International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 3, June 2014, PP 239-245

ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

FOCUS: Learning to Crawl Internet Forums

[1]M.V.Prabath Kumar

(PG scholar) , CSE ,PBR VITS, Kavali

[2]B.Grace

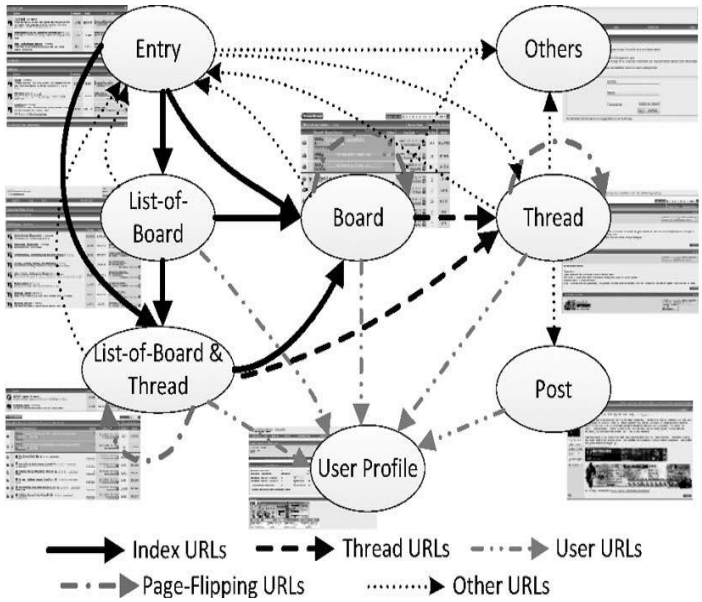
Assistant Professor**,** CSE,PBR VITS, Kavali

Abstract: *In this paper, we present Forum Crawler Under Supervision (FoCUS), a supervised web-scale forum crawler. The goal of FoCUS is to crawl relevant forum content from the web with minimal verhead. 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 we show how to learn accurate and effective regular expression patterns of implicit navigation paths from automatically created training sets using aggregated results from weak page type classifiers. Robust page type classifiers can be trained from as few as five annotated forums and applied to a large set of unseen forums. Our test results show that FoCUS achieved over 98 percent effectiveness and 97 percent coverage on a large set of test forums powered by over 150 different forum software packages. In addition, the results of applying FoCUS on more than 100 community Question and Answer sites and Blog sites demonstrated that the concept of implicit navigation path could apply to other social media sites.*

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

1. INTRODUCTION

INTERNET forums (also called web forums) are important services 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 and Liu [28], Yang et al. [27], and Song et al. [23] extracted structured data from forums. Gao et al. [15] identified question and answer pairs in forum threads. Zhang et al. [30] proposed methods to extract and rank product features for opinion mining from forum posts. Glance et al. [16] tried to mine business intelligence from forum data. Zhang et al. [29] proposed algorithms to extract expertise network in forums. To harvest knowledge from forums, their content must be downloaded first. However, forum crawling is not a trivial problem. Generic crawlers [12], 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 [13], [26]: 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 [7], 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. A forum also has many uninformative pages such as login control to protect user privacy or forum software specific FAQs. Following these links, a crawler will crawl many uninformative pages. Though there are standard-based methods such as specifying the “rel” attribute with the “no follow” value (i.e., “rel ¼ no follow”) [6], Robots Exclusion Standard (robots.txt) [10], and Sitemap [9] [22] for forum operators to instruct web crawlers on how to crawl a site effectively, we found that over a set of nine test forums more than 47 percent of the pages crawled by a breadth-first crawler following these protocols were duplicates or uninformative. This number is a little higher than the 40 percent that Cai et al. [13] reported but both show the inefficiency of generic crawlers. More information about this testing can be found in bellow Section



2. PROBLEM DEFINITION

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.

3. LITERATURE SURVEY

3.1 Introduction

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

3.2 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.

3.3 Disadvantages of Existing System

1. Consuming large amount of data’s.

2. Time wasting while crawl in the web

3.4 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 in formativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection.

3.5 ADVANTAGES OF PROPOSED SYSTEM

1. 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 five annotated forums.

