© 2017 IJSRSET | Volume 3 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Forum Crawler

Rakesh S. Mane, Gopal B. Bagga, Devendra U. Bhute, Abhijeet D. Nikam, Prof. Sonali Gaikwad

D.Y.Patil COE, Pune, Maharastra, India

ABSTRACT

A generic web crawler can be efficient in crawling the web but it is not efficient when crawling a forum. While crawling any forum the generic crawler will crawl all pages including unnecessary pages. Also generic crawlers can't maintain relation between posts of different pages. Existing forum crawlers are not easy to configure and requires too much user interaction. That's why a new type of crawler is needed for efficient forum crawling. This system aims to crawl only relevant pages from the forum with minimal overhead. For achieving that this system uses signature based approach for generating regular expressions of relevant pages URL's. Different forum softwares have different page layout but the navigation paths are mostly similar. By generating regex for relevant paths this system makes sure that it only crawls relevant pages. For generating regex first the forum software is identified using predefined signatures then predefined URL patterns are selected for identified forum software. This patterns are used to generate regex for Thread and Index pages URL's. Flipping URL's are identified by using predefined signatures. This approach allows accurate and faster crawling of forums with minimum configuration.

Keywords : Forum Crawling, Page Classification, Page Type, URL Pattern Learning, URL Type, Signatures

I. INTRODUCTION

Due to use of internet the huge amount of data is available on forums that contains very useful information. But existing systems cannot collect this data efficiently. Hence a new type of crawler is needed. Our project talks about such a new type of forum crawler. It crawls only index and thread pages to achieve efficiency and speed.

We have predefined URL Patterns for some forum softwares. Those URL patterns will be used for generating regular expressions. These regular expressions will be used while crawling to make sure only thread and index URL's will be crawled.

In order to select correct URL pattern while generating regular expressions our system first needs to detect the forum software. We have predefined body and cookies based signatures of few forum softwares. For detecting forum software it uses predefined signatures.

Currently Supported Forum Softwares:

1) phpBB
2) Discuz
3) MyBB
4) FluxBB

Remove =>

7.3.3 Temporary data structure
7.3.1 Internal software data structure
7.3.2 Global data structure

A. Web Crawler

A web crawler (also known as a web spider) is a program or an automated script which browses the World Wide Web in a methodical, automated manner. This process is called Web crawling. Many search engines use spidering as a means of providing up-to-date data.

A Web crawler starts with a list of URLs to visit. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. The large volume implies that the crawler can only download a limited number of the Web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that the pages might have already been updated or even deleted which results in generation of overhead.

I. LITERATURE SURVEY

FoCUS is a supervised web scale forum crawler which crawls relevant forum content with minimum overhead. The forum crawling problem is reduced to URL type recognition problem by using ITF-regex which specifies best navigation path by using training sets which are created automatically form page type classifiers. The goal of FoCUS is to crawl relevant forum content from the web with minimal overhead. Forum threads contain information content that is the target of forum crawlers. Although forums have different layouts or styles and are powered by different forum software packages, they always have similar implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages. Based on this observation, the web forum crawling problem is reduced to a URL-type recognition problem and classifies them as Index Page, Thread Page and Page-Flipping links [1].

iRobot first randomly samples (downloads) a few pages from the target forum site and introduces the page content layout as the characteristics to group those pre-sampled pages and reconstruct the forum sitemap. After that, it selects an optimal crawling path which only traverses informative pages and skips invalid and duplicate ones. The main idea of iRobot is to first learn the sitemap of a forum site with a few pre-sampled pages and then decide how to select an optimal traversal path to avoid duplicates and invalids. First, to discover the sitemap, those pre-sampled pages are grouped into multiple clusters according to their content layout and URL formats. In this part, it proposes a repetitive region-based layout clustering algorithm, which has been proven to be robust in characterizing forum pages. Then, the informativeness of each cluster is automatically estimated and an optimal traversal path is selected to traverse all the informative pages with a minimum cost. The major contribution in this step is to describe the traversal paths with not only their URL patterns but also their locations of the corresponding links on page layout. In such a way, it can provide a more strict discrimination between links with similar URL formats but different functions [2].

Board Forum Crawling presents a new method of Board Forum Crawling to crawl a Web forum. It first extracts all URLs from board pages then from each of these URLs it again extracts all subsequent board pages. Now, it downloads each of those subsequent pages and identifies whether it is exactly a board page and extracts links of post pages and saves them in a list. Later, all the links from that list are used to download all post pages and save those [4].

Web Forum Crawling proposes a system in which the crawler first re-constructs the sitemap of forum based on a few thousands pages randomly sampled from the target forum. The proposed solution mainly consists of the identification of skeleton links and the detection of page-flipping links. The skeleton links instruct the crawler to only crawl valuable pages and avoid duplicate and uninformative ones and the page-flipping links tell the crawler how to completely download a long discussion thread which is usually shown in multiple pages in Web forums [3].

The **GoGetIt** system takes a sample page and entry page URL of the website. In first phase, it follows all paths looking for the pages that matches the structure of the sample page and generates a TPM tree. TPM is nothing but the minimum spanning tree that represents the all minimum paths to reach the pages that match structure of provided sample page from entry page. In the second phase regular expressions are generated based on TPM tree. These regular expressions only match to the path which goes to the pages that match the structure of the given sample page [5].

II. TERMINOLOGY

PAGE TYPE

It classified forum pages into page types.

Entry Page: The homepage of a forum is contains a list of boards and is also the lowest common ancestor of all threads.

Index Page: A page of a board in a forum, which usually contains a table-like structure; each row in it contains information of a board or a thread.

