Web Forum Crawling Techniques

Namrata H.S Bamrah
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering
Pimpri Pune India

B.S. Satpute
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering
Pimpri Pune India

Pramod Patil
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering
Pimpri Pune India

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.

Keywords

Web Crawling, Web Forums, FoCUS.

1. INTRODUCTION

The World Wide Web consists of many web pages with billions of documents. As the web pages increases the contents on web pages is also growing dynamically like news, monetary knowledge, diversions, financial data, entertainment and schedules. So it is very difficult to obtain relevant information on the web pages that the user has demanded from that particular search engine. For Example the major search engine like Google that crawls innumerable pages per day takes weeks to crawl the whole web. So to find relevant information a crawler is used.

The breadth first strategy [4] crawler crawls the web and stores all the relevant data and show hyperlinks as a result. Due to which 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.

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.

Web crawlers are used for a many purposes. They are the main components of web search engines, systems that assemble a corpus of websites, index them, and permit users to issue queries against the index and find the pages i.e. web pages that match the queries.

2. USES OF WEB CRAWLING

2.1 Web Archiving

It is a service where large sets of web pages are collected periodically and archived for posterity, provided by the Internet archive.

2.2 Web Data Mining

In Web data mining web pages are analyzed for some statistical properties or where different data analytics is performed on them. Attributor, a company that monitors the web for copyright and trademark infringements can be the example.

2.3 Web Monitoring Services

The web monitoring services allows their clients to trigger or submit standing queries, and they crawl the web continuously and notify the clients of pages that match those submitted queries.

3. APPLICATION OF WEB CRAWLING

3.1 MySpiders: Query-Time Crawlers

MySpiders is a Java application programme that implements the InfoSpiders and the naïve best-first algorithms. The application programme is available online. Multithreaded crawlers are started when a user submits a query. As the crawler finds “good” pages results are displayed dynamically. The user might browse the results whereas the crawling continues in the background. Hence, every thread is more independent with non-contentious access to its frontier. Hence, every thread is more independent with non-contentions access to its frontier.

The applications programme permits the user to specify the crawling rule and therefore maximum number of pages to fetch. In order, the system uses the Google Web API11 to obtain a few seeds pages to initiate the crawl. The crawler threads are started from each of the seeds and the crawling continues till the desired numbers of pages are fetched or the frontier is empty.

3.2 Mapuccino: Building Topical Site Maps

To building website maps is to start from a seed URL and crawl in a breadth first manner till a definite range of pages have retrieved or a definite depth has been reached. The site map might then be displayed as a graph of connected pages.

However, if anyone is curious about building a website map that focuses on a definite topic, then the above mentioned approach can lead to a large range of unrelated pages as tend to crawl to larger depths or fetch more pages. Mapuccino corrects this exploitation using shark-search to guide the crawler and then build a graph that highlights the relevant pages.

3.3 Letizia: A Browsing Agent

Letizia is an agent that assists a user throughout browsing. Letizia tries to understand user interests based on the pages being browsed, while the user surfs the Web. The agent then follows the hyperlinks starting from the current page being browsed to find pages that could be of interest to the user. The hyperlinks are crawled automatically and in a breadth-first manner.

4. CHALLENGES OF WEB CRAWLING

4.1 Scale

The web is incredibly large and regularly evolving. Crawlers that get broad coverage and good freshness should achieve extremely high throughput that poses several difficult engineering issues. Modern search engine companies use thousands of computers and dozens of high-speed network links.

4.2 Content Selection/Choice Tradeoffs

Even the highest-throughput crawlers don’t crawl the entire web, maintain with all the changes. Crawling is performed selectively and in a carefully controlled order. It should acquire high-value content quickly. It ensures coverage of all affordable content, and bypass low-quality, irrelevant content, redundant content, and malicious content.

The crawler should balance competitor objectives like coverage and freshness, whereas obeying constraints like per site rate limitations. A balance should even be struck between exploration of potentially helpful content, and exploitation of content already known to be useful.

4.3 Adversaries

Some content suppliers get to inject useless or misleading content into the corpus assembled by the crawler. Such behavior is usually impelled by financial incentives, for example mis-directing traffic to business web sites.

4.4 Social Obligations

Crawlers are good citizens of the web. They should not impose large amount of burden on the web sites they crawl. In fact, a high-throughput crawler can inadvertently carry out a denial-of-service (DoS) attack without the right safety mechanisms.

5. WEB CRAWLER ARCHITECURE

Figure 1 below shows the architecture of web crawler.

URL Frontier: It contains URLs to be fetches in the current crawl. At first, in URL Frontier a seed set is stored, and by taking a URL from the seed set a crawler begins.

