WEB FORUM CRAWLING: A REVIEW

Namrata H.S. Bamrah, Prof. B.S. Satpute, Swapnil Kulkarni,

Address for Correspondence
1 M.E. Computer D.Y.P.I.E.T, University of Pune, Pune MH, India
2 D.Y.P.I.E.T, University of Pune, Pune, MH, India
3 Director, Ethika Engineering Solutions (I) Pvt.Ltd, Pune, MH, India

ABSTRACT
The web contains large data and it contains innumerable websites that is monitored by a tool or a program known as Crawler. It is very difficult to retrieve relevant information from the web. The main goal of this paper is to focus on the web forum crawling methods. In this paper, the various methods of web forum crawler are discussed. The paper also gives the overview of web forums.

KEYWORDS Web Forums, Crawling Policies, Structure-Driven Crawler Generation, iRobot, Board Forum Crawling, FoCUS.

I. INTRODUCTION
A crawler is a program that is used to download and store Web pages, often for a Web search engine. A Crawler traverses the World Wide Web in a systematic manner with the intention of gathering data or knowledge or for the aim of web indexing. Web crawler is also referred as robot or a spider. A web crawler could be a system for the bulk downloading of websites. A crawler starts off by placing an initial set of URLs, in a queue, where all URLs to be retrieved are kept and prioritized. The crawler gets a URL in some order from this queue, downloads the page, extracts any URLs within the downloaded page, and then in the queue it puts the new URLs. This whole process is continued. Finally the collected pages are used later for other applications, like for Web search engine or a Web cache.

Internet forum/message board/web forum is an online discussion site where users hold conversations in the form of posted messages which are posted on the forum known as post. Web forums are used to request and exchange information with each other. As web contains millions of web pages, it is very difficult to obtain relevant information that the user has demanded from that particular search engine. For example Google crawls innumerable pages per day but it takes weeks to crawl the whole web because of unrelated and unwanted pages. So it is a difficult job to find relevant information, for this crawler is used. The breadth first strategy crawler crawls the web and stores all the relevant data and show hyperlinks as a result, but because of this the database becomes too large to handle. If this drawback is handled then it’s simple for the user to get his desired data and also the size of database up to an extent will be reduced. So to avoid this drawback a crawler is needed that searches only the subset of the World Wide Web not the whole web, but for this a crawler has to address two problems. The first problem is, for deciding which pages to download next, it should have good strategy and the second problem is that it must have a reliable system that maintains and manages harmful results or effects of system crashes.

II. CRAWLING POLICIES
There are four crawling policies. The behavior of a Web crawler depends on the outcome of a combination of policies [9]:

A. Selection Policy
Selection Policy is used to download the appropriate pages, that is, it states which pages to download.

B. Re-visit Policy
It states when to check for changes to the pages. There are two simple re-visiting policies:

Uniform policy: All pages in the collection with the same frequency are re-visited.

Proportional policy: The pages that change more frequently are re-visited. The visiting frequency is directly proportional to the (estimated) change frequency.

C. Politeness Policy
Politeness Policy states to avoid overloading web sites.

D. Parallelization Policy
A crawler that runs multiple processes in parallel is known as parallel crawler. It avoid repeated downloads of the same page. During the crawling process the crawling system requires a policy for assigning the new URLs discovered, to avoid downloading the same page again and again, as the same URL can be found by two different crawling processes. In simple words, it states how to coordinate distributed web crawler.

III. WEB FORUMS
The Web forum, or message board, is an internet discussion site. On web forums people can hold conversations in the form of posted messages. They are not like chat rooms in that messages are at least temporarily archived. Also, a posted message might need to be approved by a moderator before it
becomes visible depending on the access level of a user or the forum set-up.

A forum consists of a tree like directory structure. The top level is "Categories". A forum is divided into categories for the relevant discussions. The sub-forums are further having more sub-forums. The topics or threads come under the lowest level of sub-forums and these are the places under which members will begin their discussions or posts. Forums are organized into a finite set of generic topics with one main topic, driven and updated by a group referred as members, and governed by a group referred as moderators. A message submitted by a user is known as post.

To visit a post that is message in a board, we usually starts from the homepage that is entrance page, then enters into a board, and finally find the post. This process states that the Web forum is structurally organized. Actually there exist 3 kinds of pages in a Web forum site: homepage and post page and board page.

- The entrance page of the site is a homepage; for example, the homepage of site www.ubuntuforums.org is shown in Fig 1.
- A post page records some messages provided by authors, and the post page contains exact information of the Web forum site.
- A board page contains a link index of some post pages in one board. For example, a board page of site www.ubuntuforums.org is shown in Fig 2.

![Fig 1. An example of Homepage](image1)

![Fig 2. An example of board page](image2)

IV. METHODS OF WEB FORUM CRAWLER

A. Board Forum Crawling

To crawl Web forum, Board Forum Crawling [6] is used. This method exploits the organized characteristics of the Web forum sites and simulates human behavior of visiting Web Forums. This technique starts crawling from the homepage, and so enters every board of the site, and so crawls all the posts of the site directly. Board Forum Crawling (BFC) will crawl most meaningful data of a Web forum site efficiently and simply. It cannot avoid duplicates links.