Thread Page: A page of a thread in a forum that contains a list of posts with user generated content belonging to the same discussion.

Other Page: A page that is not an entry page, index page, or thread page.

URL TYPE
There are four types of URL.

Index URL: A URL is on an entry page or index page and points to an index page. Its anchor text shows the title of its destination board.

Thread URL: A URL is on an index page and points to a thread page. Its anchor text is the title of its destination thread.

Page-flipping URL: A URL leads users to another page of the same board or the same thread. Correctly dealing with page-flipping URLs enables a crawler to download all threads in a large board or all posts in a long thread.

Other URL: A URL that is not an index URL, thread URL, or page-flipping URL.

Cookie Signature : Unique cookies of forum softwares

Body Signature : Unique html code of forum softwares

III. SYSTEM ARCHITECTURE

In this system the entry URL and the depth of flipping pages are taken from the user as input. The entry URL is then fetched and it is checked for signatures. Signatures are the predefined patterns that are unique for the different forum softwares. Signatures are cookies based and html body based. Once the entry URL is fetched the forum software is detected based on matched signatures. After forum software is identified the system takes predefined index and thread URL patterns of the detected forum software. Using these URL patterns it generates regular expressions for the index and thread URL.

Once the regular expressions are generated crawling process is started from the entry page. All links from each page is extracted and matched with regular expressions. URL's matching to index regex are first checked for duplication then added to the index url queue. URL's matching to thread regex are first checked for duplication then added to the thread url queue and the fetched. Thread pages contents of all flipping URL's is extracted and stored in the database. For detecting flipping URL's all the predefined signatures are used.
After getting records in the database it gets displayed to the user while crawling is still done in background and also records are updated in database simultaneously.

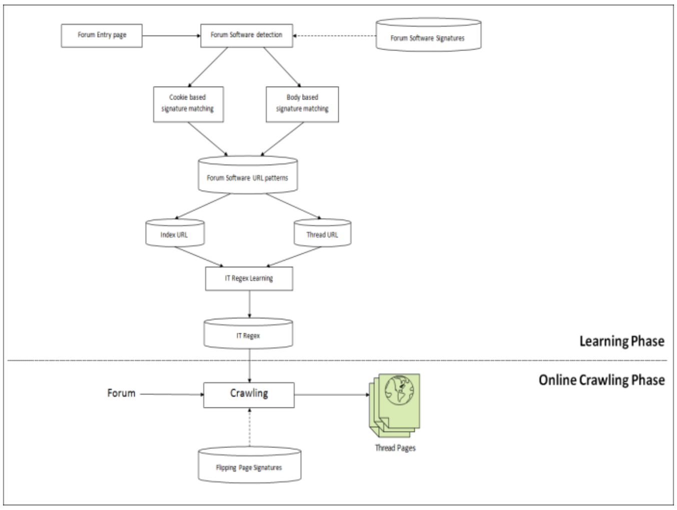

Figure 1. The overall architecture of FoCUS

IV. ALGORITHM

Detecting Forum Software :
1) Fetch the entry page of forum
2) Check for predefined signatures in the fetched page
3) Decide forum software based on signatures found in the fetched page

Generating Index URL regex :
1) Select predefined Index URL pattern of detected forum software
2) Generate regular expressions that matches selected Index URL pattern

Generating Thread URL regex :
1) Select predefined Thread URL pattern of detected forum software
2) Generate regular expressions that matches selected Thread URL pattern

Filtering URLs :
1) Check if the newly obtained URL matches with the Index or Thread URL regex
2) Ignore URL's that does not matches Index or Thread URL regex
3) Add all newly obtained thread and index URL's only if they are not already in the list

Crawling Index Pages :
1) Fetch all index URL's from the queue sequentially.
2) Extract newly obtained index URL's from fetched index pages and add it in queue

Crawling Thread Pages :
1) Fetch all thread URL's from the queue sequentially.

2) Extract newly obtained thread URL's from fetched thread pages and add it in queue
3) Fetch thread URL and add it's content to database

Crawling Flipping URL's :
1) Using predefined signatures of detected forum software extract flipping URLs
2) Fetch all extracted URLs and extract contents and add it's content to database

V. CONCLUSION

Here we developed a signature based forum crawler. We designed methods to generate regular expressions of index and thread URLs of forum. Experimental results on 5 widely used forum software packages confirm that our system can effectively crawl forums. Our system automatically detects forum software of the website using predefined signatures and selects predefined URL patterns of detected forum software for generating regular expressions of Index and Thread Urls. By following only URLs that matches with the generated regexes we can ensure that it only follows index and thread pages once. Flipping URLs are also extracted using predefined signatures.

VI. REFERENCES

[1]. Jingtian Jiang, Xinying Song, Nenghai Yu and Chin-Yew Lin, FoCUS: Learning to crawl web forums. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 6, JUNE 2013.

[2]. R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, iRobot: An Intelligent Crawler for Web Forums, Proc. 17th Intl Conf. World Wide Web, pp. 447-456,April-2008

[3]. Y. Guo, K. Li, K. Zhang, and G. Zhang, Board Forum Crawling: A Web Crawling Method for Web Forum, Proc. IEEE/WIC/ACM Intl Conf. Web Intelligence,pp. 475-478, 2006.

[4]. Y.Wang, J.-M. Yang, W. Lai, R. Cai, L. Zhang, andW.-Y. Ma,Exploring Traversal Strategy for Web Forum Crawling, Proc. 31st Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 459-466, 2008.

[5]. Mrcio L.A. Vidal, Altigran S. da Silva, Edleno S. de Moura, Joo M. B. Cavalcanti, GoGetIt!: A Tool for Generating Structure-Driven Web Crawlers, May-2006