Forum Focused Crawler

Priyanka Bandagale[1], Neha Sawantdesai[2], Rakshanda Paradkar[3], Piyusha Shirodkar[4]
[1, 2, 3, 4] Dept of IT Engineering
[1, 2, 3, 4] Finolex Academy of Management & Technology, Ratnagiri.

Abstract- *Web forums, at their best can create remarkable communities that could never have existed before the internet. They bring together a large number of people from all over the world to focus on a specific topic or interest. This gives rise to a forum web crawler which will particularly crawl only forum sites. This paper begins with defining the structure of the Web Forum Sites followed by Proposed System. It also explains the architecture and working of forum crawler. The aim of the paper is to give importance about forum web crawling.*

Keywords- web crawler, web forum, URL pattern, Index URLs, Thread URLs, ITF regex..

I. INTRODUCTION

The web contains large bulk of information on various topics. Compared to the traditional collection repositories such as libraries etc., the Web has no centrally organized content structure. This data can be downloaded using web crawler. A web crawler is a program or automated script which browses the World Wide Web in a methodical, automated manner. This process is called Web crawling. Many Legal sites in specific search engine use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for latter processing. Crawlers can also be used for automating maintenance tasks on a Website, such as validating HTML code or checking links. Also, crawlers can be used to gather specific types of information from web pages such as harvesting email addresses. It is also called as web robot or web spider. Web crawlers can be used in various areas. Most importantly it indexes a large set of pages and allow other people to search this indexes. Nowadays, web forums becoming very popular which deals with various topics. It provide space for users to access, discuss and share the information.

This paper gives the details about web forum crawler which is designed to crawl only web forums. The first section seals with Literature Survey. Next section gives general idea about web forum and its structure. In the next section, we have explained proposed system in detail. The next section explains the system architecture and implementation of the web forum crawler.

II. LITERATURE SURVEY

A. Board Forum Crawling:

This method exploits the organized characteristics of the Web forum sites and simulates human behaviour of visiting Web Forums. The method begins by crawling homepage, and after entering each board of the site it crawls all the posts of the site directly. Board Forum Crawling can crawl most meaningful information of a Web forum site efficiently and simply.

Typically to visit a post in a forum site, humans visit the home page, then find the post across multiple links. This process proves the structural organization of forums.

B. iROBOT:

This approach automatically understands the content and structure of each forum sites and then decides how to traverse to different pages in forum sites. To find out such traversal paths, it first automatically re-builds the sitemap of the target web forum and then it selects the best possible traversal paths which only traverses informative pages and skips invalid and duplicate pages.

iRobot system consist of two major parts:

1) Offline sitemap reconstructing and traversal path selection
2) Online crawling.

The offline sitemap reconstructing and traversal path selection part involves four major steps:

1) Repetitive region-based clustering
2) URL-based sub-clustering
3) Informativeness estimation
4) Traversal path selection

C. FoCUS:

FoCUS is effective for large-scale forum crawling. FoCUS defines EIT path which permit over one path and URL patterns would not be affected by a change in page structure. It

shows way to learn regular expression patterns (ITF regexes) that recognize the index uniform resource locator (URL), thread uniform resource locator (URL) and page-flipping uniform resource locator (URL) using the page classifiers.
FoCUS adopts a simple URL string de-duplication technique. The system consist of two major parts as shown below:

1) Learning Part
2) Online Crawling Part

The main advantage of FoCUS is that it can avoid duplicates without duplicate detection. This technique uses EIT path to traverse from entry pages through a sequence of index pages to thread pages. EIT means Entry – Index – Thread path. Index URLs are the links between an entry page and an index page or between two index pages. Thread URLs are the links between an index page and a thread page. Page-flipping URLs are the links connecting multiple pages of a board and multiple pages of a thread.

III. WEB FORUM STRUCTURE

Internet forum is an online discussion place on a website. Forum sites allow it's user to request and exchange information among them. In addition, the forum sites allow users to view forum postings and to post messages in it. When posting in a forum, the users can create new topics (or "threads") or post replies within existing threads. Web forums are almost available for all kinds of topics. Examples include software support, help for webmasters, and programming discussions, sports, entertainment, games, technical discussions. Due to the richness of information in forums, researchers are increasingly interested in mining knowledge from them. A link structure of forum sites as shown in fig 1.

