International Journal of Science Engineering and Advance Technology

Scuttling Web Opportunities By Application Cramming

Dhulipalla Vijaya Sree#1,Alahari Hanumat Prasad#2

#1 Student of M.Tech Department of Computer Science Engineering G.V.R&S college of Engineering & Technology, GUNTUR

#2 Department of Computer Science Engineering G.V.R&S college of Engineering & Technology, GUNTUR.

ABSTRACT

The web contains large data and it contains innumerable websites that is monitored by a tool or a program known as Crawler. The main goal of this paper is to focus on the web forum crawling techniques. In this paper, the various techniques of web forum crawler and challenges of crawling are discussed. The paper also gives the overview of web crawling and web forums.

Internet is emergent exponentially and has become progressively more. Now, it is complicated to retrieve relevant information from internet. The rapid growth of the internet poses unprecedented scaling challenges for general purpose crawlers and search engines. In this paper, we present a novel Forum Crawler under Supervision (FoCUS) method, which supervised internet-scale forum crawler. The intention of FoCUS is to crawl relevant forum information from the internet with minimal overhead, this crawler is to selectively seek out pages that are pertinent to a predefined set of topics, rather than collecting and indexing all accessible web documents to be capable to answer all possible ad-hoc questions. FoCUS is continuously keeps on crawling the internet and finds any new internet pages that have been added to the internet, pages that have been removed from the internet. Due to growing and vibrant activity of the internet; it has become more challengeable to navigate all URLs in the web documents and to handle these URLs. We will take one seed URL as input and search with a keyword, the searching result is based on keyword and it will fetch the internet pages where it will find that keyword.

Keywords

Web Crawling, Web Forums, FoCUS,EIT path, forum crawling, ITF regex, page classification, page type, URL pattern learning, URL type.

1.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.

2. PROPOSED SCHEME

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

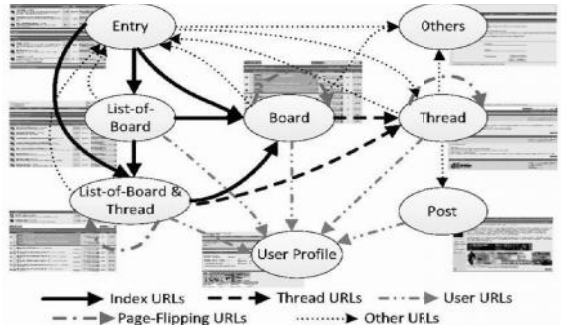

Fig 1 structure of a webform

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.

3. RESULT

In this section, we show the result which are we proposed implement in previous section. In this section, we show all the result with the help of table and graph in separate module with detailed description such as overview, Online Crawling and Entry URL Discovery.

A. Entry URL Discovery

In this module, we discuss, forum crawling assume in URL Entry. However, finding forum entry URL is not trivial. To display this, we used our URL entry discovery method with a heuristic baseline. For each forum in the test set, we randomly sampled a page and fed it to this module. Then, we checked manually if the output was indeed its entry page. In order to see whether FoCUS and the baseline were robust or not, we repeated this process 10 times with unusual sample pages. The results are shown in Table 1. The baseline had 76 percent precision and recall. On the contrary, FoCUS achieved 99 percent precision and 99 percent recall. The low standard deviation also designates that it is not sensitive to sample pages. There are two main failure cases: 1) forums are no longer in operation and 2) JavaScript generated URLs which we do not handle currently. We balanced the different types of URL for find the efficiency of thread URL and URL

discovery in terms of generic crawler in figure-02

TABLE 1- Results of Entry URL Discovery

Method	Precision %		Recall %	
	Average	Std.Dev.	Average	Std. Dev.
Baseline	76.38	1.74	76.38	1.74
FoCUS	99.31	0.20	99.13	0.32

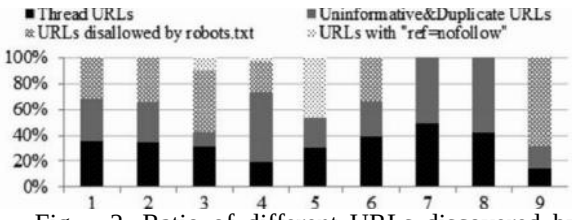

Fig.- 2. Ratio of different URLs discovered by a generic crawler

B. Evaluation of Online Crawling

In this module, we evaluate FoCUS with other existing methods for find the efficiency of result

TABLE 2- Forums Used in Online Crawling Evaluation

ID	Forum	Forum Name	Software	#Threads
1	forums.afterdawn.com	AfterDawn: Forums	Customized	535,383
2	forums.asp.net	ASP.NET Forums	Community Server	66,966
3	forum.xda-developers.com	Android Forums	vBulletin	299,073
4	bbs.cqzg.cn	春秋中文社区	Discuz!	428,555
5	forums.crackberry.com	BlackBerry Forums	vBulletin V2	525,381
6	forums.gentoo.org	Gentoo Forums	phpBB V2	681,813
7	lkcn.net/bbs	英华论坛	IP.Board	180,692
8	techreport.com/forums	Tech Report	phpBB	65,083
9	www.redandwhitekop.com/forum	Liverpool FC Forum	SMF	138,963

We preferred nine forums (Table 2) among the 190 test forums for this assessment investigation. Eight of the nine forums are popular software packages used by many forum sites this is about 53 percent of forums powered by the 200 packages deliberate in this paper, and about 15 percent of all forums we have found.

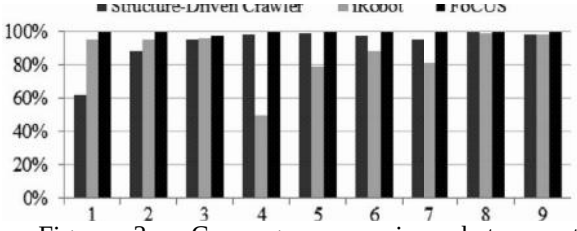

Fig.- 3 Coverage comparison between the structure-driven crawler, iRobot, and FoCUS.

In this module, we report the results of the comparison between the structure-driven crawler, iRobot, and FoCUS. Although the structure-driven crawler is not a forum crawler, it could be utilized to forums. To make a more meaningful comparison, we used it to find page-flipping URL patterns in order to increase its coverage. As to iRobot, we re-implemented it. We permit the structure-driven crawler, iRobot, and FoCUS crawl each forum until no more pages could be retrieved. After that we counted how many threads and other pages were crawled, correspondingly.

4. CONCLUSION

The web crawler collects detail information about the website and the websites links. It includes the website URL, the web page title, the meta tag information, the web page content, the links on the page. In this paper the basic of web crawling is discussed and the survey of different web forum crawling techniques is discussed. FoCUS automatically crawl the forum data and it clean up the unwanted data. After cleaning the unwanted data, FoCUS allocates that space to new queries posted by the user. Comparing with other techniques of web forum crawling, FoCUS outperforms these crawlers in terms of effectiveness and coverage. It shows that the learned patterns are effective and the resulting crawler is efficient.

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.

5. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[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.

[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.

AUTHORS:

DHULIPALLA VIJAYA SREEis a student of computer science engineering from G.V.R&S College of engineering & technology, presently pursuing M.TECH (cse) from this college. She received MCA from ANU in the year of 2010.

ALAHARI HANUMAT PRASAD is a ASSOCIATE PROFESSOR Department of CSE at G.V.R&S college of Engineering & Technology,Guntur. He received M.Tech in Computer science engineering from JNTUK.He gained 10 years Experience on Teaching. He is a good Researcher in Network Security.