Crawler with Search Engine based Simple Web Application System for Forum Mining

Parina Shah[1] Ms. Gayatri Pandi (Jain)[2]
[1]M.E Student [2]Head of the Department
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering & Technology, Ahmedabad, Gujarat, India

Abstract— A Web Crawler or a Bot or an Indexer is a program that visits Web sites for reading the content of pages and other information so that it can create index for search engine. Here, the aim of web crawler is to crawl relevant content from the Web Forum with minimal overhead. Forums are an open source portal for information exchange. Duplicate URL elimination as well as grouping of Page Flipping URLs having similar layout is done. Web Forums have navigation paths which are similar that are connected by specific URL types which lead users from entry page to thread page. Last modified date of the post, number of the threads or posts is also collected to know about the updated thread or post. The precision and recall value achieved for the entry pages were 98.03% and 96.02% respectively. Crawler achieved 98.96% coverage and 98.32% effectiveness by eliminating irrelevant information and URLs.

Key words: URL elimination, ITF, Web Crawler

I. INTRODUCTION

A web crawler is a system used by search engines for indexing of web pages. Crawlers can be used for a variety of purposes. They are one of the main components of web search engines, systems that sequences a corpus of web pages, indexes them, and allows users to issue queries against the index and find the web pages that match the queries. A web crawler crawling upon the forums is a technique called forum mining. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases and derives its name from the similarities between searching for valuable information in a large database. Crawling is a system for bulk downloading of web pages from internet. It crawls relevant content from world wide web. Web crawlers are used for indexing pages for search engines, archiving the web, analysing the web etc. Web search engines and other sites use web crawling software to adopt their web content or indexing of other site's web content. It 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. Crawlers are used for validation of hyperlinks and HTML code.

A search engine uses the web crawler. So, forums generally are topic based. User generally makes a search based on a particular topic called Query term. So, all the posts related to that topic are crawled and the user is provided with the output in particular time interval. A post is generally the reply posted by the users or members of the website. There exists certain URLs in the posts which may redirect the user to the same page. So, duplication of URLs may exist. This creates unnecessary crawling and indexing of the same page multiple times. It not only increases the search time but also reduces the efficiency.

For example, www.gtu.ac.in website may exists several times in a forum related to result or online material download. Thus, same website is crawled multiple times by the crawler. De-duplicating the pages is an important concept to be resolved.

Forums exist in many different layouts or styles and powered by a variety of forum software packages, and they always have navigation paths to lead users from entry pages to thread pages. Thus, we need to find the shortest path which may lead us from entry page to thread page.
Patterns are extracted of the Index/Thread/Flipping (ITF) URLs. Thus, this patterns are then used to find the ITF regular expressions. This ITF regexes help us to provide best results in minimal time.

So, The goal of Crawler is to crawl relevant content such as user posts from forums with minimal overhead.

In our existing system, there exist several URLs which navigate to the same page on web forums. This increases the time of a crawler due to repetitive crawling of pages. Forums have classified all the records in the web and based upon user request the information is get searched and transmitted. The posts in the forums may get constantly updated and so the crawler needs to get added in the list. Duplicate pages exists in the web crawler list. So, it reduces the efficiency of the crawler. This concept is called Page Flipping and the URLs are called Page Flipping URLs.

So, it is very much important for any crawler to uniquely identify a page URL and navigate correctly from entry page to thread page. So, pattern evaluation and regular expressions for ITF are important aspects to be focused for better crawling.
The following shows the basic elements of Forum:

A. Web Forum Structure:

A web forum is a tree like or hierarchical structure. A forum can be divided into different categories for the relevant conversations that occur and then all of the posted messages under these categories are sub-forums and these sub-forums can be further divided into more sub-forums.

B. User Groups:

A member or user of the forum can automatically get access to a more privileged user group based on conditions set by the administrator. All the anonymous users of the site are known as visitors. Visitors have privilege to be granted access to all functions that do not require break their privacy. A guest or visitor can usually view the contents and then the posted messages of the forum.

C. Posts:

A post is a conversation which user makes inside a forum upon a topic. User to writes a message which is enclosed into a block containing the user's details and the date and time it was posted. There is certain limit for submitting the post. The message should be usually of minimum 10 characters and upper limit to the length of a message is 10,000 to 20,000 characters.

The following literature surveys show different methods of crawling and techniques to make crawler efficient for forum mining.

