Supervised Web Forum Crawling

[1]Priyanka S. Bandagale, [2]Dr. Lata Ragha
[1]Student, [2] Professor and HOD
[1]Computer Department,
[1]Terna college of Engineering, Navi Mumbai, India

***Abstract* - In this paper, we present a supervised internet Forum crawler. The goal of planned methodology is to crawl optimum forum content from the net with stripped-down overhead. Forum threads contain info content that's the target of forum crawlers. though forums have completely different varieties of designs or layouts and area unit powered by numerous forum code packages, they continuously have similar implicit navigation ways connected by such uniform resource locator varieties to guide users to string pages from entry pages. supported this observation, we tend to cut back the net forum crawl drawback to a uniform resource locator (URL) kind recognition drawback victimization our crawler by demonstrating its results and pertinence. Crawler with multi-threaded downloader is chargeable for beginning threads and getting the knowledge regarding the web site being fetched. Multiple processes area unit run in parallel to perform the higher than task, so transfer rate is maximized and downloading time is decreased . finally we tend to show that our planned Naïve mathematician Classifier is best than generic BFS with the assistance of Associate in Nursing application in variety of native computer program.**

***Index Terms* – forum sites , crawling, ITF regex, URL classification, page type, URL pattern learning, URL type, EIT path.**

I. Introduction

Nowadays, there are numerous internet forums dealing with diverse topics like news, monetary knowledge, software support, programming discussion, financial data, entertainment and technical discussion. Forums have a specific set of terminology associated with them; e.g., a single discussion is called a "thread", or topic[7]. A discussion forum is tiered or tree-like in structure. A forum can contain a number of sub forums, each of which may have numerous areas. Each new discussion in the forum's topic is called a thread, and can be replied to by as many people as so desire. While forum crawling is still a demanding task due to complicated in-site link structures and it can sometimes take a long time. The generic crawlers which uses a breadth-first strategy, usually downloads many duplicate and uninformative pages from the forums and they process each page in the page flipping links individually and ignores relationship between the pages. However However, correct conformation of web site transfer will acceleration web site forum scan to reap information from forums, their content should be downloaded 1st. However, forum crawling is not a insignificant problem. Generic crawlers which adopt a breadth-first traversal strategy, are usually ineffectual and disorganized for forum crawling because it crawls duplicate links and uninformative pages and page-flipping links. A forum generally has several duplicate links that time to a standard page however with totally different URLs [3], e.g., road links inform to the newest posts or URLs for user expertise functions like ―view by date or ―view by title. A generic crawler that blindly follows these links can crawl several duplicate pages, creating it inefficient.

Forum Structure

Each page in forum site may have its own layouts. Based on their layout structure [1], the pages in forum sites are classified into four categories: Entry page: the house page of the forum site that contains an inventory of boards.

- Index page: An index page contains table-like structure, where each row in the table contains information of a board or thread.
- Thread page: A thread page contains a list of users' posts.
- Other pages like login management, about us, user profile pages, etc.

each forum web site has similar navigation ways although they differs in layout and designs structure.

