A NOVEL APPROACH TO CRAWL WEB FORUMS

A Deepthi[1], Manda Ashok Kumar[2], Betam Suresh[3]

[1]M.Tech (CSE) Scholar, Vikas Group of Institutions, Nunna, Vijayawada, A.P, (India)

[2]Asst. Professor, Department of CSE, Vikas Group of Institutions, Nunna, Vijayawada, (India)

[3]HOD, Vikas Group of Institutions, Nunna, Vijayawada, A.P, (India)

ABSTRACT

These relationships should be conserving in this project, we represent FoCUS (Forum crawl under supervision), a supervision web-scale forum crawler. The goal of FoCUS is to only get relevant forum content from the web with less overhead. Forum threads having the information content that is the target of forum crawlers. Even though forums have different layouts or styles and are powered by different forum software packages, they always have similar entire navigation paths connected to specific URL types to lead users from main pages to thread pages. Based on this concept, we reduce the web forum crawling problem to URL type identification. Here only the trusted or registered user having the permission to download the web forum, means we providing the security to web forum means this is applicable for some proprietary web forums. The register candidate only download the main page or crawling page by specifying a corresponding key associated with the file.

Keywords— *Crawl, Supervision, Web Forum, URL Pattern, Page Type.*

I. INTRODUCTION

Web forums are important platforms where users can request and exchange the advice with others. For example, the trip Advisor Travel Board is a place where people can ask and share travel tips. Because of the richness of information in forums, researchers are very interested in mining knowledge from them. According to an article from eMarketer - Where Are Social Media Marketers Seeing the Most Success? – The Forums are still part of the global social media strategy of the Top 600 Companies, and they are still getting really highly marketing success with web forums. To get the awareness from forums, their contents have to be downloaded first. A forum usually may have many duplicate links which point to a same page but with different URLs, e.g., shortcut links pointing to new posts or URLs for user experience functions such as "view by the banner". A universal that follows these links will trawl to many duplicate pages that makes inefficient. A Forum typically has many uninformative pages like login control to protect users' isolation. Following links, a crawler will trawl many uninformative pages. Though there are standard-basic methods such as describing the "rel" attribute with "nofollow" value (i.e. "rel=nofollow") 2 for forum operators to guide web crawlers on how to crawl a site effectually, we find that over a set of 9 test forums more than 50% of the pages trawled by a generic crawler following these protocols are duplicate or uninformative. This number is a little bit higher than the 40% reported but both show the inefficiency of generic crawlers.

In addition to duplicate links & uninformative pages, a long forum thread is usually divided into multiple pages these are linked by page-flipping links. Generic crawler processes each page individually and ignores the relationship between while crawling to facilitate downstream tasks such as page wrapping and content indexing.

For example, multiple pages belongs to a thread should be concatenated together in order to extract all posts of this thread as well as the reply relationships between posts.

In addition to the above challenges, there is also having the problem of entry URL discovery. A forum's entry URL points to its home page, which is the lowest common antecedent page of all threads. But entry URL discovery is not a insignificant problem. An entry URL is not necessary at root URL level of a forum hosting site and its form differ from site to site. Without entry URLs, extant crawling methods such as Vidal et al and Cai et al are less effective.

But in the existing system there is no security to download the content, this is the draw-back of nowadays anybody can download the content and share the content with others, Some malicious persons intentionally upload some illegal content or they can change original content.

In this paper we are introducing a concept FoCUS (Forum crawler under supervision), a supervision web-scale forum crawler. The goal of FoCUS is to only get relevant forum content from the web with less overhead. Forum threads having the information content that is the target of forum crawlers. Even though forums have different layouts or styles and are powered by different forum software packages, they always have similar *entire navigation paths* connected to specific URL types to lead users from main pages to thread pages. Download the content with key, that key also encrypted and store in the data base. Whenever the admin upload a file the key will goes to persons who are register to this site, they only the trusted persons tom get the details about the particular file.