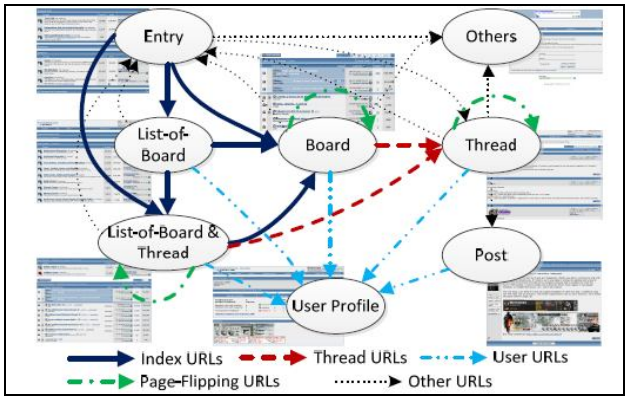

Fig 1. Typical Link structure in Web Forums

The goal of our crawler is to crawl relevant forum content from the web with minimal overhead and to organize crawled content into logical units, in order to make it easier for further processing and analysis. It avoids crawling through unproductive paths by limiting the search to a particular topic and learning features of links and paths. It uses minimum bandwidth and less storage space to store specific data. Each page in forum site may have its own layouts. Based on their layout structure, the pages in forum sites are classified into four categories:

- Entry page: The home page of the forum site which contains a list of boards.
- Index page: An index page contains table-like structure, where each row in the table contains information of a board or thread.
- Thread page: A thread page contains a list of users' posts.
- Other pages like login control, about us, user profile pages, etc.

As shown in Fig 2, the top level layer is for retrieving details of web sources such as WWW or Internet. Which has URL and their data i.e. URI of web pages. The middle layer provides the interaction with both user and web. Download all needed details to user by that crawler launched by the system server. The bottom layer performs the database functions to user as well as server. The user interacts with database to search and retrieve the information and the search engine update their database index based on the crawled details.

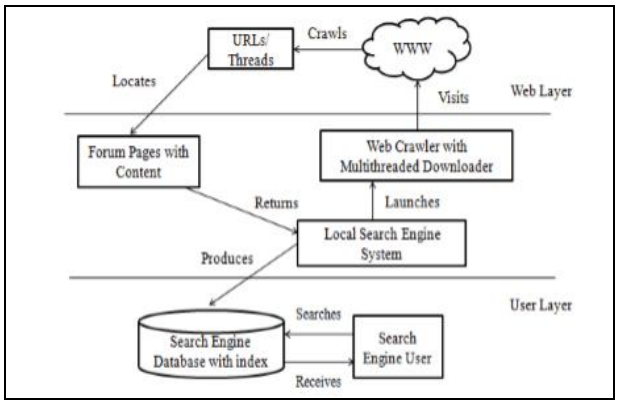

Fig 2: Web crawler with multithreaded downloader

Challenges in web forum crawling:

To harvest knowledge from forums, their contents have to be downloaded first. Generic crawlers which adopt a breadth first traversal strategy, are usually ineffective and inefficient for forum crawling.

This is mainly due to two non-crawler-friendly characteristics of forums:

1. Duplicate links & uninformative pages.
2. Page-flipping links.

IV. PROPOSED SYSTEM

Our crawler is a system that provides modules for crawling, indexing, sorting and searching of forum sites which are most commonly used.

The goal of our proposed system is to crawl relevant content, i.e., user posts, from forums with minimal overhead. Forums are always having implicit navigation paths to lead users from entry pages to thread pages. We show how to automatically learn regular expression patterns (ITF regexes) that recognize the index URL, thread URL, and page-flipping URL.

Here we show that a forum crawler should start crawling pages of forum sites from forum entry URLs. We show that, though the proposed approach is targeted at forum crawling, the implicit EIT-like path also apply to other User Generated Content (UGC).

Advantages of Proposed System:

- Avoids overhead and URL type reorganization problem.
- Provides highly precise index URL.
- No Data loss and Accuracy is maintained.
- Effective extraction of data

V. SYSTEM ARCHITECTURE

The System Architecture of Forum Crawler consists of two major parts as shown in Fig 3:

1) Learning part
2) Online crawling part.