2. 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 evaluation of this type. In addition, we show that the learned patterns are effective and the resulting crawler is efficient.

4. SOFTWARE REQUIREMENT SPECIFICATION

4.1 User Requirement

Feasibility Study

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

•Economical Feasibility

•Technical Feasibility

•Social Feasibility

Economical- Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the

system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

In this project, user is a nettiezen. Nettiezen will search in a forum based on his search the relevant content will be displayed in that page.

4.2 Software Requirements

Operating System : Windows 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.

4.3 Hardware Requirement

Processor -Pentium –III

Speed -1.1 GHz

RAM -512 MB

Hard Disk -40 GB

Floppy Drive -1.44 MB

4.4 Context Diagram

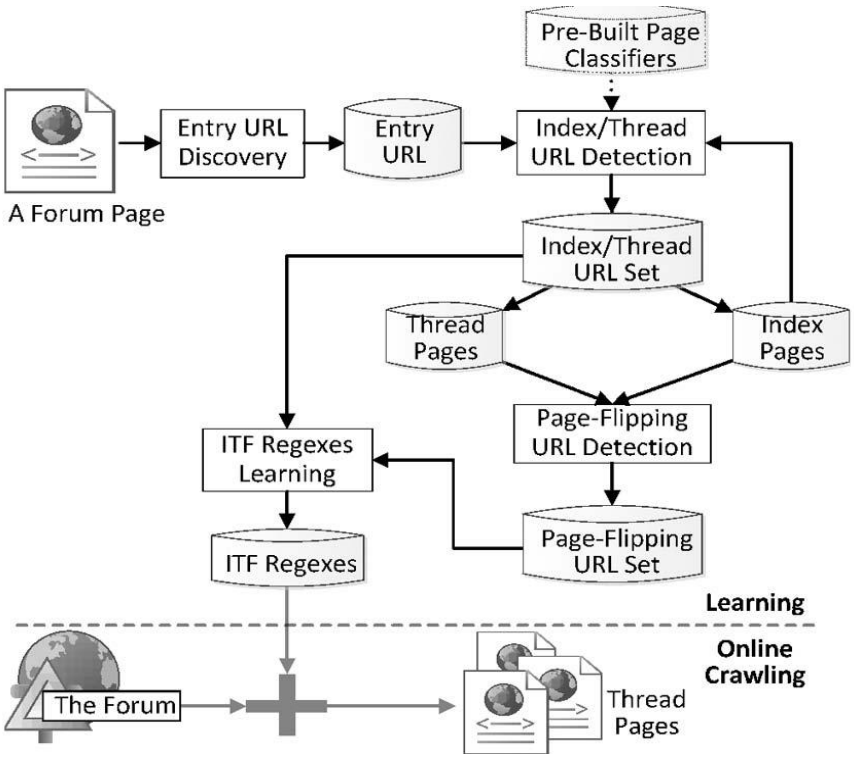


Fig. *Context Diagram*

5. ALGORITHMS

5.1 Index URL and Thread URL Detection

Algorithm: INdexUrlThreadUrlDetection

Input: sp:an entry or index page

Output: it_group:a group of index/thread URLs

1:letit_group be p:data

2:url groups=Collect URL groups by aligning HTML DOM tree of sp;

3.foreachug in url_groups do

4.uganchor_len=Total anchor text length in ug;

5. endforeach

6:if_group=argmax(ug.anchor_len)in url_groups;

7:if_group.DstPageType=Majority page type of the destination pages of URLs in ug;

8.if_group.DstPageType is INDEX_PAGE

9. if_group.Urltype=INDEX_URL;

10. else if if_group.DstPageType is THREAD_PAGE

11.if_group.Urltype=Thread_URL;

12. else

13. if_group=p;

14. end if

15:returnif_group;

5.2 Page-Flipping URL Detection

Algorithm: Page-FlippingUrlThreadUrlDetection

Input: sp:an index page or thread page

Output: if_group:a group of page_flipping URLs

1:letpf_group be

2:url groups=Collect URL groups by aligning HTML DOM tree of sp;