The major improvement of this paper is as follows:

1. We reduce the forum crawling problem to a URL type identification problem and implements a crawler, FoCUS, to demonstrate its applicability.
2. We represent how to learn automatically regular expression patterns that identify the index URL and thread URL and page-flipping URL using the page classifiers built from as few as 5 annotated forums.
3. We evaluates FoCUS on a large set of 160 unseen forum packages that covers 668,683 forum sites. To the knowing of our knowledge, this is the largest scale evaluation of this type. In reserve, we expose that the patterns are effective and the resulting crawler is efficient.
4. We compare FoCUS with the baseline generic breadth-first crawler, it's a structure driven crawler, and a state of the art child iRobot and shows that FoCUS outperforms these crawlers in terms of potency and coverage.
5. We design an effective forum URL entry discovery method. Entry URLs need to be describes to start crawling to get higher recollect. But entry page discovery is not a trivial task since entry.

II. RELATED WOK

In the previous they proposed a method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. The Target pages were found through comparing DOM trees of pages with the pre-selecting sample target page. It is very effective but it will works for only specific site from which the sample page is drawn. The same process will be repeated every time for the new site. Therefore, it is not suitable for the large- scale crawling. In divergence, FoCUS learns URL patterns across multiple sites and automatically finds forum entry page from given a page forum. An experimental result specifies that FoCUS is an effective in large scale forum crawling by leveraging crawling knowledge learned from a few annotated forum sites.

A recent and more effective work on forum crawling is iRobot by Cai et al. The aim of iRobot is automatically learn a forum crawler with the minimum human intervention by sampling forum pages, grouping them, selecting informative groups via an informativeness measure, finding a traversal path by a spanning tree algorithm. Anyway, the traversal path selection procedure requires human inspection. Proposed an algorithm to address the traversal path selection problem. However, corresponding to our appraisal, its sampling strategy and informativeness expectation is not always robust and its tree-like traversal path does not allow more than one path from a starting page node to common ending page node. For example, there are 6 paths from entry page to thread page. But iRobot will take only the first path (entry board thread). iRobot always learns URL location information to discover new URLs in crawling, but the URL location might become invalid when the page structure changes. Compare with iRobot, we explicitly define EIT (entry-index-thread) paths. FoCUS leverages page layouts to identify index pages and thread pages. FoCUS learns precise URL string patterns instead of URL locations to invent new URLs. Thus it does not need to classify new pages in crawling and would not be affected by a change in page structures. The corresponding results between iRobot and FoCUS demonstrated that the EIT path and URL string patterns are more robust than the traversal path and URL location feature in iRobot. By identifying and only following skeleton links and page-flipping links, they showed that iRobot can achieve good effectiveness and coverage.

Another related work is near-duplication detection. Forum crawling also needs to be removes duplicates. But content-based duplication detection is not bandwidth-efficient, because it can be only carried out when pages have been downloaded. URL-based duplicate detection is not helpful. It tries to reserve rules of different URLs with similar text. But such methods still need to analyze logs from target sites or results of a previous crawl. In this paper, through detecting index, thread and page-flipping URLs, FoCUS could avoid duplicates without duplicate detection and down load a page with trusted parties while downloading a page they need to enter the key, if the key is correct only the page will download otherwise it will give error saying invalid key

III. FOCUS – A SUPERVISED FORUM CRAWLER

3.1 Observations

In order to crawl forum threads effectiveness and efficiency, we investigated about 30 forums (not used in testing) and found the following features in almost all of them:

3.1.1 Navigation Path

Despite differences in layout and styles, forums always having similar implicit navigation paths leading users from their entry pages to thread pages. In general web crawling, “navigation patterns” leading to target pages. iRobot also adopted similar idea but applied page sampling and clustering techniques to find target pages. It used

3.1.2 URL Layout

Information URL layout such as the location of a URL on a page and its anchor text length is the important indication of its function. URL of the same function usually appears at the same location.

3.1.3 Page Layout

The index pages from different forums have share same layout. That same applies to thread pages. However, an index page usually has very different page layout from a thread page. An index page lean to have many narrow

records giving information about boards or threads. Thread pages has typically a few large records pages together and apply its informativeness metric to decide whether a set of pages must be crawled. FoCUS learn page type classifiers directly from a set of annotated pages based on this characteristic. This is the only part where manual annotation is requires to FoCUS.