II. RELATED WORK

In [1] Yan Guo et.al presented Board Forum Crawling (BFC) which is used to crawl Web forum. This method checks human behaviour of visiting Web Forums and exploits the organized characteristics of the Web forum sites. Thus, BFC uses a method, which starts crawling from the homepage, followed by entering each board of the site, and then crawls all the included posts of the site directly. They have also used this method in a real project, and have crawled approx. 12000 Web forum sites successfully[1]. Traditional breadth-first crawling, which is called as TBFC in the paper, is popularly used in all kinds of cases. To visit a post in a board, human usually starts from the homepage, and then it enters into a board, and then to find the post. This process says that the forum is organized in a structurally manner. Thus they have found that there exist mainly 3 kinds of pages in one Web forum site i.e homepage and board page and post page. The Precise of their BFC can reach up to 90%.[1]

In [2] Jingtian Jiang, et.al presented FoCUS, a new approach towards web crawler. They have presented a crawler named FoCUS (Forum Crawler Under Supervision). The aim of FoCUS is to crawl only relevant forum content from the web with minimal overhead. They always have similar implicit navigation paths which are connected by specific URL types to lead users from entry pages to thread pages. They call pages between the entry page and thread page which are on a breadth-first navigation path the index page. They have these implicit paths as the following navigation path (EIT path):

entry page -> index page -> thread page [2]

In their experiment [2] over 160 forum sites (10 pages each of index, thread, and other page) each powered by a different forum software package, our classifiers achieved 96% recall and 97%precision for index page and 97% recall and 98% precision for thread page with different amount of training data.

In [3] H.S. Koppula et.al have presented that the existence of duplicate documents in the World Wide Web adversely affects indexing ,relevance and crawling , which are the main building blocks of web search. Thus, in this paper, they have presented a set of techniques to mine rules from URLs and then utilize these rules for de-duplication using only the URL strings without fetching the content explicitly. Their technique is mainly composed of mining the crawl logs and further utilize these clusters of similar pages to extract transformation rules, which are used to normalize URLs which belong to each cluster. They have extended the representation of URL and the Rule presented in it. The se extensions result in better utilization of the information encoded in the URLs to generate precise Rules with more coverage.

In [4], named "Intelligent Crawler for Web forums based on improved regular expressions", the authors have presented a special crawler for Internet forums. The main data collected from forum contents are the posts, and all relevant information that goes with them. A post always contains post body i.e. name of the author of the post, text and the date when the post was created. It also contains additional information such as the link to the author's profile, gender, post number, date of joining the forum, anchor, frequency of activities, author's picture etc are also collected and can be of important use in later analysis. Thus, In most cases, users want to get the structured information or data directly instead of web pages[4].

In [7] R.Cai et.al , presented iRobot crawler for web crawling. The authors have focused upon deep web crawling and near duplicate detection. Its main aim is to understand Content and structure of a forum site. The main goal of iRobot is to automatically rebuild the graphical architecture representation, i.e., the sitemap of the target Web forum and then select an optimal traversal path which only traverses informative pages and skip invalid and duplicate ones. Thus, iRobot can significantly reduce duplicate and invalid pages, without losing the valuable ones. It only needs to pre-sample maximum of 500 pages for discovering necessary knowledge. iRobot can keep around 95% page relations in crawling, which is very useful for data mining tasks and further indexing [7]

III. TERMINOLOGY

A. Type of Pages:

Forum pages are classified into four page types according to DOM tree structure:

B. Entry Page:

A page that is on the top of the DOM tree model i.e the home page of a website is the entry page. It is lowest common ancestor of all thread pages in a forum. In forum sites, an entry page is usually the one which may have the subjects which are open for discussion or any kind of information about the website. Each and every entry page have links or URLs attached with it.

Eg- Ajax

Here, <a> is the anchor tag which has "Ajax" as the anchor text. The "href" keyword contains the link to go to the Ajax page.

Crawler extracts all such links and their text, along with their status of whether the URL exist or has been removed.

C. Index Page:

The URLs which are included under the keyword "/forum/" are extracted as Index Pages. They are called Index Pages as the hold an index for kind of topics the forum is open for.

Eg- http://www.hotscripts.com/forums/

All the URLs under the following link will be extracted. The topics open for discussion can be known by the text corresponding to these links.

D. Thread Page:

A page that contains a table-like structure; each row in it contains information on URLs on the posts with user generated content (UGC). The child pages of Index Pages are under this category. Any user on a forum site can checkout the topics open for discussion in a forum. Users can get a lot of information from various topics available. Registered users can post replies inside the thread. Each thread has a counter associated with it which shows the total number of posts or replies given by user in a particular thread. Number of views is also available in certain forum sites. A forum site contains bundles of link in a thread page.

E. Other Page:

A page that is not an index, entry or thread page.
Type of URLs: There are four types of URLs:

F. Entry URL:

A URL of the home page is included as an Entry URL.

G. Index URL:

An Index URL or Board URL is a URL that is on an index page or entry page and it points towards the index page. The anchor text of index page shows the title of its destination board.

H. Thread URL:

A Thread URL is the one that is on an index page and points towards the thread page. The title of its destination thread is shown by its anchor text.

I. Page-flipping URL:

There are URLs which lead the users to another page of the same board or the same thread. So, if we correctly deal with page-flipping URLs then it will enable a crawler to download all threads of a forum or all posts in a long thread.

J. Other URL:

A URL that is not a thread, index, or page-flipping URL.

