OAIJSE

OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

SURVEY ON FORUM ASSESSMENT SYSTEMS

Bhushan B. Zade[1], Dhiraj B. Kate[2], Prachi M. Autade[3], Prajakta V. Dhage[4], Prof. Sharad S. Adsure[5]

B. E. Students, Dept. of Computer Science, JSPM'S BSIOTR, Wagholi, Pune, Maharashtra, India[1,2,3,4]

Assistant Professor, Dept. of Computer Science, JSPM'S BSIOTR, Wagholi, Pune, Maharashtra, India[5]

Abstract: **A nonexclusive web crawler can be proficient in crawling the web however it isn't productive when creeping a gathering. While crawling any discussion the non specific crawler will creep all pages including pointless pages like client profile pages. That is the reason another kind of crawler is required for effective discussion crawling. This system introduces a gathering crawler which can crawl just pertinent substance from the forum with negligible overhead. Albeit distinctive gatherings have diverse page formats they generally have comparable circuitous route ways associated by particular URL sorts to lead clients from entry pages to thread pages. This property of gatherings is observed and forum crawling issue is decreased to URL-sort acknowledgment issue so as to take after just valuable (Thread, Index and Page-Flipping pages) URLs and disregard superfluous (User profile, External links)URLs. To perceive the URL type, the ITF regex (that matches just Index Thread and Page Flipping URLs) is found out utilizing the URL training sets. URL training sets just contains the identified URLs of thread, index and page flipping pages. To identify the URL separate and recognize thread, index and page flip-ping URLs the common qualities of those pages are used. On the off chance that user not fulfils with showed result or for any inquiry he may ask expert user.**

Keywords: *EIT path, forum crawling, ITF regex, page classification, page type*

I INTRODUCTION

Web forums are imperative stages where user can demand and exchange data with others. For example, the Trip Advisor Travel Board is where individuals can ask and share travel tips. Because of the richness of data in forum, specialists are progressively intrigued by mining information from them. We introduce Forum Assessment System, a directed web-scale forum crawler, to address existing difficulties. The objective of this system is to trawl applicable substance, i.e. user posts, from forums with negligible overhead. Forums exist in a wide range of layouts or styles and powered by an assortment of forum software bundles, however they generally have understood route ways to lead users from entry pages to thread pages. We call pages between the entry page and thread page which are on a breadth-first route way the index page. Connections between a entry page and a index page or between two index pages are referred as list URLs. Connections between a index page and a thread page are referred as thread URLs. Connections interfacing numerous pages of a board and different pages of a thread are referred as page-flipping URLs. A crawler beginning from the entry page of a forum just needs to take after index URLs, thread URLs, and page-flipping URLs to navigate EIT path and accomplish all thread pages. The test of gathering crawling is then lessened to a URL type acknowledgment issue.

II RELATED WORK

- **Jingtian Jiang, Xinying Song, Nenghai Yu and Chin-Yew Lin, FoCUS:**

  It is a supervised web scale forum crawler it crawls relent 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.

- **R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, iRobot:**

It 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, author 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.

- **Y. Guo, K. Li, K. Zhang, and G. Zhang, Board Forum Crawling:**

Author presents a new method of Board Forum Crawling to crawl Web forum. Author first extracts all URLs from board pages then from each of this URL 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 links from that list are used to download all post pages.

Sr. No.	Paper Name	Author	Year	Advantage
1	Board Forum Crawling: A Web Crawling Method for Web Forum	Yan Guo Kui Li Kai Zhang Gang Zhang	2010	This method exploits the organized characteristics of the Web forum sites and simulates human behaviour of visiting Web Forums.
2	FoCUS: Learning to Crawl Web Forums	Jingtian Jiang, Nenghai Yu Chin-Yew Lin	2012	The goal of FoCUS is to only trawl relevant forum content from the web with minimal overhead.
3	Exploring Traversal Strategy for Web Forum Crawling	Yida Wang, Jiang-Ming Yang, Wei Lai, Rui Cai, Lei Zhang, and Wei-Ying Ma	2008	It enables the crawler to completely download a discussion thread. A long thread may consists of tens or even hundreds of pages, most of which are missed in a generic crawling as their link depths are too deep;
4	GoGetIt!: A Tool for Generating Structure Driven Web Crawlers	M´arcio L.A. Vidal, Altigran S. da Silva, Edleno S. de Moura, Jo˜ao M. B. Cavalcanti	2006	The valuable information these pages implicitly contain to perform such tasks as querying, searching, data extraction, data mining and feature analysis.
5	iRobot: An Intelligent Crawler for Web Forums	Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang	2008	Effectiveness. iRobot can intelligently skip most invalid and duplicate pages, while keep informative and unique ones.

- **Y. Wang, J. M. Yang, W. Lai, R. Cai, L. Zhang Web Forum Crawling:**

  Author proposes the system which 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 meanwhile 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.

- **Mrcio L.A. Vidal, Altigran S. da Silva, Edleno S. de Moura, ”GoGetIt!:**

  This 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 Target Pages Map (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. This regular expressions only matches to the paths which goes to the pages that matches structure of the given sample page.