By the inspiration of these considerations, we develop Focus. The main aim behind this FoCUS is that index URL, thread URL, and page flipping URLs can be finding based on their layout attributes and destination pages; and forum pages can be confidential by their layouts. This awareness about URLs, pages and forum structures can be learned from a few annotated forums and then applied to unseen forums.

3.2 System Overview

Given any page of a forum, FoCUS first finds its entry URL working *Entry URL identification* module. Then, it uses the *Index/Thread URL Detection* module to identify index URLs and thread URLs on the entry level page; the identified index URLs and thread URLs are saved to the URL readying set. Next, the target pages of the identified index URLs are grub to this module again to identify more index URLs and thread URLs up to no more index URL disclosed. After that, the *Page-Flipping URL distinguish* module tries to find page-flipping URLs in both index pages and thread pages and saves them to the coaching set. Finally, the *ITF Regexes learning* modules learn a set of ITF regexes from the URL readying set.

FoCUS meet online crawling as follows: it first forcing the entry URL into a URL queue; next it getting a URL from the queue and download its page, and then pushes the outgoing

URLs those are same with any learned ITF regex into the URL queue. This step is repeated up to the URL queue is empty.

3.3 Online Crawling

Given a forum, FoCUS first knows a set of ITF regexes. Then it performs online crawling by using a breadth-first strategy. It first forcing the entry URL into a URL queue; next it getting a URL from the queue and download its pages, and then pushes the warm URLs those are same with any learned ITF regex into the URL queue. This step is repeated up to the URL queue is empty or other conditions are satisfied. To why FoCUS efficient in online crawling is that it only wishes to apply the learned ITF regexes on outgoing URLs in freshly downloaded pages. FoCUS does not need to cluster outgoing URLs, classify pages, distinguish page-flipping URLs, or learn regexes anew for that forum. Such time engrossing operations are only performed during its learning phase.

3.4 Entry URL Discovery

In the past sections, we explained how FoCUS learns ITF regexes that could be used in online crawling to resolve what URLs to follow and what URLs to avoid. However, any entry

page needs to be described to start the crawling process. To the best of our awareness, all preceding methods assumed a forum entry page is given. In process, especially in web-scale crawling, manual forum entry page annotation is not pragmatic. Forum entry page disclosure is not a trivial task since entry pages differ from forums to forums. Our experiment displays that a naïve base method can reached only about 76% recall and precision. To make FoCUS very practical and scalable in web-scale crawling, we design a simple yet effective forum entry URL discovery method based on some techniques disclosure in past sections.

We observe every page contains a link to navigate to next page and may be that page also having the link to next page, we can come to main page by clicking back. This will happen by using an algorithm breadth-first search, but in our proposed system if the user want to down load a page or share a page he must login in to his account and if he want to download a page he should give a corresponding key to a file if the key matches it download the file otherwise it gives the error message to the user. Here we providing.

IV. EXISTING SYSTEM

In the existing they are using ITF Regex and online crawling to quickly get the pages, here we provide links in main page to navigate from one page to another page, and here they are not providing any security to the publishers who are publishing the pages anybody can download the page and anybody can share the details with others this not good for credible users. Because of this there may be a chance to corrupt the original data so the users may get blunder results. To defeat this in over paper we present a concept called FoCUS with a encrypted key.

V. PROPOSED SYSTEM

In this paper we propose a technique to down load a file/page with key and this key is encrypted key, to encrypt this we use plane cipher algorithm. Before store in the data base, we encrypt the key and send to the register users. The forum has very useful content. So many software packages have these forums. All these forums uploaded by the admin at the time of uploading a forum he send the corresponding key to the register users. To provide the security to the pages the pages also to be encrypted and stores in the database and whenever user downloading the page automatically that decrypt and the original content shown to the user. After that user can share the data and he can download the forum.

