JES
Journal of Engineering Sciences

Vol 13, Issue 05, May /2022

ISSN NO:0377-9254

# FoCUS: Learning to Crawl Web Forums

*K. Vishnu vardhan raju, T. Ragu trivedi*
*DEPARTMENT OF MCA*
*Sri Padmavathi College Of Computer Sciences & Technology*

**ABSTRACT :**

In this paper, we present FoCUS (Forum Crawler Under Supervision), a supervised web-scale forum crawler. The goal of FoCUS is to only trawl 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, we reduce the web forum crawling problem to a URL type recognition problem and show how to learn accurate and effective regular expression patterns of implicit navigation paths from an automatically created training set using aggregated results from weak page type classifiers. Robust page type classifiers can be trained from as few as 5 annotated forums and applied to a large set of unseen forums. Our test results show that FoCUS achieved over 98% effectiveness and 97% coverage on a large set of test forums powered by over 150 different forum software packages.

## I. INTRODUCTION

Internet forums are important platforms where users can request and exchange information with others. For example, the Trip Advisor Travel Board is a place where people can ask and share travel tips. Due to the richness of information in forums, researchers are increasingly interested in mining knowledge from them. Zhai et al., Yang et al. and Song et al. extracted structured data from forums. Glance et al. tried to mine business intelligence from forum data. Zhang et al. proposed algorithms to extract expertise network in forums. Gao et al. identified question and answer pairs in forum threads. According to an article from eMarketer - Where Are Social Media Marketers Seeing the Most Success? - Forums are still part of the global social media strategy of the Top 500 Companies, and they are still getting really high marketing success with forums1.

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 and (2) page-flipping links. A forum usually has many duplicate links which point to a common page but with different URLs, e.g., shortcut links pointing to latest posts or URLs for user experience functions such as “view by title”. A generic crawler that blindly follows these links will trawl many duplicate pages that make it inefficient. A Forum typically has many uninformative pages such as login control to protect users’ privacy. Following these links, a crawler will trawl many uninformative pages. Though there are standard-based methods such as specifying the “rel” attribute with “nofollow” value (i.e. “rel=nofollow”)2, Robots Exclusion Standard (robots.txt)3, and Sitemap4, for forum operators to instruct web crawlers on how to crawl a site effectively, we found that over a set of 9 test forums more than 47% of the pages trawled by a generic crawler following these protocols are duplicate or uninformative. This number is a little higher than the 40% that Cai et al. reported but both show the inefficiency of generic crawlers.

Besides duplicate links & uninformative pages, a long forum board or thread is usually divided into multiple pages which are linked by page-flipping links. Generic crawlers process each page individually

and ignore the relationship between such pages. These relationships should be preserved while crawling to facilitate downstream tasks such as page wrapping and content indexing. For example, multiple pages belonging 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 the problem of entry URL discovery. A forum's entry URL points to its home page, which is the lowest common ancestor page of all threads. Our experiment in Section 5.3.2 shows that a crawler starting from an entry URL could achieve much higher performance than starting from other URLs. Previous works by Vidal et al. and Cai et al. assumed that an entry URL is given. But entry URL discovery is not a trivial problem. An entry URL is not necessary at the root URL level of a forum hosting site and its form varies from site to site. Without entry URLs, existing crawling methods such as Vidal et al. and Cai et al. are less effective.

In this paper, we present FoCUS (Forum Crawler Under Supervision), a supervised web-scale forum crawler, to address these challenges. The goal of FoCUS is to trawl relevant content, i.e. user posts, from forums with minimal overhead. Forums exist in many different layouts or styles and powered by a variety of forum software packages, but they always have implicit navigation paths to lead users from entry pages to thread pages. Figure 1 illustrates a typical page and link structure in a forum. For example, a user can navigate from the entry page to a thread page through the following paths:

1. Entry ⇨board ⇨read
2. entry ⇨list-of-board ⇨rd ⇨ad
3. entry ⇨list-of-board & thread t⇨d
4. entry ⇨list-of-board & thread ⇨rd th⇨l
5. entry ⇨list-of-board ⇨of-board & thread boa⇨ threa⇨
6. entry ⇨list-of-board ⇨-of-board & thread thre⇨