From a board page, a small link index of post pages in one board is extracted. From all board pages of a Web forum site, a whole link index of all post pages in the site is created. By the whole link index, the post pages of the site are downloaded efficiently and simply.

**Algorithm:**

Input: an entrance page that is a homepage of a Web forum site

Output: most post pages in the site

Method:

Step1. From the homepage, extract board page seeds.

Step2. For each board page seed, get a link queue of all subsequent board pages in the same board with the input seed.
Step 3. For each queue, download each page in the queue, and identify whether it is exactly a board page and extract links of post pages from the board page. At last create a whole link index of all post pages in all board pages.

Step 4. Download post pages linked by the whole link index gained in Step 3.

**B. Structure–Driver Crawler Generation**

For learning regular expression patterns of URLs that lead a crawler from an entry page to target page, Structure–Driver Crawler Generation [7] technique is used. By comparing DOM trees of pages with a preselected sample target page, target pages were found. It is very effective. The main drawback of this technique is that it only works for the specific site from which the sample page is drawn. For a new site the same process has to be repeated every time. Therefore, it is not appropriate for large-scale crawling.

**C. iRobot**

It is an intelligent crawler for Web Forums. The fundamental step in many web applications is web forum crawling problem, such as search engine and web data mining. Web forum crawling is not a trivial issue due to the in-depth link structure, the massive amount of duplicate pages and many invalid pages caused by login failure problems. For this, prototypes of an intelligent forum crawler is proposed and build known as iRobot [9], which has intelligence to grasp the content and therefore the structure of a forum site, and then decide the way to select traversal paths among different kinds of pages. This technique includes:

- **Repetitive Region based Clustering**: which automatically group forum pages with a similar content layout that has same template. By investigating the sampled pages it discover all possible repetitive patterns, and then it generates a description in the feature space for each page.

- **URL-based Sub-clustering Pages**: The main problem is that, in same layout cluster the pages can have different URL format and this is caused because of invalid or duplicate pages. Different URLs have almost the same contents and page layout. To avoid this problem, each layout clusters are split further into subsets or sub-cluster by grouping those pages with similar URL formats.

The main drawback of this technique is its tree-like traversal path which does not allow more than one path and its URL location might become invalid when the page structure changes. iRobot does not deal with the frequent thread updating in forum. No clear segregation of page identification is carried out in iRobot.

**D. FoCUS (Forum Crawler under Supervision)**

To Forum Crawler under Supervision (FoCUS) [11] is a supervised web-scale forum crawler. FoCUS crawls relevant forum content from the web with minimal overhead. FoCUS crawls the selected pages based on predefined set of topics. FoCUS learns uniform resource locator (URL) patterns across multiple sites and automatically finds a forum’s entry page given a page from the forum. FoCUS is effective for large-scale forum crawling. FoCUS defines EIT path which permit over one path and URL patterns would not be affected by a change in page structure.

This technique uses EIT path to traverse from entry pages through a sequence of index pages to thread pages. EIT means Entry – Index – Thread path. Index URLs are the links between an entry page and an index page or between two index pages. Thread URLs are the links between an index page and a thread page. Page-flipping URLs are the links connecting multiple pages of a board and multiple pages of a thread. To traverse EIT paths that lead to all thread pages a crawler should starts from the entry URL and needs to follow index URL, thread URL and page-flipping URL. EIT paths and URL patterns are more robust than the traversal path and URL location feature in iRobot. The main advantage of FoCUS is that it can avoid duplicates without duplicate detection. This technique includes:

**Entry URL Discovery**: An entry URL is needed to start the crawling process. Manual forum entry URL annotation is not practical in web-scale crawling. Entry URLs vary from forums to forums. A heuristic rule is developed to find entry URL as a baseline to demonstrate this. The heuristic baseline tries to search out the keywords ending with “/” in a URL: forum, community, board and discuss the path from the URL host to this keyword is extracted as its entry URL if a keyword is found; if not, the URL host is extracted as its entry URL. Forum entry URL discovery method is designed to make FoCUS more scalable, practical, simple yet effective based on some techniques introduced in previous sections.

**Algorithm: EntryUrlDiscovery**

Input: url1: a URL pointing to a page from a forum
Output: e_url: Entry URL of this forum

1. b_url = GetNaiveEntryUrl(url1);
2. p2 = download(b_url);
3. urlss = download all outgoing URLs in p2 that begin with b_url;
4. sam_urls = Randomly sample a few URLs from urls;
5. Add the host of url1 into sam_urls;
6. foreach u1 in sam_urls do
7. p2 = Download(u);
8. urlss = urlss ∩ (outgoing URLs in p2 starting with b_url);
9. endforeach
10. let e_url be b_url, index_urls be ϕ, count be 0;
11. foreach url1 in urlss do
12. if url1 is in index_urls continue;
13. p2 = Download(url1);
14. i_urls = Detect index URLs ⊂ i_urls;
15. if count < | i_urls |
16. count = | i_urls |
17. e_url = url1;
18. endif
19. endforeach
20. return e_url;

**V. 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 different methods of web forum crawling 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.

REFERENCES