IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Special Issue-2, April 2016 E-ISSN: 2347-2693

A Supervised Forum Crawler

Sreeja S R[1], Sangita Chaudhari[2]

[1]Computer Engineering, A. C Patil College of Engineering, India
[2]Computer Engineering, A. C Patil College of Engineering, India

Available online at: www.ijcseonline.org

Abstract— Web Forums or Internet Forums provide a space for users to share, discuss and request information. Web Forums are sources of huge amount of structured information that is rapidly changing. So crawling Web Forums requires special softwares. A Generic Deep Web Crawler or a Focused Crawler cannot be used for this purpose. In this paper, we propose an effective Web Crawler especially for Internet Forums. This Forum Crawler overcomes the drawbacks of many of the existing Forum Crawlers. It has the ability to detect the Entry URL (Uniform Resource Locator) of a Forum site, given any page of it. Crawling process starting from Entry URL increases the coverage. Different URLs in the Web Forums are classified into four categories. The entire crawling process is divided into a learning part and an online crawling part. Learning part will create regular expressions based on URLs and crawling part actually crawls the Web pages.

Keywords— Forum Crawling; URL Type; Page Classification; Crawling Strategy; Javascript-based URLs;

I. INTRODUCTION

A Web Crawler also known as "web robot" is a software that simulates the human behavior of visiting Web pages [2]. The main purpose of using Web Crawlers is to validate Web pages, analyze them, notify about changes in the Web pages, and visualize Web pages. A Deep Web Crawler does not consider the relationship among the Web pages while crawling. Deep Web Crawlers generally use Breadth First Strategy (BFS) for crawling. Focused crawlers are another type of Web Crawlers which retrieve Web pages that are specific to a pre-defined topic. Focused Crawlers consume lesser amounts of system resources when compared to Generic Deep Web Crawler. Focused crawlers use a variety of techniques for domain-specific retrieval. Some of them use semantic-based techniques while others use some properties of the Web pages which include number of links between the pages and similarity of their contents.

Web Forums are also called Internet Forums [1] and they provide a space for users to share knowledge. Web Forums are very important sources of information. It is a place where a user can post his queries and can get their answers. Also he can get large amount of other useful information related to his query.

This paper is organized as follows. Section II contains a review of related crawlers like Generic Deep Crawlers, Focused Crawlers and other Forum Crawlers. Section III explains the system architecture. In Section IV, the Results of the experiments performed on the Supervised Crawler are discussed and the last section is the Conclusion & Future Work.

II. RELATED WORK

Majority of the Web Crawlers existing today can be categorized into two; Generic Deep Web Crawlers and Focused Crawlers. A Generic Deep Web Crawler retrieves all the Web pages as it crawls by following the hyperlinks. The basic procedure followed by such a crawler is shown in Figure 1.

An example of such a crawler is PyBOT [10]. PyBOT starts by taking the 'seed URL' and from that URL, it gets all other URLs. This is done by scanning the Web page pointed to by that URL. Using the collected URLs, it crawls again until a point where no new URLs are found.

This type of crawlers generally uses *Breadth First Crawling (BFC)* strategy which has many drawbacks:

- After a large number of Web pages are fetched, it will start losing its focus which will result in the introduction of a lot of noise into the final collection.
- It may crawl many redundant and duplicate pages and many times misses useful pages.
- Download of large amount of useless pages wastes network bandwidth and negatively affects repository quality.
- It crawls the pages without understanding the correlation among them and so it cannot be used to crawl Web Forums.