K. Internal URLs:

Any URL which matches the URL of selected Website is Internal URL.

L. External URLs:

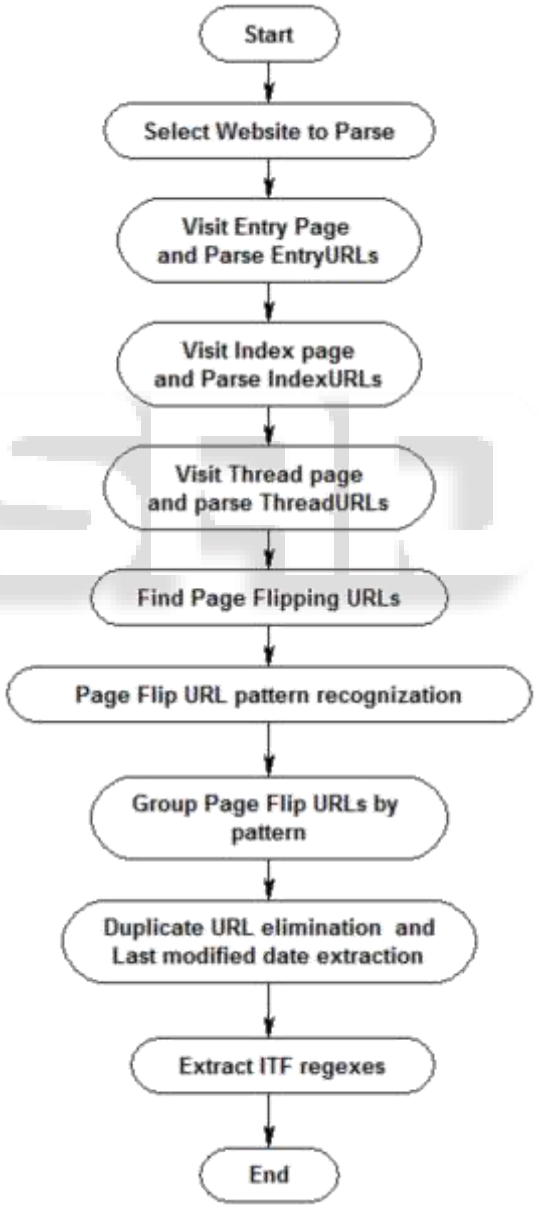
Any URL which is an outlink or referenced by the website is an External URL.

IV. METHODOLOGY

The main disadvantage of our existing system is crawling of URLs which may lead users to same thread or page inside a forum. Thus, it increases the crawling time of the crawler. Page Flipping and Clone mining are very much required to make the crawler more efficient and increase the coverage of the crawler. Page Flipping URL's destination pages have similar layout as source pages. Clone mining is done to identify the URLs which have same structure but different data. Index-Thread-Flipping URLs are analyzed and patterns are extracted. These Regular expressions are used to eliminate Page Flipping URLs. This crawling technique thus will help to improve the precision and recall value of a crawler.

A. Process is as Follows:

1) Select a Website to parse: The website to be parsed must be selected.
2) Visit Entry page and Entry URL: Parse Entry URLs from Entry page along with their anchor text.
3) Visit Index page and Index URL: Parse Index URLs from Index page along with their anchor text.
4) Visit Thread page and Thread URL: Parse Thread URLs from Thread page along with their anchor text.
5) Extract Page Flipping Pattern and group them according to their similarity.
6) Extract last modified date and ITF regular expression



B. Website Selected

Two websites are selected i.e www.hotscripts.com and www.forumsoftware.ca . Now one website is selected to show multi-level page-flipping, while the later one is to show single-level page-flipping.

According to the website selected, further extraction of pages and URLs occur.

C. Entry page and Entry URL

The homepage of the selected website is fetched and all the links on that page are fetched and stored. Anchor text is extracted of corresponding URLs. A URL is internal or external URL is also specified by the crawler.

Entry Pages are crawled by the crawler and the crawler displays information :