VI. CONCLUSION

In this paper, we proposed and implemented FoCUS with encrypted key, a supervised forum crawler and download forum using key. By using this we reduced the forum crawling problem to a URL type recognition problem and showed how to import implicit navigation paths of forums, i.e. entry-index-thread path, and designed methods for learning ITF regexes explicitly. Experimental results on 150 forum sites each of them powered by a different forum software packages confirm that FoCUS can efficiently learn knowledge of EIT path and ITF regexes from as few as 5 expound forums. We also exposed that FoCUS can effectively apply learned forum crawling knowledge on 160 unseen forums is to automatically collect index URL and thread URL and page-flipping URL string training sets and learn the ITF regexes from the training sets. These readying regexes could be applied directly in online crawling. Training and examination on the basis of forum package makes our experiments manageable and our results applicable to lot of forum sites. Moreover, FoCUS can start from any page, while all previous works expect an entry page is given. Our test results on 8 unseen forums show that FoCUS is absolutely very effective and efficient and outperforms the state-of-the-art forum crawler, we just add advantage here to download the forums for trusted users only, and we are providing security by giving a key.

REFERENCES

[1] Forum Software Reviews. http://www.forumsoftware.org/forum-reviews

[2] Forum Matrix. http://www.forummatrix.org/index.php

[3] Z. Bar-Yossef, I. Keidar, and U. Schonfeld. Do not crawl in the DUST: different URLs with similar text. In *Proc. of 16th WWW*, pages 111-120, 2007.

[4] Message Boards Statistics. http://www.bigboards.com/statistics/

[5] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang. iRobot:An Intelligent Crawler for Web Forums. In *Proc. of 17th WWW*, pages 447-456, 2008.

[6] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117, 1998.

[7] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song. Finding Question-Answer Pairs from Online Forums. In *Proc. of 31st SIGIR*, pages 467-474, 2008.

[8] A. Dasgupta, R. Kumar, and A. Sasturkar. De-duping URLs via rewrite rules. In *Proc. of 14th KDD*, pages 186-194, 2008.

[9] Y. Guo, K. Li, K. Zhang, and G. Zhang. Board Forum Crawling: a Web Crawling Method for Web Forum. In *Proc. of 2006 IEEE/WIC/ACM WI*, pages 475-478, 2006.

[10] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving Marketing Intelligence from Online Discussion. In *Proc. 11th SIGKDD*, pages 419-428, 2005.

[11] H. S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg and A. Sasturkar. Learning URL Patterns for Webpage De-duplication. In *Proc. of 3rd WSDM*, pages 381-390, 2010.

[12] M. Henzinger. Finding near-duplicate Web pages: a largescale evaluation of algorithms. In *Proc. of 29th SIGIR*, pages 284-291, 2006.

[13] G. S. Manku, A. Jain, and A. D. Sarma. Detecting near duplicates for Web crawling. In *Proc. of 16th WWW*, pages 141-150, 2007.

[14] K. Li, X.Q. Cheng, Y. Guo, and K. Zhang. Crawling Dynamic Web Pages in WWW Forums. *Computer Engineering*, 33(6): 80-82, 2007.

[15] X.Y. Song, J. Liu, Y.B. Cao, and C.-Y. Lin. Automatic Extraction of Web Data Records Containing User-Generated Content. In *Proc. of 19th CIKM*, pages 39-48, 2010.

[16] U. Schonfeld , N. Shivakumar. Sitemaps: above and beyond the crawl of duty. In *Proc. of the 18th WWW*, pages 991-1000, 2009.

[17] Y. Wang, J.-M. Yang, W. Lai, R. Cai, L. Zhang, and W.-Y.Ma. Exploring Traversal Strategy for Web Forum Crawling. In *Proc. of 31st SIGIR*, pages 459-466, 2008.

[18] V. N. Vapnik. The Nature of Statistical Learning Theory.Springer, 1995.

[19] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proc. of 16th WWW*, pages 221-230, 2007.

[20] J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W.-Y. Ma. Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums. In *Proc. of 18th WWW*,pages 181-190, 2009.

AUTHORS PROFILE

A Deepthi, pursuing M.Tech(CSE) from Vikas Group of Institutions, Nunna, Vijayawada. Affiliated to JNTU-Kakinada, A.P., India

Manda Ashok Kumar, working as an Asst. Professor of CSE department at Vikas Group of Institutions, Nunna, Vijayawada, Affiliated to JNTU-Kakinada, A.P., India

Betam Suresh, is working as an HOD, Department of Computer science Engineering at Vikas Group of Institutions, Nunna, Vijayawada, Affiliated to JNTU-Kakinada, A.P., India