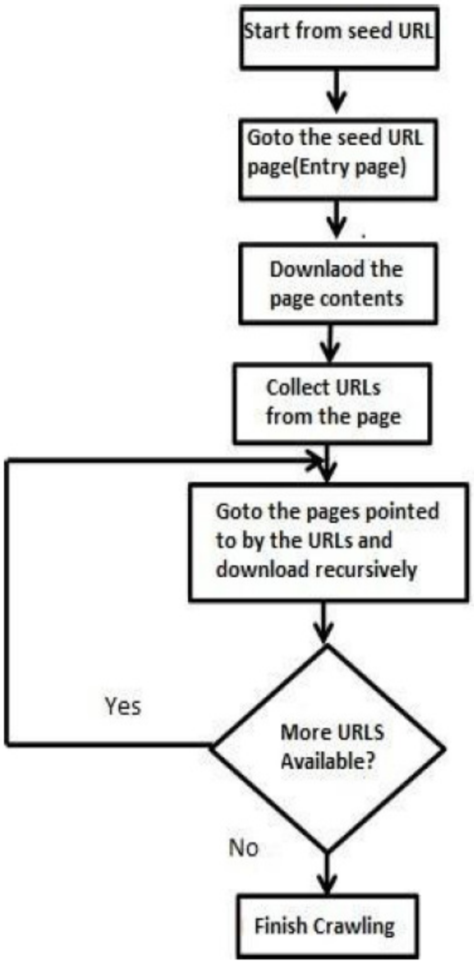
Start from seed URL
Goto the seed URL page(Entry page)
Downlaod the page contents
Collect URLs from the page
Goto the pages pointed to by the URLs and download recursively
More URLS Available?
Yes
No
Finish Crawling

Figure-1. Generic Deep Web Crawler

The second category of Web Crawlers falls under Focused Crawlers. A Focused Crawler is designed to crawl Web pages relevant to a pre-defined topic [6]. There are several Focused Crawlers available each using different techniques.

A Semantic Focused Crawler [5] is a Focused Crawler that makes use of Semantic Web technologies for performing the crawling. An Ontology-based Semantic Focused Crawler links Web documents with related Ontology concepts for the purpose of categorizing them. It makes use of Ontologies to analyze the semantic similarity between URLs of Web pages and topics. The limitation of this type of crawler is that most of these crawlers fetch the surrounding texts of URLs as the descriptive texts of the URLs and compute the similarity between the URLs and ontology concepts based on these texts. But, the surrounding texts cannot be used to correctly or sufficiently describe the URLs. A Metadata Abstraction based Semantic Focused Crawler is a Focused Crawler that extracts meaningful information or metadata from relevant Web pages and annotates the metadata with Ontology Mark-up Languages. Many of these supervised classification models use predefined classifiers based on plain texts without enough semantic support. This will decrease the performance of document classification.

Another Focused Crawler combines the features of Ontology-based Focused Crawlers and the Metadata Abstraction Focused Crawlers. Since this crawler is built by combining the characterictics of Ontology-based and Metadata-based crawlers, it has all the disadvantages of both types of crawlers. A Focused Crawler based on Link Structure and Content Similarity makes use of a combination of link-structure and similarity of contents between the Web pages for performing the crawling [12]. This crawler starts with an initial 'Seed'. If the initial Seed page does not relate to the domain, then the number of related pages will be very less in the beginning stages. This will affect the overall efficiency of the crawler.

Generic Deep Web Crawlers and Focused Crawlers cannot be used for crawling Web Forums because a Web Forum has a specific structure which a normal Web Crawler cannot follow. So special softwares are required for crawling Web Forums. One of such Web Forum Crawlers is a Board Forum Crawler (BFC) [7]. It exploits the characteristic of Web Forum that they have an organized structure. It starts from the home page, extracts board page seeds, board page links and post page links and stores them in a link queue. Crawling is performed by retrieving the links in the queue. The drawback of this crawler is that it is designed for specific Forum Sites which have fixed structure. Another Forum Crawler based on Links and Text Properties uses the outgoing links and text information in the Forum pages [8]. Using this information, it designs traversal strategies and the best traversal path will be selected for crawling. It gives best performance only when the URLs are keyword based. The performance is negatively affected for Forum sites with verbose-based URL structures.

iRobot [4] is another Forum Crawler which automatically rebuilds the sitemap of the target Web forum site and then selects an optimal traversal path which only traverses informative pages and skips invalid and duplicate ones[8]. This crawler does not have the capability to detect Entry URLs which will affect the coverage and the overall performance of the Crawler. FoCUS (Forum Crawler Under Supervision) is a Forum Crawler which integrates a learning module and an online crawling module [9]. The learning module classifies the pages and detects URLs present in those pages. It forms regular expressions based on the detected URLs. The online crawling part uses these regular expressions to detect the URLs and to perform the crawling.