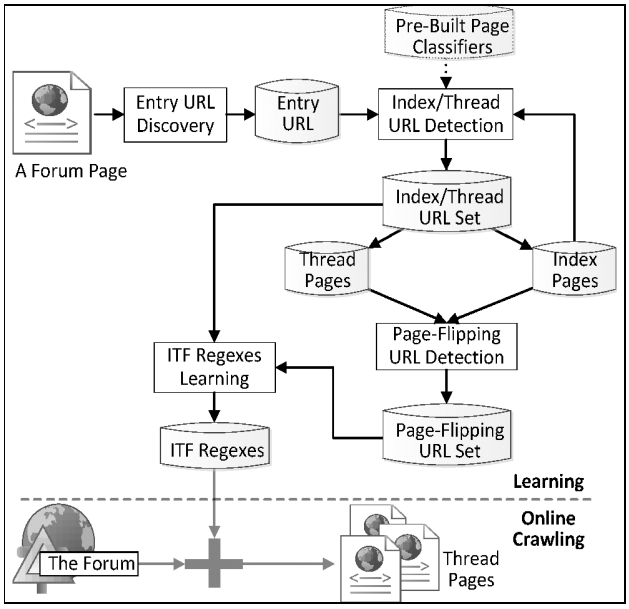

Fig 3: System Architecture

The major part in the system architecture is online crawling part that applies learned ITF regexes to crawl all thread pages efficiently. Using Entry URL Discovery module, our crawler finds its entry URL. The Index/Thread URL Detection module is used to detect index URLs and thread URLs on the entry page. Then, detected index URLs and thread URLs are saved to training set of URLs .

Next, to detect more index URLs, the destination pages of the detected index URLs are feed to this module again. The major task of Page-Flipping URL Detection module is to find out Page-flipping URLs in both index pages and thread pages and then add all that URLs to the training set. Finally, the ITF Regexes Learning module learns a set of regular expressions from the URL training set. It performs online crawling as follows: After pushing the entry URL into a URL queue it fetches a URL and downloads its page; then it pushes the outgoing URLs that are matched with any learned regular expression into the URL queue until URL queue gets empty.

VI. IMPLEMENTATION

We have divided our project in 5 Modules:

- Online Forum Crawling
- Page Classification
- Recognizing the URL
- Implementation of ITF Regex
- Experimental Analysis.

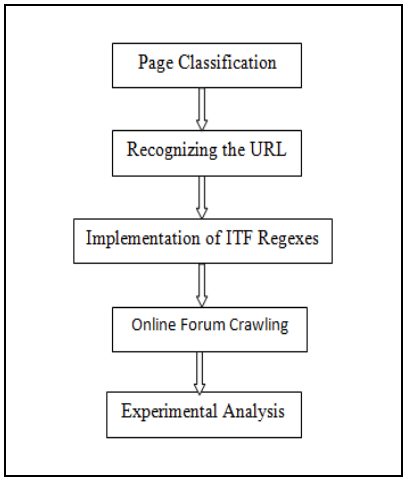

Fig 4: Implementation Plan

A. Page Classification:

The structure of forum sites varies from the structure of other sites. Forum sites consists of three major categories of pages named as

I. Entry Page
II. Index Page
III. Thread Page

Based on the properties of each type of page we will classify the pages.

B. Recognizing the URL:

After the classification of pages, we will now detect the different types of URLs using URL detection algorithms such as:

I. Entry URL detection algorithm.
II. Index/thread URL detection algorithm.
III. Page Flipping URL detection algorithm.

The detected URLS are stored for further processing.

C. Implementation of ITF Regexes:

The pattern of detected URL set is analysed and according to that we will generate ITF Regex (Index-Thread-Page Flipping) which is used to recognize index, thread or page Flipping URLs on EIT path i.e. Entry Index Thread path.

D. Online Forum Crawling:

Given a forum, our system will first learns a set of ITF regular expressions following the procedure mentioned above. Then it first performs online crawling using a breadth first strategy. After pushing the entry URL into a URL queue it fetches a URL and downloads its page; then it pushes the outgoing URLs that are matched with any learned regular expression into the URL queue until URL queue gets empty. Here system need not to group outgoing URLs, classify pages, detect page Flipping URLs or learn regexes again for that forum. Such time consuming operations are only performed during its learning phase. Here only apply the learned ITF regexes on outgoing URLs in newly downloaded pages.

E. Experimental Analysis:

To carry out meaningful evaluations that are good indicators of web-scale forum crawling, we will select different forum software packages. Among them, we will select some forums as our training set and leave the remaining for testing.

VI. OUTPUT

The homepage of our project is shown in the fig 5. It displays the links of two forum sites from which the user has to select any one at a time for crawling.

