 2017-05-25 17:34:35 | 38622.015000000001 | 38622.015000000001 |
| 114 | http://www.drmoehnderdentalclinic.com/our-profile | http://www.drmoehnderdentalclinic.com/ | 3 | 5 | 2017-05-25 17:34:51 | 18979.915000000005 | 18979.915000000005 |
| 115 | http://www.himachaltouristguide.com/index | http://www.himachaltouristguide.com/index.php/dist... | 39 | 2 | 2017-05-25 17:46:01 | 231958.880000000003 | 181958.880000000003 |
| 116 | http://www.himachaltouristguide.com/index | https://www.google.co.in/ | 881 | 2 | 2017-05-25 17:52:29 | 58443833.815 | 58333833.815 |
| 117 | http://www.manimaheshyatra.co.in/ | https://www.google.co.in/ | 24 | 9 | 2017-05-25 17:54:08 | 28956.004999999997 | 18956.004999999997 |
| 118 | http://www.himachaltouristguide.com/index.php/kull... | android-app://com.google.android.googlequicksearch... | 14 | 2 | 2017-05-25 17:54:13 | 25975.675000000007 | 15975.675000000007 |
| 119 | http://www.manimaheshyatra.co.in/legends-about-man... | http://www.manimaheshyatra.co.in/ | 16 | 9 | 2017-05-25 17:54:26 | 12665.060000000001 | 12665.060000000001 |
| 120 | http://www.manimaheshyatra.co.in/lord-shiva-11-rud... | http://www.manimaheshyatra.co.in/legends-about-man... | 31 | 9 | 2017-05-25 17:54:38 | 7489.69 | 7489.69 |
| 121 | http://www.manimaheshyatra.co.in/manimahesh-route | http://www.manimaheshyatra.co.in/lord-shiva-11-rud... | 19 | 9 | 2017-05-25 17:55:35 | 14414.7 | 5585.299999999999 |

Fig. 29. Visits of Links, UAT and PRT data on crawl search.xyz server

Y. Snapshot of EWPRvolT Scores of pages

As per our proposed approach, after 45 days, our server will execute a con job, which will calculate the EWPR_{volT} score of the stored pages in the database (i.e. crawled data). If EWPR_{volT} score of any page is less than the threshold (i.e. EWPR_VOL_THRESHOLD), then that page will be deleted from the database. Pages with an EWPR_{volT} score less than

EWPR_VOL_THRESHOLD are those pages that are never browsed (accessed) or not actively used. Deletion of such unused data is an optimization to that crawled data.

Before the execution of our Script, no. of saved pages in the database are 579. After the Script execution browsed pages or less actively browsed pages deleted are 64. Total Pages left in the database are 515.


```

crawlsearchxyz@vps:~/public_html/isha
[crawlsearchxyz@vps isha]$ php -f ewpr_vol.php
-----
Extended WPR VOL ENHANCEMENT STARTED AT 06/13/2017 10:50:04 am
-----

check page: http://www.vvguptaandco.com
check page: http://www.vvguptaandco.com/services.html
check page: http://www.vvguptaandco.com/taxation.html
check page: http://www.vvguptaandco.com/accounting.html
check page: http://www.vvguptaandco.com/auditing.html
check page: http://www.vvguptaandco.com/company-law-matters.html
check page: http://www.vvguptaandco.com/project-financing.html
check page: http://www.vvguptaandco.com/allied-services.html
check page: http://www.mca.gov.in/MCA21/Download_eForm_choose.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2017-18.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2016-17.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2015-16.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2014-15.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2013-14.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2012-13.html
check page: http://www.vvguptaandco.com/it-form-assessment-year-2011-12.html
check page: http://www.vvguptaandco.com/income-tax-rates-for-assessment-year-2012-13.html
check page: http://www.vvguptaandco.com/income-tax-rates-for-assessment-year-2011-12.html
check page: http://www.vvguptaandco.com/vat-rates.html
check page: http://www.vvguptaandco.com/itr-form.html
check page: http://www.vvguptaandco.com/contact.html
check page: https://incometaxindiaefiling.gov.in/e-Filing/Services/LinkAdhaarHome.html
check page: http://www.lcai.org
check page: http://www.lcsi.edu/
check page: http://www.mycwai.com/
check page: http://www.incometaxindiaefiling.gov.in
check page: http://www.cbec.gov.in/
check page: https://www.tin-nsdl.com/
check page: http://www.sebi.com/
check page: https://www.payumoney.com/paybypayumoney/#/47869
check page: http://www.sodiztechnologies.com
check page: http://www.pcomat.com
check page: http://www.pcomat.com/about-us/
check page: http://www.pcomat.com/our-mission/
check page: http://www.pcomat.com/courses/
check page: http://www.pcomat.com/faculty-details/
check page: http://www.pcomat.com/facilities/
    
```