Entry page→ Index page→ Thread page

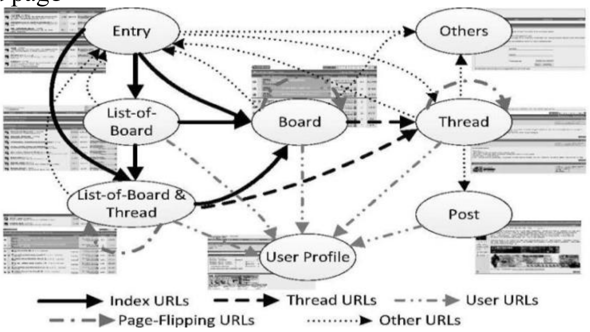


Figure 1: Different Forum URL categories

II. RELATED WORK

To make a balance between the "performance and cost", most of the generic Web crawlers implement the breadth-first strategy (BFS)[4] and limit depth of crawling. However, practically it is difficult to select an appropriate crawling depth for each site. A narrow crawling strategy does not give assurance to access all valued content, whereas a deep crawling strategy may cause in numerous duplicate and invalid pages. Few research works tried to uncover more effective crawling strategy than the BFS to improve the quality of content. A new method is Board Forum Crawling (BFC) to crawl Web forum[4]. This method utilizes the structured characteristics of the Web forum sites and simulates human behavior of visiting Web Forums. This method starts its crawl from the homepage, and then enters each board of the site, and then eventually crawls all the posts of the site directly. The Board Forum Crawling (BFC) can crawl most considerable information of a Web forum site efficiently and in a simple way. A recent and more comprehensive work on forum crawling is iRobot by Cai et al. [1]. The goal of this method is to automatically learn a forum crawler with minimum human interference by sampling forum pages, clustering them, selecting informative clusters via an in formativeness measure. This method uses spanning tree algorithm for finding a traversal path and these selection procedure requires human inspection. Follow up work by Wang et al. [2] proposed an algorithm to address the traversal path selection problem. They introduced the concept of skeleton link and page-flipping link. The most important link underneath the structure or architecture of a forum site known as skeleton link. Another related work is template-independent approach to extract structured data from web forum sites. The method introduces an automatic approach to extract Site -level information with the reconstructed sitemap. The site-level knowledge includes[2]: 1. Linkages among pages inter vertexes of the sitemap should have certain functionalities, such as the title link between list page and post page; and 2. Interrelationships of pages sharing the similar layout deign, such as the post contents appear in the same Document Object Model (DOM) path of all post pages. In this way, we can leverage the mutual information among pages inner or inter vertices of the sitemap.

III. PROBLEM STATEMENT

We proposed a supervised web-scale forum crawler. The goal of crawler is to only trawl important forum content from the web with least overhead. Forum threads contain information content that is the target of forum crawlers. Although forums have different layouts or designs and are high 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 which is different from other normal websites. Based on this observation, we proposed a method that reduce the web forum crawling problem to a URL sort recognition drawback 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.

Generic Crawler Fails in case of web forum because of the following reasons:

- Presence of many functional links
- Inability to index relationship among post pages.
- Avoids crawling deep inside a web site.
- Inefficient and ineffective

IV. METHODOLOGY

The proposed system consist of two major parts as shown in figure 2

1. learning phase

2. online crawling part

1. The learning part learns ITF regexes of a given forum from automatically constructed URL examples. The online crawling part applies learned ITF regexes to crawl all threads efficiently. The online crawling part then tries to crawl all thread pages that matches the learned ITF regexes. The learning phase is composed of four modules:

1) Entry URL discovery, 2) Index/Thread URL detection, 3) Page-Flipping URL detection, and 4) ITF regexes learning.

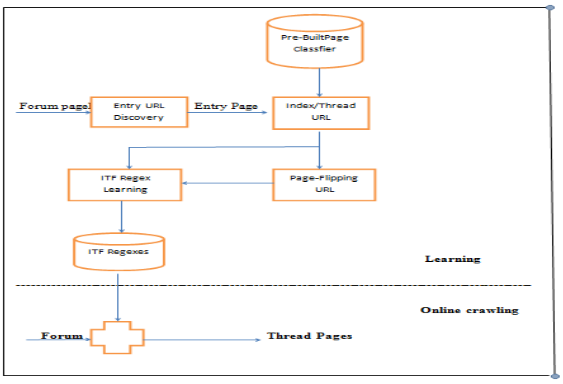


Figure 2: Flow of the Proposed System

Naive Bayes model is used for learning from train dataset it never contains more attributes than components available and it guarantee that assumption will be brought up effectively. Naive Bayes classifier is basically a probabilistic classifier based on assumption. On the basis of assumption and learning from train set; it finds out most suitable assumption based on previous assumptions and initial knowledge.

Proposed work is done for classification of forum pages using Naïve Bayes Algorithm shown in figure 3.

Proposed framework works in three steps as,
Step 1: consist of training data collection,
Step 2: supervised learning with the training data,
Step 3: URL Classification & Recognition.

These stages can operate repeatedly as in batched learning, or in an interleaving manner: additional data is collected to incrementally train the classification model while the model is used in detection and recognition.

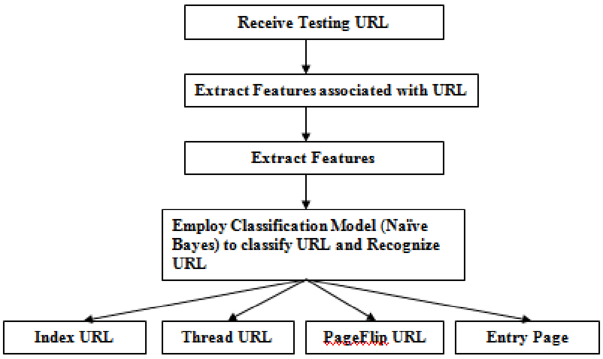

Receive Testing URL
Extract Features associated with URL
Extract Features
Employ Classification Model (Naïve Bayes) to classify URL and Recognize URL
Index URL
Thread URL
PageFlip URL
Entry Page

Figure 3: Proposed System Process

INPUT: Training set, URLs to be tested.
OUTPUT: classification of URLs
Step 1: For given feature calculate its sub features for training purpose using the training set.
Step 2: The classifier is created from the training set using by calculating mean and variance of each sub feature.
Step 3 : Probability of individual class is calculated.
Step 4: Testing sample with their calculated feature is taken for classification.
Step 5: Probability for each class (entry, index, page-flip, thread) is calculated.
Step 6: Analyze probability values of each class.
Step 7: Among Four classes, class with greater value of Probability is assigned to testing URL.