3:foreachug in url_groups do

4:if the anchor texts of ug are digit strings

5:pages=Download(URLs in ug);

6:if pages=have the similar layout to sp and ug appears at same location of pages as in sp

7:pf_group=ug;

8:break;

9:end if

10:end if

11:endforeach

12:ifpf_group is

13:foreachurl in outgoing URLs in sp

14:P=Download(url);

15:pf_url=Extract URL in p at the same location as url in sp;

16:ifpf_url exists and pf_url.anchor==url.anchor and pf_url.UrlString|=url.UrlString

17:Addurl and cand_url into pf_group;

18:break;

19:end if

20:endforeach

21:end if

22:pf_group UrlType=PAGE_FLIPPING_URL;

23:returnpf_group;

6. 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] Blog, http://en.wikipedia.org/wiki/Blog, 2012.

[2] "ForumMatrix," http://www.forummatrix.org/index.php, 2012.

[3] Hot Scripts, http://www.hotscripts.com/index.php, 2012.

[4] Internet Forum, http://en.wikipedia.org/wiki/Internet_forum,2012.

[5] "Message Boards Statistics," http://www.big-boards.com/statistics/, 2012.

[6] nofollow, http://en.wikipedia.org/wiki/Nofollow, 2012.

[7] "RFC 1738—Uniform Resource Locators (URL)," http://www.ietf.org/rfc/rfc1738.txt, 2012.

[8] Session ID, http://en.wikipedia.org/wiki/Session_ID, 2012.

[9] "The Sitemap Protocol," http://sitemaps.org/protocol.php, 2012.

[10] "The Web Robots Pages," http://www.robotstxt.org/, 2012.

[11] "WeblogMatrix," http://www.weblogmatrix.org/, 2012.

[12] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine." Computer Networks and ISDN Systems, vol. 30,nos. 1-7, pp. 107-117, 1998.

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

[14] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-Duping URLs via Rewrite Rules," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.

[15] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song, "Finding Question-Answer Pairs from Online Forums," Proc. 31st Ann. Int'l ACMSIGIR Conf. Research and Development in Information Retrieval, pp. 467-474, 2008.

[16] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, "Deriving Marketing Intelligence from Online Discussion," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 419-428, 2005.

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

[18] M. Henzinger, "Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 284-291, 2006.

[19] H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg, and A. Sasturkar, "Learning URL Patterns for Webpage De-Duplication," Proc. Third ACM Conf. Web Search and Data Mining, pp. 381-390, 2010.

[20] K. Li, X.Q. Cheng, Y. Guo, and K. Zhang, "Crawling Dynamic Web Pages in WWW Forums," Computer Eng., vol. 33, no. 6, pp. 80-82, 2007.

[21] G.S. Manku, A. Jain, and A.D. Sarma, "Detecting Near-Duplicates for Web Crawling," Proc. 16th Int'l Conf. World Wide Web, pp. 141-150, 2007.

[22] U. Schonfeld and N. Shivakumar, "Sitemaps: Above and Beyond the Crawl of Duty," Proc. 18th Int'l Conf. World Wide Web, pp. 991-1000, 2009.

[23] X.Y. Song, J. Liu, Y.B. Cao, and C.-Y. Lin, "Automatic Extraction of Web Data Records Containing User-Generated Content," Proc. 19th Int'l Conf. Information and Knowledge Management, pp. 39-48,2010.

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

[25] M.L.A. Vidal, A.S. Silva, E.S. Moura, and J.M.B. Cavalcanti, "Structure-Driven Crawler Generation by Example," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 292-299, 2006.

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

[27] 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," Proc. 18th Int'l Conf. World Wide Web, pp. 181-190, 2009.

[28] Y. Zhai and B. Liu, "Structured Data Extraction from the Web based on Partial Tree Alignment," IEEE Trans. Knowledge Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[29] J. Zhang, M.S. Ackerman, and L. Adamic, "Expertise Networks in Online Communities: Structure and Algorithms," Proc. 16th Int'l Conf. World Wide Web, pp. 221-230, 2007.

[30] L. Zhang, B. Liu, S.H. Lim, and E. O'Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 1462-1470, 2010.