DNS: DNS is domain name service resolution and it look up IP address for domain names.

Fetch: It is used to fetch the URL for that it uses the http protocol.

Parse: It is used to parse the page. In this text, images, videos etc. and Links are extracted.

Content Seen? : It is used to test whether a web page with the same content has already been seen at another URL or not. It develops a way to measure the fingerprint of a web page.

URL Filter: it tells whether the extracted URL should be excluded from the frontier (robots.txt) or not. URL should be normalized (relative encoding).

Dup URL Elim: Dup URL Elim is used to check the URL for duplicate elimination.

The process of crawling consists of many threads and every thread performs work cycles repeatedly. The following describes the work cycle of the thread:

First the user thread obtains a URL from URL Frontier. Then the Frontier according to its priorities and politeness policies pops out a URL. After that, the Fetcher module calls the DNS module and the DNS Module resolves the host address of the corresponding web server. Once you got the address, the Fetcher module connects to the server and checks for robots exclusion protocol and then tries to download the web page.

If the web page is downloaded, the web page is sent to Link Extractor. Link Extractor will extract the outgoing links. To the URL filter the extracted URLs are passed. The URL filter, filters out unwanted URLs like file extensions of no interest, black listed sites etc. Then to duplicate eliminator these filtered URLs are passed, which removes the URLs already present in the Frontier. Finally the URLs reach the URL Frontier, allots them to the specific positions in its data structure according to some fixed priority rules.