2. Online crawling

Given a forum, the proposed system first learns a set of ITF regexes then it executes online crawling using a breadth-first strategy(BFS). It first pushes the entry URL into a queue; next it fetches a URL from the URL queue and downloads its page; and then it pushes the outgoing URLs that are accorded with any learned regex into the URL queue.

Proposed Application Specifications

We can use our proposed system (supervised web forum crawler) to local search engine for one particular topic like education, medical forums. In what way means Download and maintain their web pages using this local server. Filter some kind of URLs which is unwanted while surfing their request based on some more combination of algorithms and techniques with this.
The proposed application using above crawling strategy has the following modules as shown in figure 4.

- Simple Crawler module for URL Detection
- Crawled Page Storage
- Local Search Engine System

Initially a simple crawler design for forum sites will detect and classify both index and page flipping URL and make them as training data. When the entry URL entered which will check with previous data and produce the resultant URLs that is stored in the database index. If the user search data is similar to that database detail retrieve the needed content by the URL and HTML based search engines.

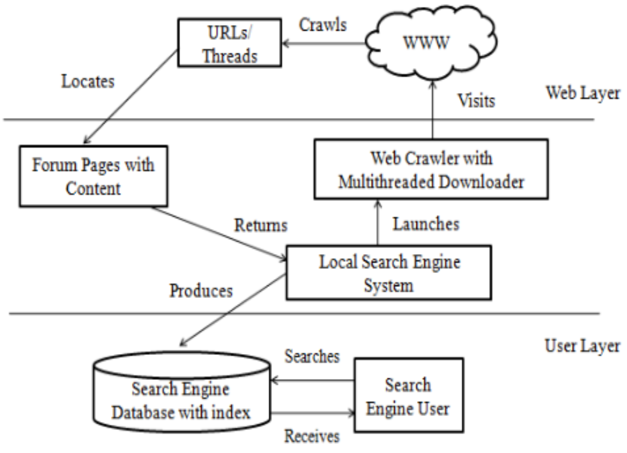
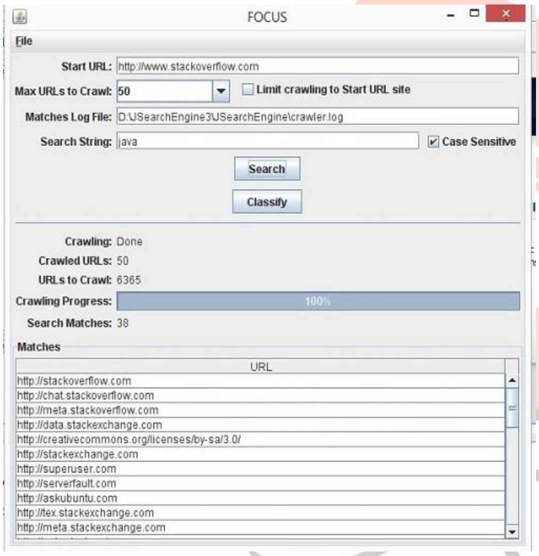


Figure 4: Application

V. IMPLEMEMTATION

The naïve Bayes algorithm has been implemented for forum URL classification.

The forum crawling process is shown in figure 5 for forum site **www.stackoverflow.com** and forums URL are classified into index URL, thread URLs, Page flip URLs and other URL and untagged URLs are shown in figure 6.

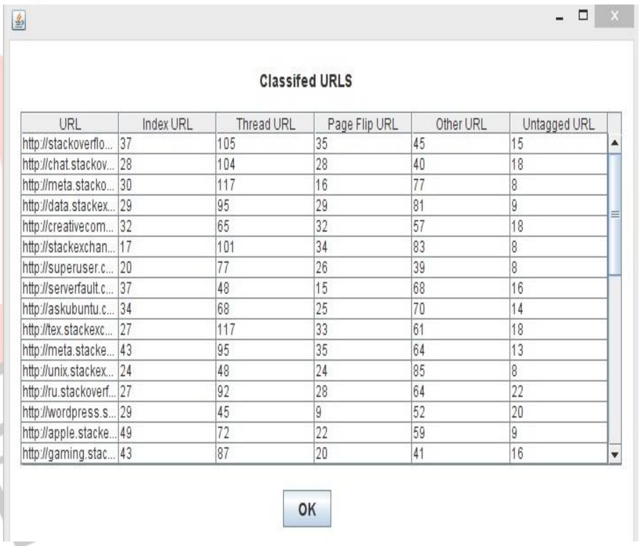


Start URL: http://www.stackoverflow.com
Max URLs to Crawl: 50
□ Limit crawling to Start URL site
Matches Log File: D:\JSearchEngine3\JSearchEngine\crawler.log
Search String: java
☑ Case Sensitive