Fig. 30. Extended VOL Enhancement started

```

crawlsearchxyz@vps:~/public_html/isha
check page: http://www.manimaheshyatra.co.in/what-to-offer-shivling-to-worship-lord-shiva/?share=google-plus-1
check page: http://pinterest.com/pin/create/button/?url=http%3A%2F%2Fwww.manimaheshyatra.co.in%2Fwhat-to-offer-shivling-to-worship-lord-shiva%2F&media=http%3A%2F%2Fwww.manimaheshyatra.co.in%2Fwp-content%2Fuploads%2F2016%2F06%2Fmanimahesh_shivai1.jpg
cnt=579
---> Analyzing : http://www.vvguptaandco.com
EWPR VOL . . . . : 0.402404
---> Analyzing : http://www.vvguptaandco.com/services.html
EWPR VOL . . . . : 0.510815
---> Analyzing : http://www.vvguptaandco.com/taxation.html
EWPR VOL . . . . : 0.380208
---> Analyzing : http://www.vvguptaandco.com/accounting.html
EWPR VOL . . . . : 0.414602
---> Analyzing : http://www.vvguptaandco.com/auditing.html
EWPR VOL . . . . : 0.265650
---> Analyzing : http://www.vvguptaandco.com/company-law-matters.html
EWPR VOL . . . . : 0.495312
---> Analyzing : http://www.vvguptaandco.com/project-financing.html
EWPR VOL . . . . : 0.419512
---> Analyzing : http://www.vvguptaandco.com/allied-services.html
EWPR VOL . . . . : 0.514286
---> Analyzing : http://www.mca.gov.in/MCA21/Download_eForm_choose.html
EWPR VOL . . . . : 0.452222
---> Analyzing : http://www.vvguptaandco.com/it-form-assessment-year-2017-18.html
    
```