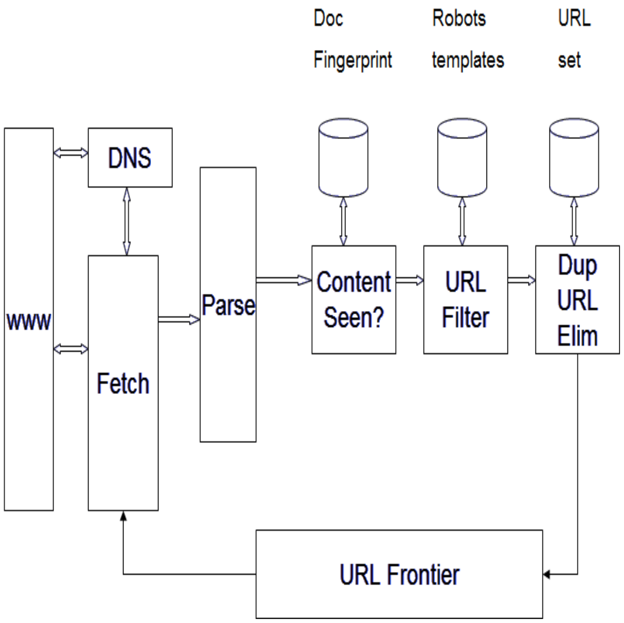


Fig 1: Architecture of Web Crawler

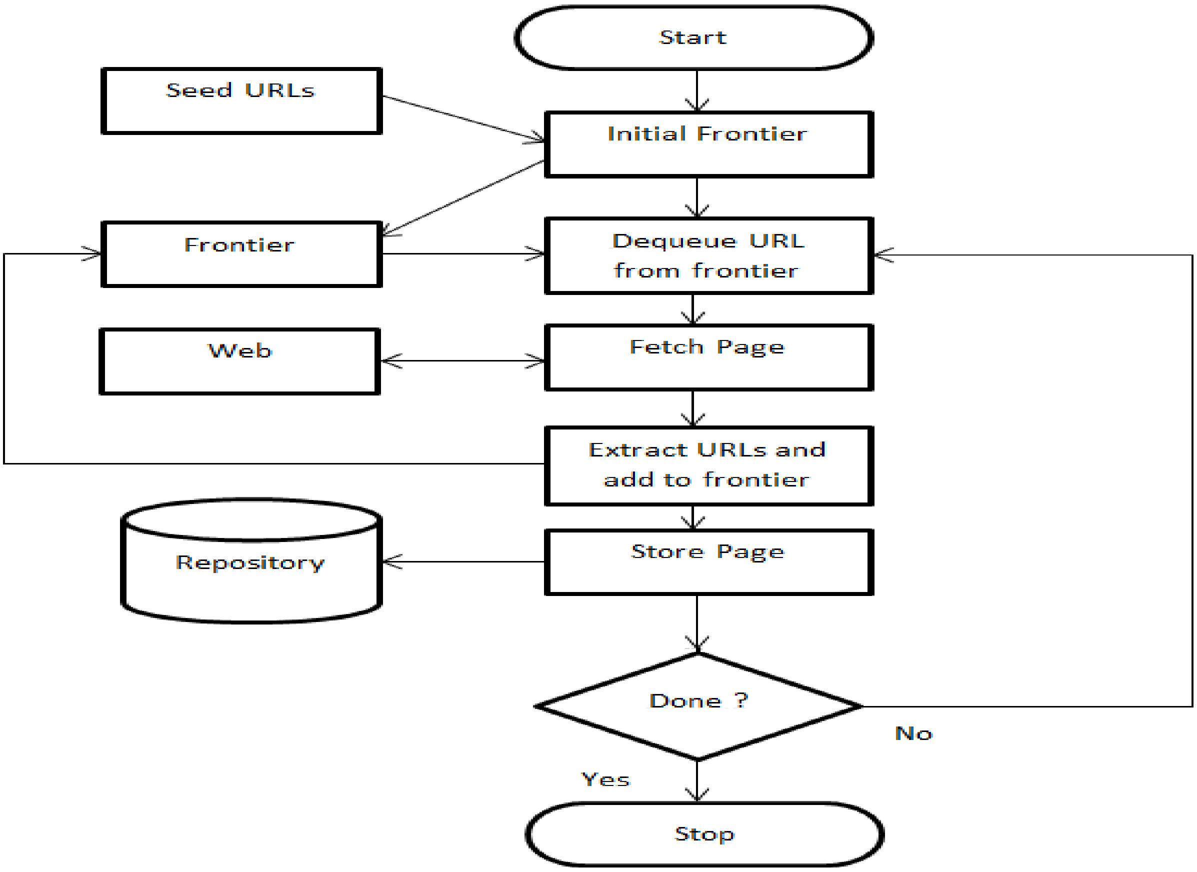


Fig 2: Process of Crawler

6. CRAWLING POLICIES

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

6.1 Selection Policy

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

6.2 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.

6.3 Politeness Policy

Politeness Policy states t to avoid overloading web sites.

6.4 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.

7. PROCESS OF CRAWLING

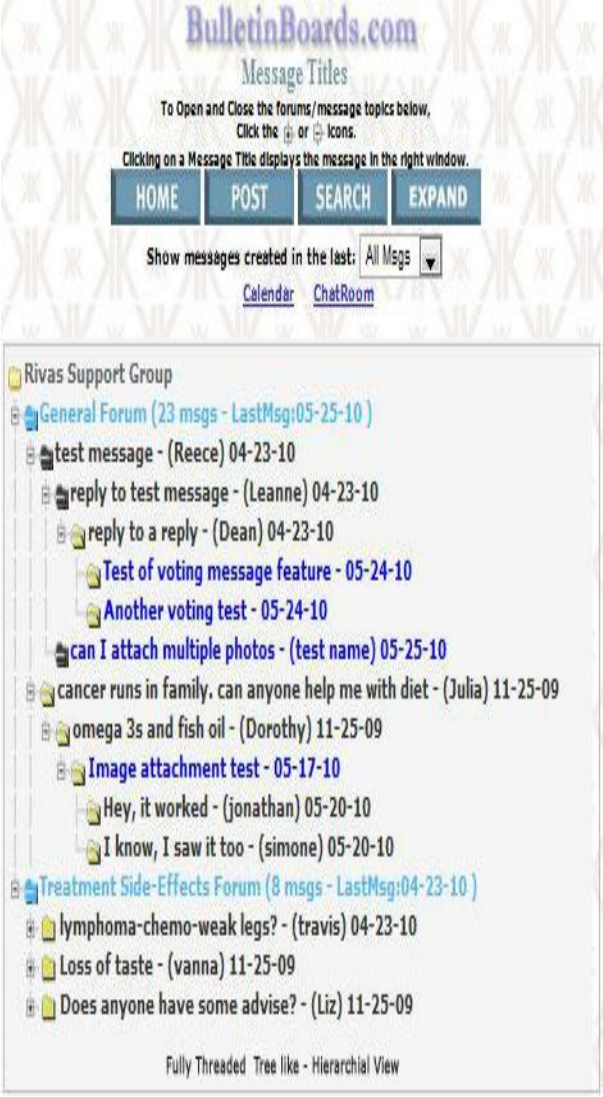
Above Fig 2 is a sequential crawler. The list of starting URLs can be a seeds. Frontier data structure is used to determine order of page visits. Stop criterion can be anything. A generic crawler adopts a BFS strategy. A generic crawler is usually inefficient and inefficient for forum crawling.

8. 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. The Fig 3 below shows the example of tree structure of forum.