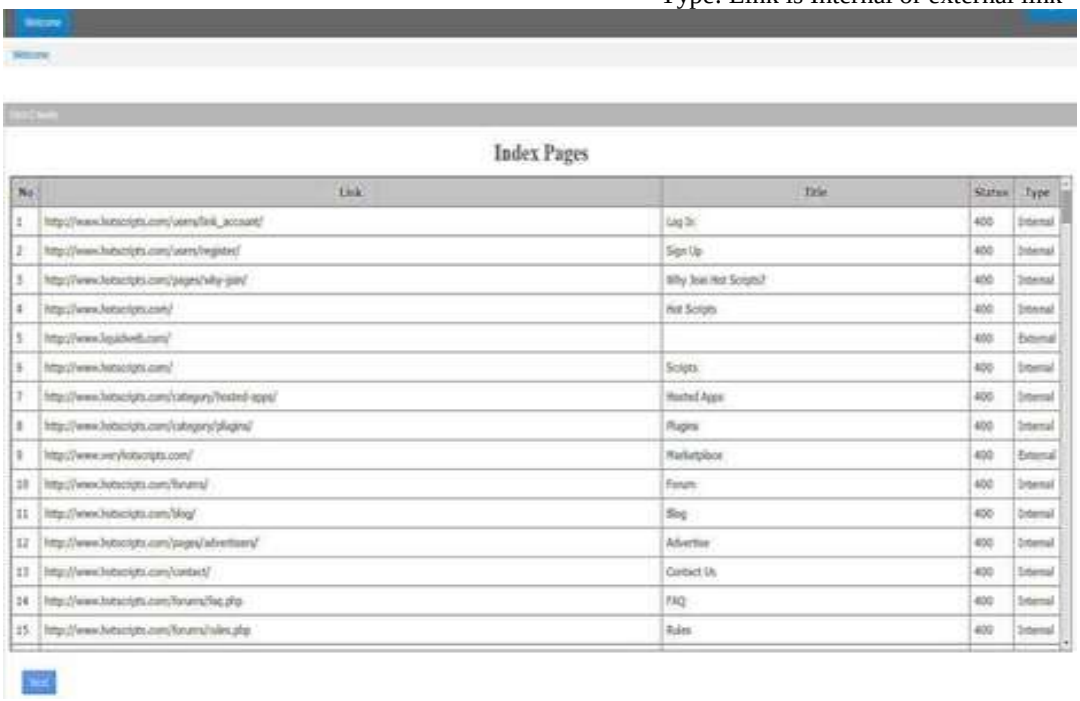
- Link: URL crawled
- Title: Anchor text attached with URL
- Status: Status of working URL
- Type: Link is Internal or external link

D. Index Page and Index URL

The URL having "/forum/" and the anchor text as "forum" is Index URL of an Index page. Varied kind of topics are included under this page. Title of the URL, status and type of URL are shown.

Index Pages are crawled by the crawler and the crawler displays information :

- Link: URL crawled
- Title: Anchor text attached with URL
- Status: Status of working URL
- Type: Link is Internal or external link



E. Thread Page and Thread URL

Child pages of Index page are called Thread pages. It contains:

1) Index URL name: Name of Index URL the threads are attached.
2) Index URL link: Link of Index URL the threads are attached.
3) Number of Threads: number of threads in particular Index page
4) Last modified: Last modified date of thread posted.
5) Posts: Number of posts inside a thread

Thread Pages

Total Threads : 45801

No	Index Url Name	Index Url Link	Threads	Last Modified	Posts
1	Hot Scripts Forum Questions, Suggestions and Feedback	http://www.hotscripts.com/forums/hot-scripts-forum-questions-suggestions-feedback/	442	06-30-13 08:10 PM	1,896
2	General HotScripts Site Discussion	http://www.hotscripts.com/forums/general-hotscripts-site-discussion/	409	06-29-13 11:35 PM	1,737
3	HotScripts Site Bug Reports	http://www.hotscripts.com/forums/hotscripts-site-bug-reports/	137	07-01-13 04:56 AM	585
4	PHP	http://www.hotscripts.com/forums/php/	15,542	01-19-14 07:11 PM	72,281
5	Perl	http://www.hotscripts.com/forums/perl/	748	05-16-13 04:31 AM	2,688
6	ASP	http://www.hotscripts.com/forums/asp/	1,446	11-21-12 09:33 AM	5,275
7	ASP.NET	http://www.hotscripts.com/forums/asp-net/	638	06-30-13 08:28 PM	2,054
8	C/C++	http://www.hotscripts.com/forums/c-c/	586	06-30-13 11:47 AM	1,675
9	Visual Basic	http://www.hotscripts.com/forums/visual-basic/	724	06-30-13 07:57 PM	2,640
10	Windows .NET Programming	http://www.hotscripts.com/forums/windows-net-programming/	548	06-29-13 01:18 PM	1,705
11	Everything Java	http://www.hotscripts.com/forums/everything-java/	618	06-30-13 08:37 PM	1,786
12	Other Languages	http://www.hotscripts.com/forums/other-languages/	178	06-30-13 07:53 PM	416
13	HTML/XHTML/DML	http://www.hotscripts.com/forums/html-xhtml-xml/	1,370	07-01-13 04:51 AM	5,743
14	HTML5	http://www.hotscripts.com/forums/html5/	108	05-17-13 08:07 AM	158
15	JavaScript	http://www.hotscripts.com/forums/javascript/	3,475	06-28-13 01:31 AM	13,133

