© 2017 IJSRSET | Volume 3 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Review-FoCUS: Learning to Crawl Web Forums

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 websites but it is not efficient when crawling a web forum. While crawling any forum, the generic crawler will crawl all pages including unnecessary pages like user profile pages, advertisement pages or redirection pages which might result in duplication. That's why a new type of crawler is required for efficient forum crawling. This system will crawl only relevant contents from the forum with minimal overhead and maximum efficiency. Although different kinds of forums have different page layouts, they always have similar indirect navigation paths connected by specific URL types to lead users from entry pages to thread pages. This property of forums is observed and forum crawling problem is reduced to URL-type recognition problem in order to follow only useful (Thread, Index and Page-Flipping pages) URLs and ignore unnecessary (User profile, External links) URLs. To recognize the URL types, the ITF-regex (that matches only Index, Thread and Page Flipping URLs) is learned by using the URL training sets. URL training sets just contains the detected URLs of thread, index and page flipping pages. To detect the kind of URL, differentiate and detect thread, index and page flipping URLs. The common characteristics of those pages are used to detect the page type.

Keywords : EIT path, forum crawling, ITF-regex, page classification, page type, URL pattern learning, URL type

I. INTRODUCTION

An online forum is a web application for holding discussions and posting user generated content in a specific domain, such as sports, recreation, techniques, travel etc. Forums contain a huge amount of valuable user generated content on a variety of topics and it is highly desirable if the human knowledge contained in user generated content in forums can be extracted and reused.

To harvest knowledge from forums, their content must be downloaded. However, forum crawling is not a minor problem. Generic crawlers are usually ineffective and inefficient for forum crawling. This is mainly due to two non crawler friendly characteristics of forums 1) duplicate links and uninformative pages and 2) page-flipping links. A forum typically has many duplicate links that point to a common page but with different URLs e.g. shortcut links pointing to the latest posts or URLs for user experience functions such as "view by date" or "view by title." A generic crawler that blindly follows these links will crawl many duplicate pages, making it inefficient and resulting in generation of unnecessary data. A forum also has many uninformative pages such as login control to protect user privacy or forum software specific FAQs or advertisement pages. Following these links, a crawler will crawl many uninformative pages.

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.

B. Forum Crawler

Forum Crawler is a supervised web-scale crawler to crawl forums and address the challenges in crawling such as avoiding duplicate links, uninformative pages and crawling only relevant pages which cannot be done efficiently by typical web crawler.

II. 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].

III. 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.

EIT Path: An entry-index-thread path is a navigation path from an entry page through a sequence of index pages (via index URLs and index page-flipping URLs) to thread pages (via thread URLs and thread page-flipping URLs).

ITF Regex: An index-thread-page-flipping regex is a regular expression that can be used to recognize index, thread, or page-flipping URLs. ITF-regex is what FoCUS aims to learn and applies directly in online crawling. The learned ITF-regexes are site specific, and there are four ITF-regexes in a site: one for recognizing index URLs, one for thread URLs, one for index page-flipping URLs, and one for thread page-flipping URLs. A perfect crawler starts from a forum entry URL and only follows URLs that match ITF-regexes to crawl all forum threads. The paths that it traverses are EIT paths.

IV. SYSTEM ARCHITECTURE

The fig.1 shows the overall architecture of FoCUS. It consists of two major parts: the learning part and the online crawling part. The learning part first learns ITF regexes of a given forum from automatically constructed URL training examples. The online crawling part then applies learned ITF regexes to crawl all threads efficiently. Given any page of a forum, FoCUS first finds its entry URL using the Entry URL Discovery module. Then, it uses the Index/Thread URL Detection module to detect index URLs and thread URLs on the entry page; the detected index URLs and thread URLs are saved to the URL training sets. Next, the destination pages of the detected index URLs are fed into this module again to detect more index and thread URLs until no more index URL is detected. After that, the Page-Flipping URL Detection module tries to find page-flipping URLs from both index pages and thread pages and saves them to the training sets. Finally, the ITF Regexes Learning module learns a set of ITF regexes from the URL training sets. Once the learning is finished, FoCUS performs online crawling as follows: starting from the entry URL, FoCUS follows all URLs matched with any learned ITF regex. FoCUS continues to crawl until no page could be retrieved or other condition is satisfied.

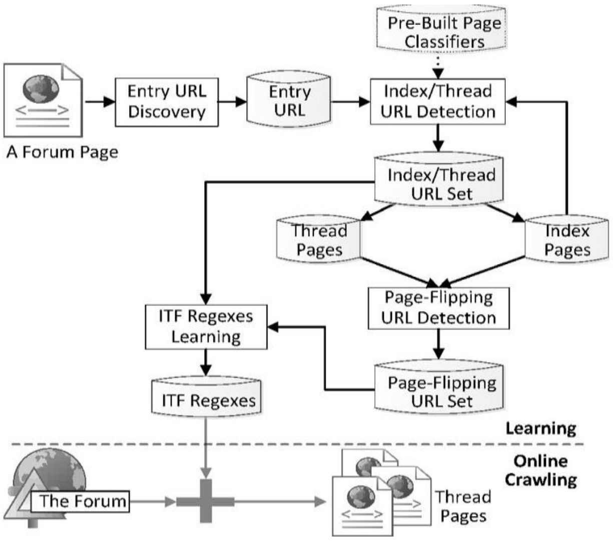

Figure 1. The overall architecture of FoCUS

V. CONCLUSION

The forum crawling problem is reduced to a URL type recognition problem and portrayed how to leverage implicit navigation paths of forums, i.e., EIT path, and designed methods to learn ITF-S regexes explicitly. FoCUS can learn about knowledge of EIT paths from very few type of forums and then can apply this learned knowledge on large set of unseen forums to collect Index URLs, Thread URL's and page flipping URL's training sets and learn ITF-regexes from these training sets. After this, the FoCUS will crawl only those pages whose URLs will match the Pattern described by the ITF-regex resulting in reduction of overhead and duplication. Unlike other crawlers, FoCUS does not expect an entry URL and can start from any page of the forum. It also managed to outperform iRobot and hence proving that it is the most efficient forum crawler.

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.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.

[4] 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.

[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