Forums have a selected set of jargon related to them. Thread or topic is a single conversation. A discussion forum is a tree-like in structure: a forum will contain a number of sub-forums, each of which may have several topics. Inside a forum's topic, every new discussion started is called a thread, and may be replied to by as many people as so wish. Users can be anonymous or have to register with the forum and then subsequently log in, in order to post messages depending on the forum's settings. In many forums, users do not have to log in to read existing messages.

The Fig 4 below shows the example of link relations in forums that is it shows the typical page and link structure in a forum [8]. In the fig 4 entry URL is not necessarily at the root URL level of a forum hosting site. Its form varies from site to site. The user can navigate from entry page to thread page by many paths. For example entry – board – thread, entry – list of board – board – thread or entry – list of board – list of board & thread - board – thread.



BulletinBoards.com

Message Titles

To Open and Close the forums/message topics below, Click the ⊞ or ⊟ Icons.

Clicking on a Message Title displays the message in the right window.

Show messages created in the last: All Msgs

Calendar ChatRoom

- Rivas Support Group
  - General Forum (23 msgs - LastMsg:05-25-10)
    - test message - (Reece) 04-23-10
      - reply to test message - (Leanne) 04-23-10
        - reply to a reply - (Dean) 04-23-10
          - Test of voting message feature - 05-24-10
          - Another voting test - 05-24-10
      - can I attach multiple photos - (test name) 05-25-10
    - cancer runs in family. can anyone help me with diet - (Julia) 11-25-09
      - omega 3s and fish oil - (Dorothy) 11-25-09
        - Image attachment test - 05-17-10
          - Hey, it worked - (jonathan) 05-20-10
          - I know, I saw it too - (simone) 05-20-10
  - Treatment Side-Effects Forum (8 msgs - LastMsg:04-23-10)
    - lymphoma-chemo-weak legs? - (travis) 04-23-10
    - Loss of taste - (vanna) 11-25-09
    - Does anyone have some advise? - (Liz) 11-25-09

Fully Threaded Tree like - Hierarchial View

Fig 3: Example of Forum tree Structure

9. FORUM STRUCTURE

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*.

It can even have a graph structure. All message boards can use one of 3 possible display formats. Each of the 3 basic message board display formats: Non-Threaded / Semi-Threaded / Fully-Threaded, has some advantages and disadvantages. If messages aren't associated with each other at all a Non-Threaded format is best. If a user includes a message topic and multiple replies to that message topic a semi-threaded format is best. If a user includes a message topic and replies to that message topic, and replies to replies, then it is fully threaded.

Thread: - A collection of posts are known as thread or topic. It is displayed from oldest to latest or newest to oldest. The format of the thread includes thread title, an additional description that describes the summarization of the intended discussion and an opening or original post/poster (OP). OP opens whatever dialogue or makes whatever announcement the poster wished. A thread contains any number of posts. A thread can include multiple posts from the same members, even if they are one after another.

When any member posts in a thread it will appear on the top because it is the latest updated thread. Forum contains a thread. Forum also contains the date of the last post. An important thread which rarely received posts are known as stickyed or pinned, which is always appear in front of normal threads. The reply counts on forums are used to measure threads. Some forums also use or track page views.

Post: - A message submitted by a user is known as post. This message is enclosed into a block. This block contains user details, date and time of the submitted message. Threads can contain posts. Members can edit or delete their own posts. In thread posts appear as blocks one after other.

The first post is known as thread starter (TS) or original post (OP). Forums can also keep the track of user postcount, which is measurement of how many posts a user has made. User with has highest postcounts are more reputable than user with lowest postcounts, but this not always considered.

Moderators: - The moderators also referred as mod are users or employees of the forum. The moderators have granted access to the posts and it keeps the forum clean from the spam, spambots etc. Moderators respond to the specific complaints of the members, answers the general questions of the members. Moderators can delete, move, split, and merge the posts and threads. It can also rename, lock, ban, unban, suspend, un-suspend, add, edit, warning the members and remove the polls of the threads. Moderators should mange day-to-day details of the forum or board.

Administrator: - The administrators also referred as admin are very important. For running the site they manage technical details. They are the main person of the site. They can promote or demote members to or from moderators. They also create sections, sub-sections, mange the rules, change the skin of the forum that is appearance of the forum, performs any database operations like database backup in the forum. They can act as moderators. Administrators also share their knowledge in many forums.

10. TECHNIQUES OF WEB CRAWLER

10.1 Board Forum Crawling

To crawl Web forum, Board Forum Crawling [5] 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 without duplicate detection.