Fig. 31. EWPR_{volT} Score Computation

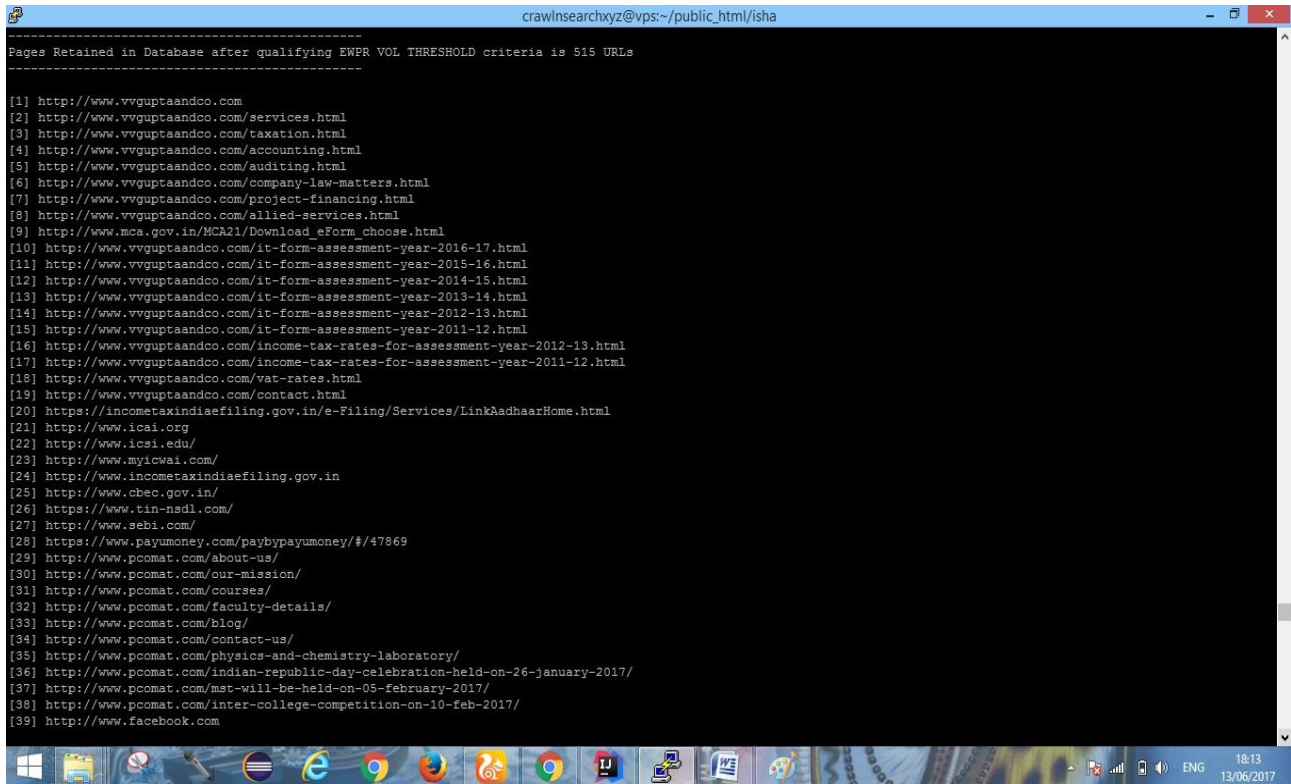


Fig. 32. Pages retained in database after EWPR_{volT} Threshold check

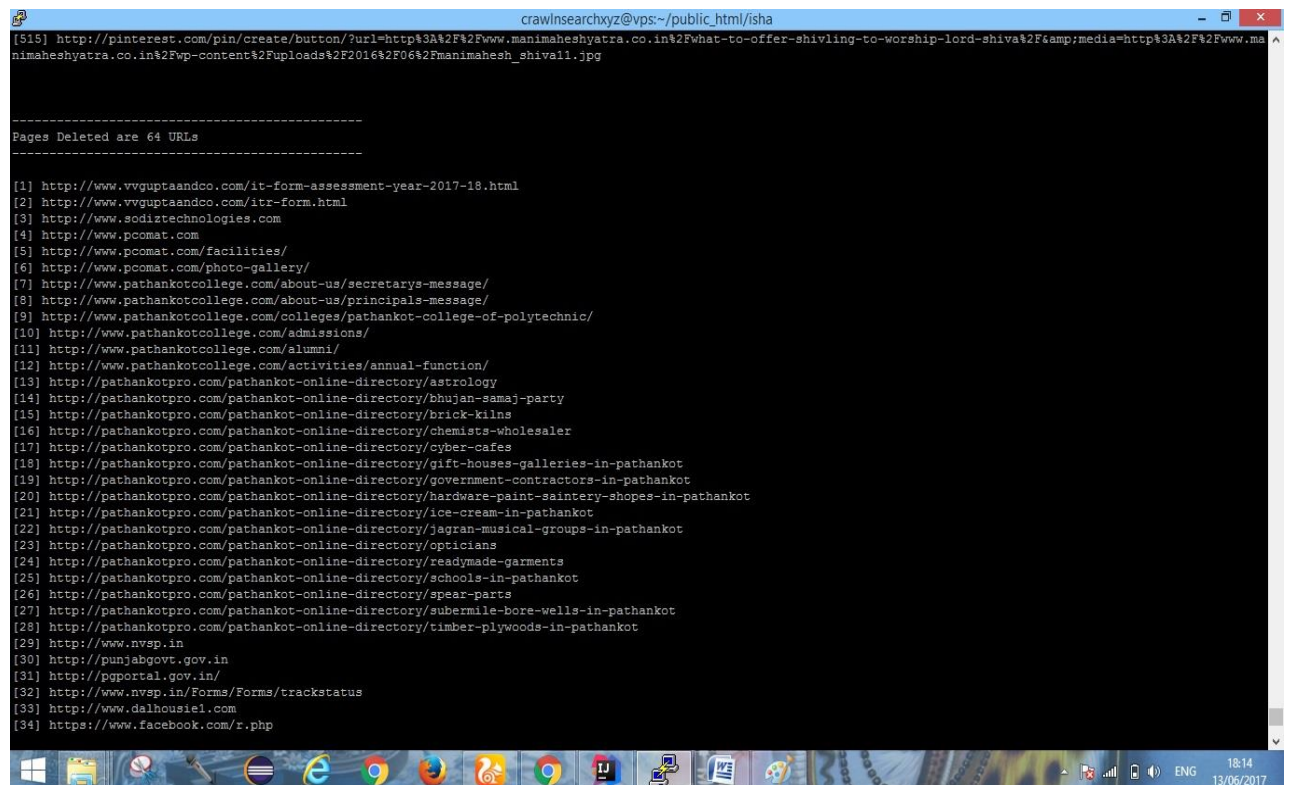


Fig. 33. Pages deleted in database after EWPR_{volT} Threshold check

Z. Snapshot of Reduced Execution time while searching

Following is a snapshot of the reduced execution time of search engine, to provide fast search results to

users. Our proposed crawling and searching approach when implemented together, takes 0.0364508628845 seconds which is less than the 0.0863118171692 seconds of the general search engine.