III. SUPERVISED FORUM CRAWLER ARCHITECTURE

FoCUS gives better precision, accuracy and coverage compared to other Forum Crawlers that were discussed above. Moreover, it has the capability to detect the Entry URL of a Forum site. But there are some limitations to FoCUS. First one is that it is using SVM (Support Vector Machine) [3] with a linear kernel setting for index/thread page classification. Linear kernel setting has several disadvantages:

- A linear kernel setting provides less accuracy.
- Linear kernel needs more convergence time.
- Cannot be used for data that is not linearly separable.

The second limitation is FoCUS uses a weak page classifier (SVM^{light}) along with majority voting method for index/thread URL detection. The issues related with this combined method are that the outcome of a weak classifier may be erroneous. If the URL group contains very less number of URLs (two or four) and if majority of them are misclassified, then that will affect the accuracy of the crawling process.

One more drawback is that if the URL group contains a few URLs and if the number of URLs is even, then the majority voting method will fail if half of the URLs in the group are classified erroneously.

In this paper, we propose a Supervised Forum Crawler which is an enhanced Forum crawler that overcomes the above disadvantages.

The major enhancements are:

- Instead of using SVM^{light} with a linear kernel setting for index/thread classification, we are using the Weka tool with J48 classifier which is giving better results in terms of accuracy and precision [11].
- The weak classifier combined with majority voting method is removed and a strong page classifier is used which decreases number of misclassifications.
- A new feature HasReplyBtn is added to the list of features used for index/thread page classification

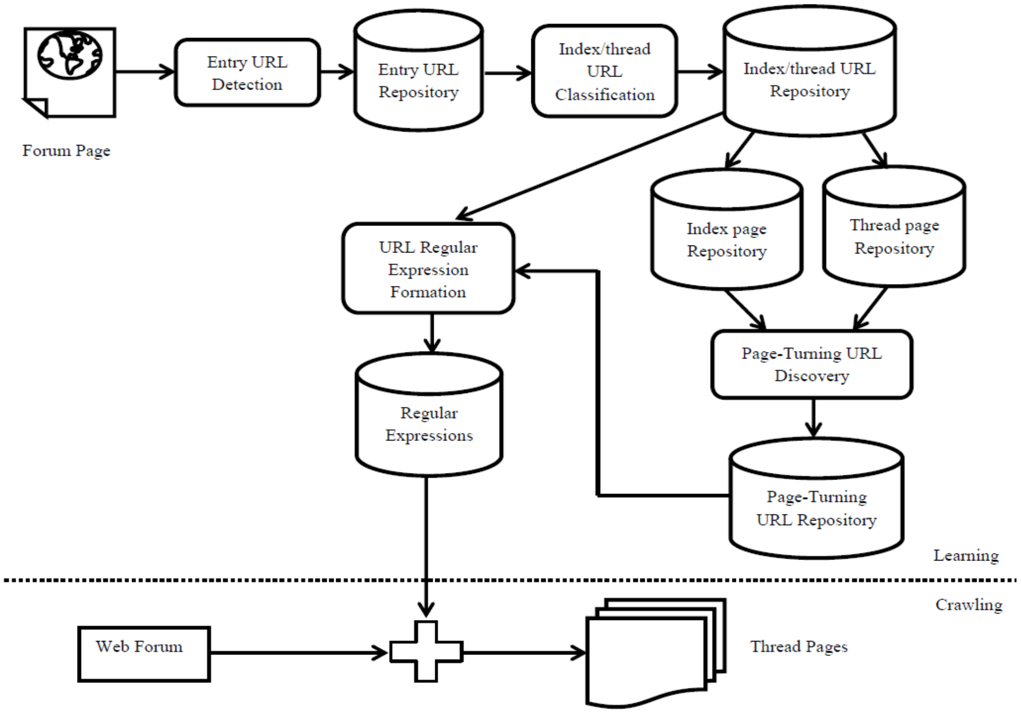

Figure-2. Supervised Forum Crawler Architecture

The architecture of the Supervised Forum Crawler is shown in Figure 2. This crawler consists of two major modules: one learning module and a crawling module.

Learning module forms regular expressions by analyzing the different URLs in different pages. The crawling module performs the online crawling by using these regular expressions.