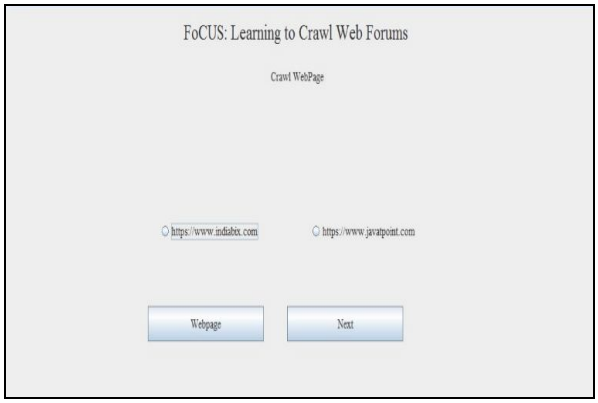

Fig 5: Home Page

The Fig 6 displays the list of thread URLs which are crawled.

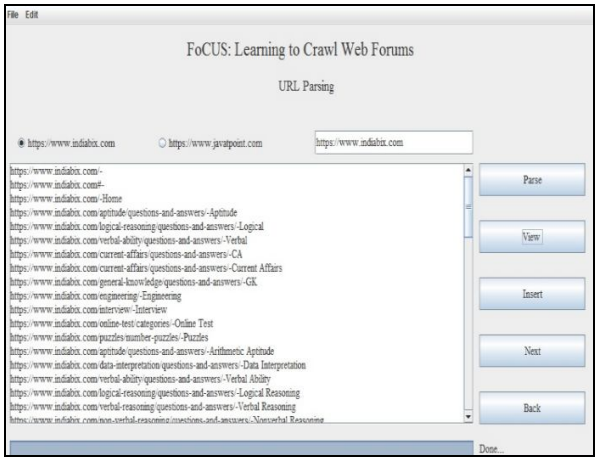

Fig 6: URL parsing

VII. CONCLUSIONS

We also showed that concept of FoCUS can effectively apply learned forum crawling knowledge of unseen forums. To automatically collect index URL, thread URL, and page-flipping URL training sets and learn ITF regexes from the training sets. These learned regexes can be applied directly in online crawling. Training and testing on the basis of the forum package makes our experiments manageable and our results applicable to many forum sites.

VIII. ACKNOWLEDGEMENT

In this project we have taken much of our efforts however, this was not possible without help of many individuals and organization. I am thankful to all of them. This research paper was made possible by the support of Dr. Vinayak Bharadi, HOD, IT Department, FAMT; Ms. Priyanka Bandagale, Project Guide. We would like to express our great gratitude to Ms. Priyanka Bandagale for her kind advice on the project and precious information.

REFERENCES

[1] Jingtian Jiang†* , Nenghai Yu , Chin-Yew, ”FoCUS: Learning to Crawl Web Forums”, IEEE Transactions on knowledge and data engineering vol:25 no:6, by 2013

[2] Priyanka Bandagale and Dr. Lata Ragha "Supervised Web Forum Crawling", International Journal of Engineering development and Research (IJEDR) Volume 4 Issue 1, ISSN: 2321-9939, by 2016

[3] T. MaharaJothi, K.Thirumoorthy ”A Survey on Web Forum Crawling Techniques”,IEEE International Conference on Innovations in Engineering and Technology (ICIET’14) .

[4] Priyanka Bandagale, Neha Sawantdesai, Rakshanda Paradkar and Piyusha shirodkar, “Survey on Effective web crawling technique” International Journal of Computer and Mathematical science(IJCMS) volume 6 Issue 10, ISSN:2347-8527, by 2017.

[5] Yan Guo, Kui Li, Kai Zhang and Gang Zhang , “A Web Crawling Method for Web Forum” 17th Int‟l Conf. World Wide Web, pp. 447-456, 2008.

[6] K.Vidhya and Ms .E.AnnalSheebaRanil ”A survey on crawling web forums”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2 Issue 11, November 2013

[7] M.Maheswari, N.Tharminie, “Crawler with Search Engine based Simple Web Application System for Forum Mining” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Mar-Apr. 2014

[8] V.Rajapriya, “FoCUS – Forum Crawler Under Supervision” by International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 3, Issue. 8, August 2014.

[9] Mrs. S. Nithyapriya and Dr. T. Kalaikumaran,Dr. S. Karthik “A survey on focused crawler for seed url selection” by , IJAICT Volume 1, Issue 8, December 2014.