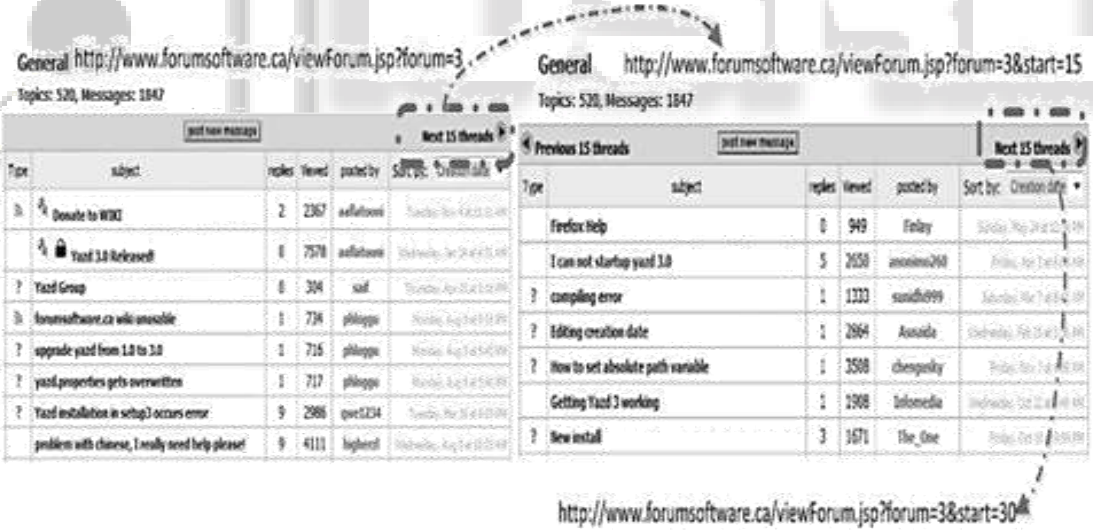
F. Page Flipping Page

1) Types

There are two types of Page flips:

1) Single Level:

The user can see only one level of pages at a time and not directly jump between different levels. The single page-flipping URLs that appear in their source and destination pages have the same anchor text but different URL strings.



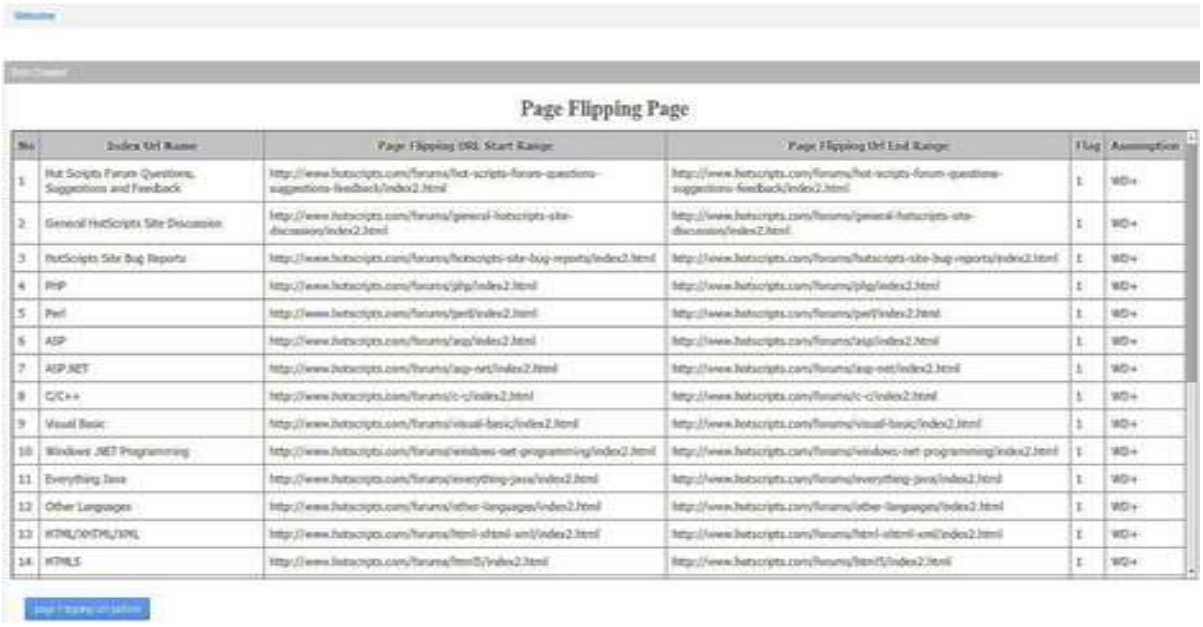
1) Multi-Level:

In multi-level Page Flipping, user can view number of pages in a thread and can also jump between various levels of pagination as shown below.