III SYSTEM ARCHITECTURE

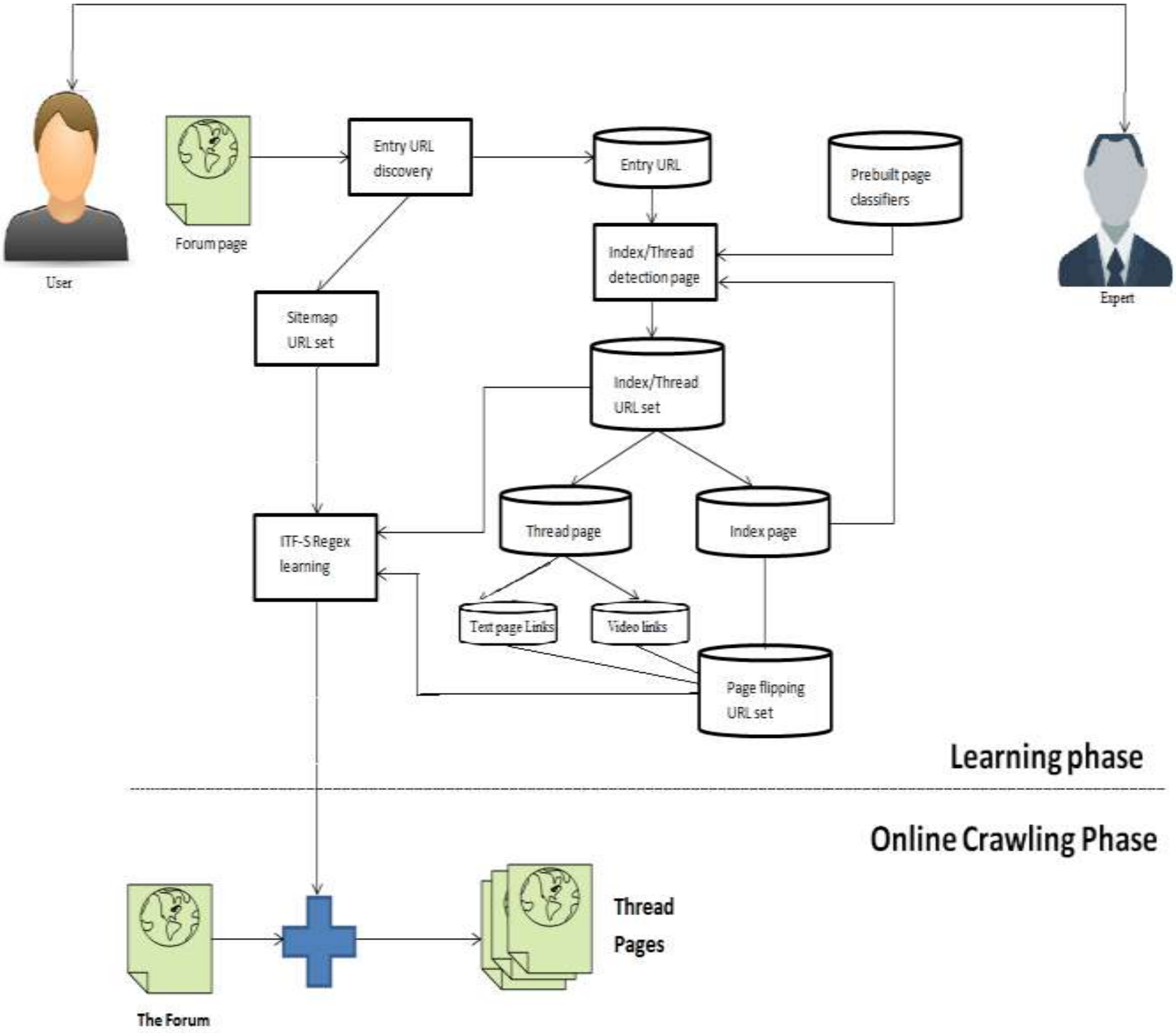

Figure 1: System Architecture

IV METHODOLOGY

Modules:

1. Constructing URL Training Set

It automatically create sets of highly precise index URL, thread URL, and page-flipping URL string samples for regex learning.

2. Learning ITF Regexes

It creates a URL regular expression pattern. Each pattern matches a subset of URLs. These patterns are refined recursively until no more specific patterns could be generated. These patterns are final output as they cannot be refined further. a refined pattern is retained only if the number of its matching URLs is greater than an empirically determined threshold.

3. Expert Chatting

After the learning and crawling part results will stored to database and displays to user. If user wants expert suggestion then he can chat with expert.

4. Online Crawling

Online crawling is done using a breadth-first strategy. It first pushes the entry URL into a URL queue; next it fetches a URL from the URL queue and downloads its page; and then it pushes the outgoing URLs that are matched with any learned regex into the URL queue. Forum Assessment System repeats this step until the URL queue is empty or other conditions are satisfied.

Applications

- Forum websites.
- QA sites

V CONCLUSION AND FUTURE WORK

The forum crawling problem is reduced to a URL type recognition problem and portrayed how to leverage implicit navigation paths of forums. The proposed system proved that Forum Assessment System is the most efficient forum crawler among all the currently existing crawlers.

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,

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

[5] Ali Bahrami, Chapter 6, Object Oriented Analysis Process, in Object Oriented System Development.

[6] H. M. Harmain and R. Gaizauskas, CM-Builder: An Automated NL Based CASE tool, in IEEE International Conference on automated software engineering (2000)

[7] Mich L., NL-OOPS: From natural language to object oriented requirement using natural language processing system (1996)

[8] Overmyer, S. P., Benoit, L. and Owen R., Conceptual modeling through linguistic analysis using LIDA. International Conference of Software Engineering (ICSE), (2001)

[9] Hector G perez-Gonzalez and Jugal K. Kalita, GOOAL : A Graphical Object Oriented Analysis laboratory, ACM 1-58113-626-9/02/0011 (2002)