We call pages between the entry page and thread page which are on a breadth-first navigation path the *index page*. We represent these implicit paths as the following navigation path (EIT path):

entry page ⇨index page ⇨read page

Links between an entry page and an index page or between two index pages are referred as *index URLs*. Links between an index page and a thread page are referred as *thread URLs*. Links connecting multiple pages of a board and multiple pages of a thread are referred as *page-flipping URLs*. A crawler starting from the entry page of a forum only needs to follow index URLs, thread URLs, and page-flipping URLs to traverse EIT path and achieve all thread pages. The challenge of forum crawling is then reduced to a URL type recognition problem. In this paper, we show how to learn regular expression patterns, i.e. ITF regexes, recognizing these three types of URLs from as few as 5 annotated forum packages and apply

them to a large set of 160 nseen forums packages. Note that we specifically refer to "forum package" rather than "forum site". A forum software package such as vBulletin5 can be deployed by many forum sites.

The major contributions of this paper are as follows:

1. We reduce the forum crawling problem to a URL type recognition problem and implement a crawler, FoCUS, to demonstrate its applicability.

2. We show how to automatically learn regular expression patterns (ITF regexes) that recognize the index URL, thread URL, and page-flipping URL using the page classifiers built from as few as 5 annotated forums.

3. We evaluate FoCUS on a large set of 160 unseen forum packages that cover 668,683 forum sites. To the best of our knowledge, this is the largest scale evaluation of this type. In addition, we show that the patterns are effective and the resulting crawler is efficient.

4. We compare FoCUS with a baseline generic breadth-first crawler, a structure-driven crawler, and a state-of-the-art crawler iRobot and show that FoCUS outperforms these crawlers in terms of effectiveness and coverage.

5. We design an effective forum entry URL discovery method. Entry URLs need to be specified to start crawling to get higher recall. But entry page discovery is not a trivial task since entry pages vary from forums to forums. Our evaluation shows that a naïve baseline can achieve only 76% recall and precision; while our method can achieve over 95% recall and precision.

The rest of this paper is organized as follows. Section 2 is a brief review of related work. we define terms used in this paper. We report our observations which motivate our method and describe the detail of the proposed method in Section 0. In Section 5, we report results of our experiments. In the last section, we draw conclusions and point out future directions of research.

## II. Existing System:

The existing system is a manual or semi automated system, i.e. The Textile Management System is the system that can directly sent to the shop and will purchase clothes whatever you wanted.

The users are purchase dresses for festivals or by their need. They can spend time to purchase this by their choice like color, size, and designs, rate and so on.

They But now in the world everyone is busy. They don't need time to spend for this. Because they can spend whole the day to purchase for their whole family. So we proposed the new system for web crawling.

**Disadvantages:**

1. Consuming large amount of data's.
2. Time wasting while crawl in the web.

## III. Proposed System:

We propose a new system for web crawl as **FoCUS: Learning to Crawl Web Forums.** It is a system overcome by existing crawl systems. In this method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. Target pages were found through comparing

DOM trees of pages with a pre-selected sample target page. It is very effective but it only works for the specific site from which the sample page is drawn. The same process has to be repeated every time for a new site. Therefore, it is not suitable to large- scale crawling. In contrast, FoCUS learns URL patterns across multiple sites and automatically finds forum entry page given a page from a forum. Experimental results show that FoCUS is effective in large scale forum crawling by leveraging crawling knowledge learned from a few annotated forum sites. A recent and more comprehensive work on forum crawling is iRobot. iRobot aims to automatically learn a forum crawler with minimum human intervention by sampling forum pages, clustering them, selecting informative clusters via an informativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection.

## IV. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

## V. MODULES :

1. **Signup & Login**
2. **Upload New Files**
3. **Crawl On Web**

**MODULE DESCRIPTION**:

### 1. Signup & Login:

In this module, we have two sub modules. They are,

- **User signup & login:** In this module user can create account with our site by filling details. And then they can login with our site using this user name and password
- **Admin login:** The owner of this system have a own user name and password for login with the page.

### 2. Upload File:

In this module the owner of the site have to upload a new file for crawl in this site. The user of the page wants to crawl in the site. So the admin should upload a maximum of files for the users need.

Also the admin can view the user details those are having account in his page. And they can view files which they are already uploaded in database.

### 3. Crawl in Web:

The goal of this paper is crawl on the web. So the user can view files in this site which they are uploaded by admin. The users can search a files what they need to know about that.

Also they can view the related searches based on their search. The search contains additional links of its contents also. This web crawling proposed like tree search.

And then user can view their own details which they already gave while signup with this site. They also can change / modify the details.