Crawling: Done
Crawled URLs: 50
URLs to Crawl: 6365
Crawling Progress: 100%
Search Matches: 38

Matches

URL
http://stackoverflow.com
http://chat.stackoverflow.com
http://meta.stackoverflow.com
http://data.stackexchange.com
http://creativecommons.org/licenses/by-sa/3.0/
http://stackexchange.com
http://superuser.com
http://serverfault.com
http://askubuntu.com
http://tex.stackexchange.com
http://meta.stackexchange.com



Classifed URLS

URL	Index URL	Thread URL	Page Flip URL	Other URL	Untagged URL
http://stackoverflo...	37	105	35	45	15
http://chat.stackov...	28	104	28	40	18
http://meta.stacko...	30	117	16	77	8
http://data.stackex...	29	95	29	81	9
http://creativecom...	32	65	32	57	18
http://stackexchan...	17	101	34	83	8
http://superuser.c...	20	77	26	39	8
http://serverfault.c...	37	48	15	68	16
http://askubuntu.c...	34	68	25	70	14
http://tex.stackexc...	27	117	33	61	18
http://meta.stacke...	43	95	35	64	13
http://unix.stackex...	24	48	24	85	8
http://ru.stackoverf...	27	92	28	64	22
http://wordpress.s...	29	45	9	52	20
http://apple.stacke...	49	72	22	59	9
http://gaming.stac...	43	87	20	41	16

Figure 5: Crawling process

Figure 6: URL classification

VI. RESULT ANALYSIS

Proposed method is efficient in learning ITF regexes and is effective in detection of index URL, thread URL, page-flipping URL, and forum entry URL. Experimental results on 5 forum sites each powered by a different forum software package confirm that We also showed that the proposed method can effectively apply learned forum crawling knowledge on forums to automatically collect index URL, thread URL, and page-flipping URL string training sets.

ID	Forum name	Threads	Crawled threads	Crawled pages
1.	http://forums.afterdawn.com/	581137	578097	578097
2.	http://forums.asp.net/	83443	72927	91219
3.	http://forum.xdadevelopers.com/	383043	331962	412696
4.	http://forums.crackberry.com/	601230	564525	751726
5.	http://forums.gentoo.org/	869001	741068	990689

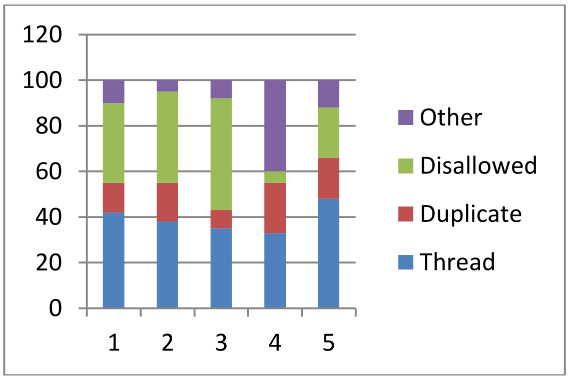


Table 1: Forum sites used in online crawling evaluation

Figure 7: Ratio of different URLs discovered

VII. CONCLUSION

In this paper, we proposed and implemented 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 regular expression ITF regexes unambiguously. These learned regexes could be applied directly in online crawling. Though, the method introduced in this paper is targeted at forum crawling, the contained EIT-like path also apply to other sites, such as community Q&A sites, blog sites, and so on. In this work no need to consider page score and weights for analyzing the web pages of forum sites. By using some data mining approaches with web sources the crawler function was discussed and developed. We would like to conduct comprehensive experiments to further verify our approach and improve upon it.

REFERENCES

[1] 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.

[2] 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.

[3] J. Jiang, X. Song, and N. Yu, “FoCUS: Learning to Crawl Web Forums,” IEEE Trans. Knowledge and Data Engg, pp. 1293-1306,2013.

[4] 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.

[5] Gautam Pant, Padmini Srinivasan, and Filippo Menczer, “Crawling the Web,” Department of Management Sciences.

[6] Namrata H.S Bamrah , B.S. Satpute, Pramod Patil “Web Forum Crawling Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.

[7] Web Crawler, http://en.m.wikipedia.org/wiki/Web_Crawler, 2015