2) Page-Flipping Page and Page-Flipping URL

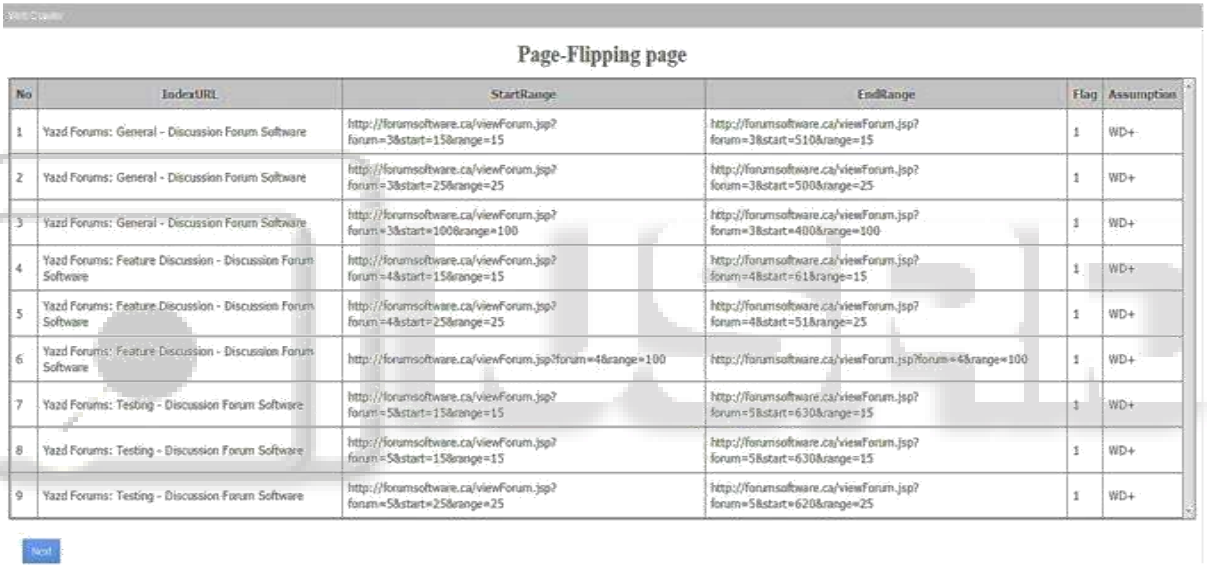
All the URLs having similar layout come under page-flip URL and are grouped together. It includes:

1) Index URL name: Name of Index URL the threads are attached.
2) Index URL link: Link of Index URL the threads are attached.
3) Page flipping URL start range: Starting range of page flips.
4) Page flipping URL end range: Ending range of page flips.
5) Flag: 1 if the corresponding Index is crawled else 0
6) Assumption: W for word replacement, d for digit replacement



Page Flipping Page

No	Index Url Name	Page Flipping URL Start Range	Page Flipping Url End Range	Flag	Assumption
1	Hot Scripts Forum Questions, Suggestions and Feedback	http://www.hotscripts.com/forums/hot-scripts-forum-questions-suggestions-feedback/index2.html	http://www.hotscripts.com/forums/hot-scripts-forum-questions-suggestions-feedback/index2.html	1	WD+
2	General HotScripts Site Discussion	http://www.hotscripts.com/forums/general-hotscripts-site-discussion/index2.html	http://www.hotscripts.com/forums/general-hotscripts-site-discussion/index2.html	1	WD+
3	HotScripts Site Bug Reports	http://www.hotscripts.com/forums/hotscripts-site-bug-reports/index2.html	http://www.hotscripts.com/forums/hotscripts-site-bug-reports/index2.html	1	WD+
4	PHP	http://www.hotscripts.com/forums/php/index2.html	http://www.hotscripts.com/forums/php/index2.html	1	WD+
5	Perl	http://www.hotscripts.com/forums/perl/index2.html	http://www.hotscripts.com/forums/perl/index2.html	1	WD+
6	ASP	http://www.hotscripts.com/forums/asp/index2.html	http://www.hotscripts.com/forums/asp/index2.html	1	WD+
7	ASP.NET	http://www.hotscripts.com/forums/asp-net/index2.html	http://www.hotscripts.com/forums/asp-net/index2.html	1	WD+
8	C/C++	http://www.hotscripts.com/forums/c-c/index2.html	http://www.hotscripts.com/forums/c-c/index2.html	1	WD+
9	Visual Basic	http://www.hotscripts.com/forums/visual-basic/index2.html	http://www.hotscripts.com/forums/visual-basic/index2.html	1	WD+
10	Windows .NET Programming	http://www.hotscripts.com/forums/windows-net-programming/index2.html	http://www.hotscripts.com/forums/windows-net-programming/index2.html	1	WD+
11	Everything Java	http://www.hotscripts.com/forums/everything-java/index2.html	http://www.hotscripts.com/forums/everything-java/index2.html	1	WD+
12	Other Languages	http://www.hotscripts.com/forums/other-languages/index2.html	http://www.hotscripts.com/forums/other-languages/index2.html	1	WD+
13	HTML/XHTML/XML	http://www.hotscripts.com/forums/html-xhtml-xml/index2.html	http://www.hotscripts.com/forums/html-xhtml-xml/index2.html	1	WD+
14	HTML5	http://www.hotscripts.com/forums/html5/index2.html	http://www.hotscripts.com/forums/html5/index2.html	1	WD+