The learning part in turn contains several sub modules. This crawler detects the Entry URL of a Web Forum, given any page of it. This task is performed by the *Entry URL Detection* sub module. The detected Entry URL is stored in the Entry URL repository. Next sub module of the learning module is the *Index/Thread URL Classification* module. All the URLs from the index and thread pages are collected and these URLs are classified as either index or thread URLs. The URLs are classified according to which page it is pointing to. The page classification is done by SVM with a Gaussian kernel setting. Several page features are given as inputs to the classifier in the baseline system. In addition to these features, one more parameter that we are using in our system is a HasReplyBtn feature which indicates the presence of a Reply button in post pages. Almost 95% of the Forums contain a Reply button in the post page. The classified index/thread URLs are stored in a index/thread URL repository.

From each of the index and thread pages, the *Page-Turning URL Discovery* sub module will collect the page-turning URLs. Next sub module of the learning module is the *URL Regular Expression Formation* module. This module analyses all the collected URLs and forms Regular Expressions from them. When the learning module finishes its task, actual crawling process will be performed.

IV. EXPERIMENTS AND RESULTS

The performance of the Supervised Forum Crawler is evaluated using different parameters like precision, recall and crawling time. We have compared the results with two traditional crawlers which use Breadth First and Depth First strategies for crawling. For calculating precision and recall for the Weka tool, seven different classifiers like J48, PART, OneR, Decision Table etc. are used. Crawling time is calculated first for 50 pages and then it is increased to 100, 200, 300, 400 and 500 in each step. For each case, the crawling time taken by both the proposed system and the traditional crawlers are calculated.

A. *Precision and Recall*

We have used seven different classifiers available in Weka tool to evaluate the precision and recall given by the index/thread classification module. Inorder to obtain the precision and recall values, we have to provide the entry URL to the system and should collect the training set URLs. Here we have used the ASP.Net forum (http://forums.asp.net) as the testing forum. The graph plotted for Classifier v/s Precision & Recall is shown in Figure 3. J48, PART and OneR are the classifiers which are giving the highest precision values and Decision Table, KStar and IBK are giving lesser precision values. In this system, we have chosen J48 as the classifier since it is one of the classifiers which are giving the highest values. In the graph, we can observe that SMO (Sequential Minimal Optimization) is giving the least precision of 64.2%. SMO is the SVM (Support Vector Machine) implementation in Weka and FoCUS uses a weak version of SVM for doing the index/thread classification. If we consider the recall parameter, all the classifiers in Weka are giving 100% recall values whereas SMO is giving a recall value of just 45% which is very less.

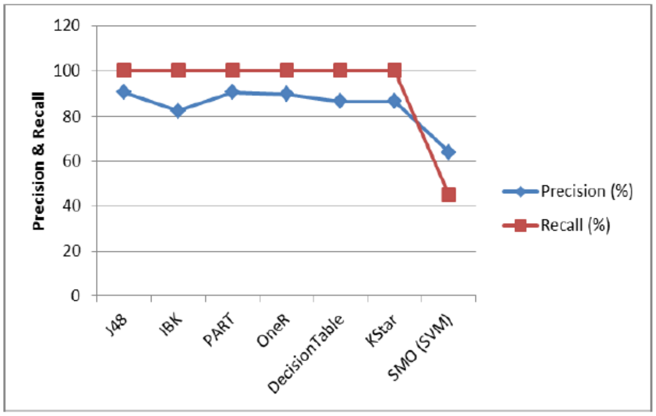

Figure-3. Precision and Recall v/s Weka Classifiers

B. *Crawling Time*

We have calculated the crawling time required by the proposed system for different no. of pages which is shown in Table 1. Crawling time is calculated as the sum of time taken to train the crawler and the time taken to do the actual crawling. The test forum is the ASP.Net forum. The crawling time for the proposed system is initially more than that of the traditional crawlers. But, as the no. of pages is increased it is taking lesser time than BFS and DFS. So the proposed system will take lesser time compared to the traditional crawlers for crawling a forum completely.

Corresponding Author: *Sreeja S R, Sreeja.sr@acpce.ac.in*
Department of Computer Engineering, A. C. Patil College of Engineering, India

Table-1. Time for Different No. of Pages for the Proposed System.