Method:

Input of BFC is a homepage of a Web forum site. Output of the BFC is most post pages in the site. In BFC, first it will extract from homepage the board page seeds. Then it will select a link queue of all subsequent board pages in the same

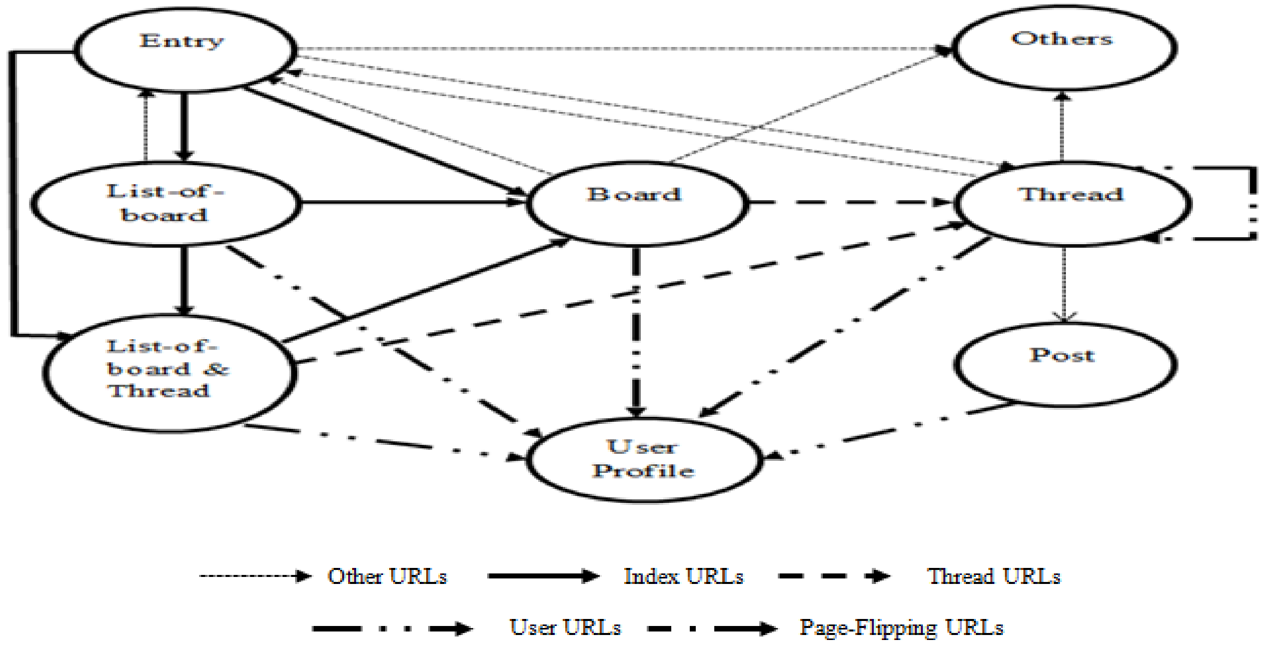


Fig 4: Example of Link relations in Forums

board with the input seed for each board page seed. Then it will download each page in the queue, and then it will identify whether it is a board page or not for each queue. And then creates a link index. The link index contains all the post pages in all board pages. At last it will download post pages linked by the whole link index gained.

10.2 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 [6] 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.

10.3 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 [7], 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.

10.4 FoCUS (Forum Crawler under Supervision)

To Forum Crawler under Supervision (FoCUS) [8] is a supervised web-scale forum crawler. FoCUS crawls relevant forum content from the web with minimal overhead. FoCUS learns uniform resource locator 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.

It shows way to learn regular expression patterns (ITF regexes) that recognize the index uniform resource locator (URL), thread uniform resource locator (URL) and page-flipping uniform resource locator (URL) using the page classifiers. FoCUS adopts a simple URL string de-duplication technique. The main advantage of FoCUS is that it can avoid duplicates without duplicate detection.

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.

11. 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.

12. REFERENCES

[1] Chakrabarti, S. (2003). Mining the Web: Discovering Knowledge from Hypertext Data. San Francisco, CA: Morgan Kaufmann.

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

[3] Castillo, Carlos (2004), “Effective Web Crawling”, University of Chile.

[4] R. A. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez, May 2005, “Crawling a country: better strategies than breadth-first for Web page ordering” In Proc. 14th WWW, pages 864−872, Chiba, Japan.

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

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

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

[8] Jingtian Jiang, Xinying Song, Nenghai Yu and Chin-Yew Lin, 2013 “FoCUS: Learning to Crawl Web Forums,” Proc. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6.

[9] Raja Iswary, Keshab Nath, October 2013, “Web Crawler”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2.