Page-Flipping page

No	IndexURL	StartRange	EndRange	Flag	Assumption
1	Yazd Forums: General - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=3&start=15&range=15	http://forumsoftware.ca/viewForum.jsp?forum=3&start=510&range=15	1	WD+
2	Yazd Forums: General - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=3&start=25&range=25	http://forumsoftware.ca/viewForum.jsp?forum=3&start=500&range=25	1	WD+
3	Yazd Forums: General - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=3&start=100&range=100	http://forumsoftware.ca/viewForum.jsp?forum=3&start=400&range=100	1	WD+
4	Yazd Forums: Feature Discussion - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=4&start=15&range=15	http://forumsoftware.ca/viewForum.jsp?forum=4&start=61&range=15	1	WD+
5	Yazd Forums: Feature Discussion - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=4&start=25&range=25	http://forumsoftware.ca/viewForum.jsp?forum=4&start=51&range=25	1	WD+
6	Yazd Forums: Feature Discussion - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=4&range=100	http://forumsoftware.ca/viewForum.jsp?forum=4&range=100	1	WD+
7	Yazd Forums: Testing - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=5&start=15&range=15	http://forumsoftware.ca/viewForum.jsp?forum=5&start=630&range=15	1	WD+
8	Yazd Forums: Testing - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=5&start=15&range=15	http://forumsoftware.ca/viewForum.jsp?forum=5&start=630&range=15	1	WD+
9	Yazd Forums: Testing - Discussion Forum Software	http://forumsoftware.ca/viewForum.jsp?forum=5&start=25&range=25	http://forumsoftware.ca/viewForum.jsp?forum=5&start=620&range=25	1	WD+

Next

V. EVALUATION

This Crawler is efficient in learning ITF regexes and is effective in detection of entry URL, index URL, thread URL and page-flipping URL. In this section, comparison of Crawler is done with FoCUS crawler in terms of effectiveness and coverage.

We selected 2 forums having popular software packages used by many forum sites.

The effectiveness is calculated as:

$$Effectiveness = \frac{\#Crawled\ threads}{\#Crawled\ pages} \times 100\%$$

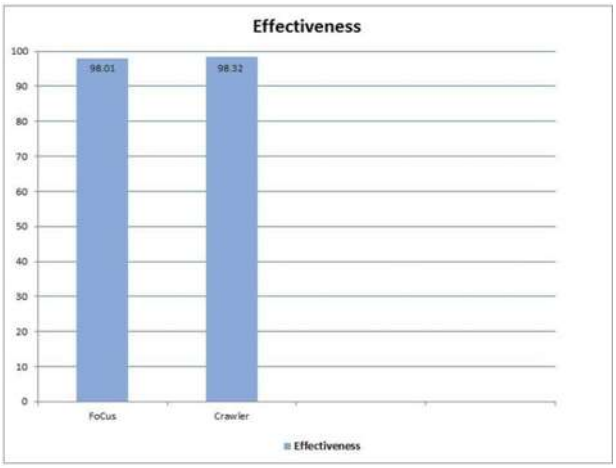
Crawled threads are the total threads extracted by the crawler. Crawled pages are the total number of URLs parsed from all pages that are included in forum website.

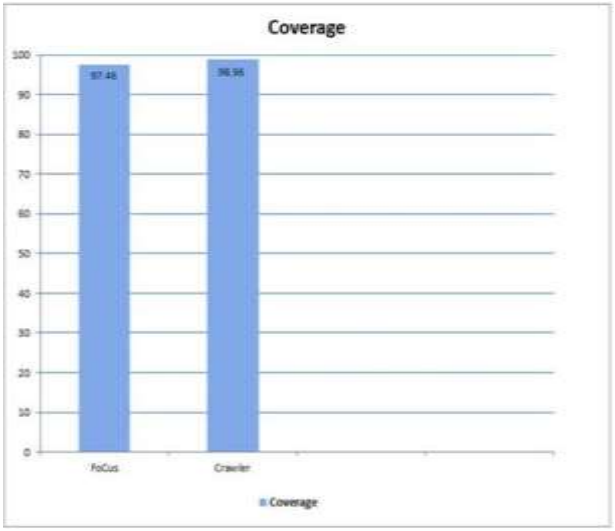
The coverage is calculated as follows:

$$Coverage = \frac{\#Crawled\ threads}{\#Threads\ in\ all} \times 100\%$$

Crawled threads are the total threads extracted by the crawler. The total number of threads present in thread pages are taken as threads in all.

Crawler achieved 98.96% coverage and 98.32% effectiveness by eliminating irrelevant information and URLs.