No. of Pages	Training Time	Crawling Time
50	1 min 44 sec	1 min 25 sec
100	1 min 44 sec	2 min 4 sec
200	1 min 44 sec	2 min 23 sec
300	1 min 44 sec	3 min 29 sec
400	1 min 44 sec	4 min 14 sec
500	1 min 44 sec	6 min 5sec

V. CONCLUSION & FUTURE WORK

Web crawlers are softwares used mostly by search engines to update their database. Web crawlers can be generic or focused. Generic crawlers retrieve all the web pages of a website. Focused crawlers crawl websites to retrieve Web pages which are related to a particular topic. Web forum site is a discussion site where people can discuss about a common subject. Among the forum crawlers, FoCUS is a crawler which is giving best performance. A detailed study of FoCUS disclosed some drawbacks in it. So a new Supervised Forum Crawler is developed which overcomes these drawbacks. The results show that the newly developed forum crawler gives better performance in terms of precision and crawling time. In future, we would like to extend this crawler to other sites like Question & Answer (Q & A) sites, blog sites and other social media sites also.

REFERENCES

[1] Internet forum. http://en.wikipedia.org/wiki/Internet forums.

[2] Web Crawler. http://en.wikipedia.org/wiki/Webcrawler.

[3] Asa Ben-Hur and JasonWeston. A user's guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010N.B. Salem, and J-P Hubaux, "Securing Wireless Mesh Networks", IEEE Wireless Communications, Vol.13, Issue-2, **2006**, pp.**50-55**.

[4] Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang. irobot: An intelligent crawler for web forums. In Proceedings of the 17th international conference on World Wide Web, pages **447–456**. ACM, **2008**.

[5] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R Scott Cost, Yun Peng, Pavan Reddivari, VC Doshi, and Joel Sachs. Swoogle: A semantic web search and metadata engine. In Proc. 13th ACM Conf. on Information and Knowledge Management, pages **65–69**, **2004**.

[6] Hai Dong and Farookh Khadeer Hussain. Focused crawling for automatic service discovery, annotation, and classification in industrial digital ecosystems. *Industrial Electronics,IEEE Transactions on*, 58(6):**2106–2116**, **2011**.

[7] Yan Guo, Kui Li, Kai Zhang, and Gang Zhang. Board forum crawling: a web crawling method for web forum. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages **745–748**. IEEE Computer Society, **2006**.

[8] Amit Sachan, Wee-Yong Lim, and Vrizlynn LL Thing. A generalized links and text properties based forum crawler. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, pages **113–120**. IEEE Computer Society, **2012**.

[9]] Jingtian Jiang, Nenghai Yu, and Chin-Yew Lin. Focus: learning to crawl web forums. In Proceedings of the 21st international conference companion on World Wide Web, pages **33–42**. ACM, **2012**.

[10] Alex Goh Kwang Leng, KP Ravi, Ashutosh Kumar Singh, and Rajendra Kumar Dash.Pybot: An algorithm for web crawling. In Nanoscience, Technology and Societal Implications (NSTSI), 2011 International Conference on, pages **1–6**. IEEE, **2011**.

[11] Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations. **1999**.

[12] Jamali, Mohsen, et al. "A method for focused crawling using combination of link structure and content similarity." Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, **2006**.

AUTHORS PROFILE

Sreeja S R received the M.E degree in Computer Engineering from Mumbai University, Mumbai, India, in 2015 and B.Tech degree in Computer Engineering from Mumbai University, Mumbai, India, in 2006.

She is currently working as an Assistant Professor with Computer Engineering Department, A. C. Patil College of Engineering, Kharghar, Navi Mumbai. Her research interests include Web Crawlers and Data Mining Techniques.

Sangita Chaudhari received the Ph.D degree from IIT Bombay, Mumbai, India, in 2016 and ME degree in computer engineering from Mumbai University, Mumbai, India, in 2008.

She is currently working as an Assistant Professor with Computer Engineering Department, A. C. Patil College of Engineering, Kharghar, Navi Mumbai. Her research interests include digital image processing, advanced databases, Geographical Information Systems, High Performance Computing, and information security techniques.