TABLE III. CRAWLED DATA OPTIMIZATION RESULTS OF WPR_{VOL}

| Pages in Database | Pages Retained | Pages Deleted |
|-------------------|----------------|---------------|
| 579 | 523 | 56 |
| 1000 | 917 | 83 |
| 1500 | 1382 | 128 |

TABLE IV. CRAWLED DATA OPTIMIZATION RESULTS OF EWPR_{VOL,T}

| Pages in Database | Pages Retained | Pages Deleted |
|-------------------|----------------|---------------|
| 579 | 515 | 64 |
| 1000 | 905 | 95 |
| 1500 | 1353 | 147 |

Conclusion

The Internet is one of the easiest sources available in present days for searching and accessing any data from the entire world. The structure of the World Wide Web is a graphical structure, and the links given in a page can be used to open other web pages. Web crawlers are the programs or software that uses the graphical structure of the Web to move from page to page. Here, we have briefly discussed crawlers and algorithms to enhance the quality of crawling process by crawling important and relevant pages. The crawler is an important module of a search engine. The quality of a crawler directly affects the searching quality of search engines. An another approach has been suggested in this paper, to optimize the crawled data by deleting the never browsed or least actively browsed pages data from those crawled data from search engine database. Also, use of EWPR_{VOL,T} rank at the time of giving search results to the user. An extended architecture is proposed in this paper to make the crawling and searching process effective.

8. Future Scope

For future work, we can also plan to add a dashboard facility for the webmasters, who will add our Extended VOL Analytics Script on their websites. With Dashboard, they can see from which resources (like search engine, blogs, forums and websites) visitors are coming to their websites. How many no. of times a same page is requested from another page or same resource. Also from which location, browser and Operating System, etc their page are accessed. How much time the user has actively devoted on his page. To improve the quality of crawling more, work can be done on finding the better threshold values of Weighted Page Rank and Extended Weighted Page Rank Based on Visits of Links algorithms. User Attention Time of page can be made better as an enhancement in this work. Our Proposed Crawling Architecture work needs a huge amount of Computer Memory for its operation if that can be reduced that will be a great achievement in the enhancement of this work.

Acknowledgment

I would like to express my sincere gratitude to Ms. Harjinder Kaur, Asstt. Prof. and Deptt. Head at SSIET Dinnagar and Mr. Sachin Gupta, Director of Sodiz

Technologies who gave their heart whelming full support in the completion of this research paper with their stimulating suggestions and encouragement to go ahead in all the time. Mr. Harjinder Kaur has beacons light to me as a guide at all stages of preparation of my research work. Mr. Gupta has always been a source of inspiration and confidence for me. I express my thanks from the core of my heart to my parents and friends for encouragement, cooperation and also help in challenging circumstances. At last I am very thankful to my GOD who has given me this golden opportunity to do M.Tech as well as to do research work.

References

- [1] Internet World Stats survey report available at - << <http://www.internetworldstats.com/stats.htm> >>.
- [2] Pew Research center's Internet and American Life Project Survey report available at - << <http://www.pewinternet.org/2012/03/09/main-findings-11/> >>.
- [3] Average Traffic a website receives from a Search Engine is << <http://moz.com/community/q/what-is-the-average-percentage-of-traffic-from-search-engines-that-a-website-receives> >>
- [4] Size of World Wide Web is available at << <http://www.worldwidewebsite.com/> >>
- [5] Carlos Castillo, Mauricio Marin, Andrea Rodrigue and Ricardo Baeza-Yates, "Scheduling Algorithms for Web Crawling" Proceedings of the Web Media & LA-Web 2004, 0-7695-2237-8 ©2004 IEEE, Pages 10-17.
- [6] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280 (5360) : 98–100, 1998.
- [7] Introduction to Web Crawler is available at - << http://en.wikipedia.org/wiki/Web_crawler >>
- [8] Introduction to Web Crawler is available at - << <http://searchsoa.techtarget.com/definition/crawler> >>
- [9] Amit Chawla and Rupali Ahuja, "Crawling the Web : Discovery and Maintenance of Large-Scale Web Data", International Journal of Advances in Engineering Science (IJAES), ISSN: 2231- 0347, Volume-3, Pages 62-66, July 2013.
- [10] Sachin Gupta, Sashi Tarun and Pankaj Sharma, "Controlling access of Bots and Spamming Bots", International Journal of Computer and Electronics Research (IJCER), ISSN: 2278-5795, vol. 3,issue 2, April 2014.
- [11] Sonal Tuteja, "Enhancement in Weighted PageRank Algorithm Using VOL", IOSR Journal of Computer Engineering (IOSR- JCE), ISSN: 2278-0661, vol. 2, issue 6, pp. 135-141, Sept-Oct 2013.
- [12] Shweta Agarwal and Bharat Bhushan Agarwal, "An Improvement on Page Ranking Based on Visits of Links", International Journal of Science and Research (IJSR), ISSN: 2319-7064, vol. 2, issue 6, pp. 265-268, June 2013.
- [13] S. Brin, and Page L., "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, vol. 30, issue 1-7, pp. 107-117, 1998.
- [14] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
- [15] Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCT)-2011, 978-1-4577-1385-9.
- [16] Neelam Tyagi and Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, vol. 2, issue 3, pp. 441–446, July 2012.