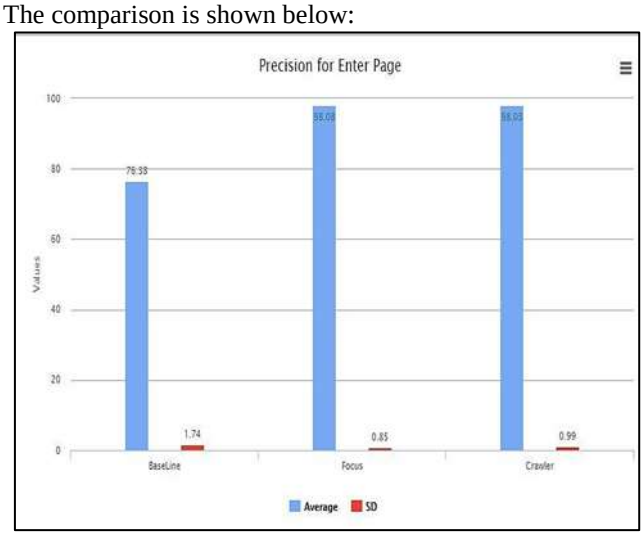
Precision and Recall of Entry Pages is calculated by comparing it with FoCUS and baseline crawler.

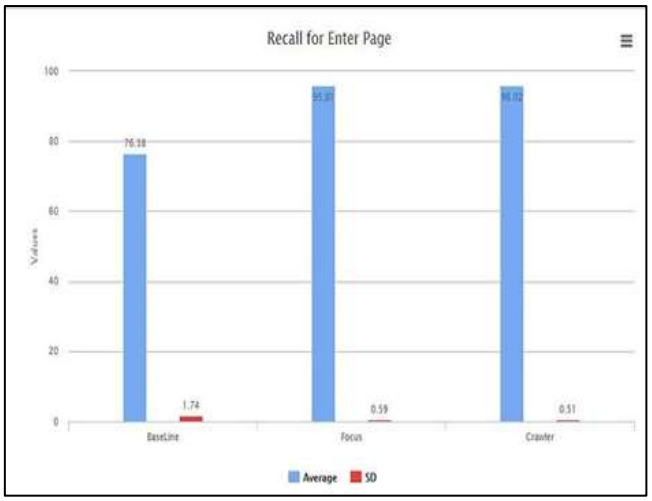
Name of crawler	Precision Average %	Precision SD	Recall Average %	Recall SD
Baseline	76.38	1.74	76.38	1.74
Focus	98.08	0.85	95.81	0.59
Crawler	98.03	0.99	96.02	0.51

Precison is fraction of retrieved pages that are relevant.
Precision = Pr/Tn
Where, Pr = number of pages visited that are relevant
Tn = total number of pages visited.
Recall is fraction of relevant pages that are retrieved.

$$\text{target\_recall} = \frac{|Ptg \cap Pcr|}{|Ptg|}$$

Where, Ptg = target pages
Pcr = crawled pages.
The comparison is shown below:





No loss of data can occur via this technique as only the URLs are processed. No tampering is done on the content of the pages. Forum sites have different software packages and resources, so some of the forum sites will be used for experimenting this approach.

VI. Conclusion

The paper illustrates various methods and techniques which are used for making an efficient web crawler. Web Forums have different layouts, different structure and different types of pages. Any kind of web crawler needs to improve its efficiency, coverage and precision by understanding the structure of a web forum. Learning of ITF regex pattern is necessary by constructing the DOM structure. Page-flip type must recognized in order to perform grouping of various Page-flipping URLs. It must be able to eliminate irrelevant pages or any kind of unnecessary information to utilize the time in efficient way. Thus, Techniques used for Parsing and extraction of URLs play an important role in improving the crawler efficiency upon a forum site.

In future, a more normalized URL pattern can be created for unconventional type of URLs and find ways to crawl JavaScript generated URLs.

References

[1] Yan Guo, Kui Li, Kai Zhang, Gang Zhang, "Board Forum Crawling: A Web Crawling Method for Web Forum", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.
[2] Jingtian Jiang, Nenghai Yu, Chin-Yew Lin, "FoCUS: Learning to Crawl Web Forums",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:6 YEAR 2013
[3] H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg, and A. Sasturkar, "Learning URL Patterns for Webpage DeDuplication," Proc. Third ACM Conf. Web Search and Data Mining, pp. 381-390, 2010.
[4] Miloš Pavković, Prof. Jelica Protić, "Intelligent Crawler for Web Forums based on Improved Regular Expressions",21st Telecommunications forum TELFOR 2013/IEEE/YEAR 2013
[5] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, "Study Of Web Crawler and Its Different Types", IOSR Journal of Computer Engineering (IOSR-

JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05
[6] Bergholz, B. Chidlovskii,"Crawling for domain-specific Hidden Web resources", Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE'03). pp.125-133,IEEE Press, 2003
[7] R.Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An Intelligent Crawler for Web Forums", In Proc. of 17thWWW, pages 447-456, 2008
[8] http://www.slideshare.net/iamthevictory/web-crawler
[9] http://hunyadi.info.hu/levente/en/publications/6-crawlernet
[10] http://mias.uiuc.edu/files/tutorials/mercator.pdf
[11] http://en.wikipedia.org/wiki/CRAWLER.