## VI. SYSTEM CONFIGURATION:-

H/W System Configuration:-

**Processor - Pentium –III**

| | |
|---|---|
| Speed | - 1.1 Ghz |
| RAM | - 256 MB (min) |
| Hard Disk | - 20 GB |
| Floppy Drive | - 1.44 MB |
| Key Board | - Standard Windows Keyboard |
| Mouse | - Two or Three Button Mouse |
| Monitor | - SVGA |

S/W System Configuration:-

- Operating System :Windows95/98/2000/XP
- Application Server : Tomcat5.0/6.X
- Front End : HTML, Java, Jsp
- Scripts : JavaScript.
- Server side Script : Java Server Pages.
- Database : Mysql
- Database Connectivity : JDBC.

\

## VII. CONCLUSION

In this paper, we proposed and implemented FoCUS, a supervised forum crawler. We reduced the forum crawling problem to a URL type recognition problem and showed how to leverage implicit navigation paths of forums, i.e. entry-index-thread (EIT) path, and designed methods to learn ITF regexes explicitly. Experimental results on 160 forum sites each powered by a different forum software package confirm that FoCUS could effectively learn knowledge of EIT path and ITF regexes from as few as 5 annotated forums. We also showed that FoCUS can effectively apply learned forum crawling knowledge

on 160 unseen forums to automatically collect index URL, thread URL, and page-flipping URL string training sets and learn the ITF regexes from the training sets. These learned regexes could be applied directly in online crawling. Training and testing on the basis of forum package makes our experiments manageable and our results applicable to many forum sites. Moreover, FoCUS can start from any page of a forum, while all previous works expect an entry page is given. Our test results on 9 unseen forums show that FoCUS is indeed very effective and efficient and outperforms the state-of-the-art forum crawler, iRobot. The results on 160 forums show that FoCUS can apply the learned knowledge to a large set of unseen forums and still achieve a very good performance. Though, the method introduced in this paper is targeted at forum crawling, the implicit EIT-like path also apply to other sites, such as community Q&A sites, blog sites, and so on.

In the future, we would like to handle forums which use JavaScript, include incremental crawling, and discover new threads and refresh crawled threads in a timely manner. The initial results of applying FoCUS-like crawler to other social media are very promising. We would like to conduct more comprehensive experiments to further verify our approach and improve upon it.

**REFERENCES**

[1] CISCO, "Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update , 2011-2016," Tech. Rep., 2012.

[2] Y. Li, Y. Zhang, and R. Yuan, "Measurement and Analysis of a Large Scale Commercial Mobile Internet TV System," in *ACM IMC*, pp. 209–224, 2011.

[3] T. Taleb and K. Hashimoto, "MS2: A Novel Multi-Source Mobile-Streaming Architecture," in *IEEE Transaction on Broadcasting*, vol. 57, no. 3, pp. 662–673, 2011.

[4] X. Wang, S. Kim, T. Kwon, H. Kim, Y. Choi, "Unveiling the BitTorrent Performance in Mobile WiMAX Networks," in *Passive and Active Measurement Conference*, 2011.

[5] A. Nafaa, T. Taleb, and L. Murphy, "Forward Error Correction Adaptation Strategies for Media Streaming over Wireless Networks," in *IEEE Communications Magazine*, vol. 46, no. 1, pp. 72–79, 2008.

[6] J. Fernandez, T. Taleb, M. Guizani, and N. Kato, "Bandwidth Aggregation-aware Dynamic QoS Negotiation for Real-Time Video Applications in Next-Generation Wireless Networks," in *IEEE Transaction on Multimedia*, vol. 11, no. 6, pp. 1082–1093, 2009.

[7] T. Taleb, K. Kashibuchi, A. Leonardi, S. Palazzo, K. Hashimoto, N. Kato, and Y. Nemoto, "A Cross-layer Approach for An Efficient Delivery of TCP/RTP-based Multimedia Applications in Heterogeneous Wireless Networks," in *IEEE Transaction on Vehicular Technology*, vol. 57, no. 6, pp. 3801–3814, 2008.

[8] K. Zhang, J. Kong, M. Qiu, and G.L Song, "Multimedia Layout Adaptation Through Grammatical Specifications," in *ACM/Springer Multimedia Systems*, vol. 10, no. 3, pp.245–260, 2005.

[9] M. Wien, R. Cazoulat, A. Graffunder, A. Hutter, and P. Amon, "Real-Time System for Adaptive Video Streaming Based on SVC," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1227–1237, Sep. 2007.

[10] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.