- [17] Animesh Tripathy and Prashanta K Patra, "A Web Mining Architectural Model of Distributed Crawler for Internet Searches Using PageRank Algorithm", Asia-Pacific Services Computing Conference, 978-0-7695-3473-2/08 © 2008 IEEE, Pages 513-518.
- [18] Lay-Ki Soon, Yee-Ern Ku and Sang Ho Lee, "Web Crawler with URL Signature – A Performance Study", 4th Conference on Data Mining and Optimization (DMO) 978-1-4673-2718-3/12 ©2012 IEEE, Pages 127-130.
- [19] Farha R. Qureshi and Amer Ahmed Khan, "URL Signature with body text normalization in a web crawler", International Journal of Societal Applications of Computer Science (IJSACS), ISSN 2319 – 8443, vol. 2, issue 3, Pages 309-312, March 2013.
- [20] Saurabh Pakhidde , Jaya Rajurkar and Prashant Dahiwal, "Content Relevance Prediction Algorithm in Web Crawlers to Enhance Web Search", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), ISSN: 2278 – 1323, vol 3, issue 3, March 2014.
- [21] Prashant Dahiwal, Pritam Bhowmik, Tejaswini Bhorkar and Shraddha Shahare, "Rank Crawler : A Web Crawler with Relevance Prediction Mechanism for True Web Analysis", International Journal of Advance Foundation and Research in Computer (IJAFRC), ISSN: 2348-4853, vol. 1,issue 4, April 2014.
- [22] Information on HTTP_REFERER is available at - << http://en.wikipedia.org/wiki/HTTP_referer >>.
- [23] Information on Url Normalization is available at - << http://en.wikipedia.org/wiki/Url_normalization >>.
- [24] Information on MD5 Hashing Algorithm is available at - << <http://en.wikipedia.org/wiki/MD5> >>.
- [25] Introduction to WHOIS is available at - << <http://en.wikipedia.org/wiki/Whois> >>.
- [26] Sachin Gupta and Pallvi Mahajan, "Improvement in Weighted Page Rank based on Visits of Links (VOL) Algorithm", International Journal of Computer and Communications Engineering Research (IJCCER), ISSN: 2321-4198, Vol. 2, Issue 3, Pages 119-124, May 2014.
- [27] Sachin Gupta and Sashi Tarun, "Extended Architecture of Web Crawler", International Journal Of Computer & Electronics Research (IJCER), ISSN: 2278-5795, Vol. 3, Issue 3, Pages 147-169, June 2014.
- [28] Isha Mahajan, Harjinder Kaur and Dr. Darshan Kumar, "Extended Weighted Page Rank based on VOL by finding User Activities Time and Page Reading Time", International Journal of Engineering Works (IJEW), ISSN: 2409-2770, Vol. 7, Issue 2, Pages 41-48, Feb 2017.
- [29] Introduction to Code Minification is available at - << <https://developers.google.com/speed/docs/insights/MinifyResources> >>.
- [30] Javascript Code Minification Api is available at - << <https://javascript-minifier.com/> >>.
- [31] Introduction to Cron Jobs is available at - << <https://code.tutsplus.com/tutorials/scheduling-tasks-with-cron-jobs--net-8800> >>.
- [32] Domain age calculating Api is available at - << <https://github.com/99webtools/PHP-Domain-Age> >>.