ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

A Survey on Web Forum Crawling Techniques

Anu K P[1], Rejimoan R[2]

P.G. Student, Department of CSE, SCTCE, Trivandrum, Kerala, India[1]

Asst. Professor, Department of CSE, SCTCE, Trivandrum, Kerala, India[2]

ABSTRACT: Web forum has become an important source on the Web due to its rich information contributed by millions of users. But still web forum crawling is a challenging task due to complex in-site link structures and login controls of most forum sites. The main goal of this paper is to focus on various web forum crawling techniques. Board Forum Crawling is a technique which exploits the organized characteristics of the Web forum sites and simulates human behaviour of visiting Web Forums. A recent and more comprehensive work on forum crawling is iRobot. iRobot aims to automatically learn a forum crawler with minimum human intervention by sampling pages, clustering them, selecting informative clusters via an informativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection. Forum Crawler Under Supervision (FoCUS) is a supervised web-scale forum crawler. The goal of this technique is to crawl relevant forum content from the web with minimal overhead. Forum threads contain information content that is the target of forum crawlers.

KEYWORDS: WebCrawler, Forum, iRobot, FoCUS.

I. INTRODUCTION

Nowadays Web forum (also called bulletin or discussion board) is very popular all over the world for opening discussions. With millions and millions of user contribution, forum data usually has tremendous collection of highly valuable knowledge and information. Commercial search engines such as Google, Yahoo!, have begun to leverage information from forums to improve the quality of search results. For this, search engines have to first download pages from various forums and build a locally indexed repository [1]. High quality search results highly depend on a high quality repository. This requires the crawler to fetch as much as possible valuable data from various Web forum pages [2], using the limited bandwidth and storage space.

However, traditional generic crawlers [1] which basically adopt the breadth-first traversal strategy [3] are usually inefficient in crawling forum sites. This is mainly due to noncrawlerfriendly characteristics of forums [4], [5]: 1) duplicate links 2) uninformative pages 3) page-flipping links and 4) Entry URL discovery.

A forum typically has many duplicate links that point to a common page but with different URLs , e.g., shortcut links pointing to the latest posts or URLs for user experience functions such as “view by date” or “view by title.” A generic crawler will crawl many duplicate pages if they blindly follows these links, making it inefficient.

A forum also has many uninformative pages such as login control to protect user privacy or forum software specific FAQs. Following these links, a crawler will crawl many uninformative pages thus making it inconsistent for forum crawling.

Usually a long forum boards or threads are divided into multiple pages that are linked by page-flipping links. Those links should be preserved to facilitate tasks such as page wrapping and content indexing etc. Besides the above challenges, there is the problem of entry URL discovery. The entry URL of a forum points to its homepage. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

crawling should start with the entry URL rather than from nonentry URLs for the efficient forum crawling process. In addition, for a forum crawling technique to yield high performance, forum entry URL discovery is required.This paper gives the overview of various forum crawling techniques like iRobot, FOCUS. The rest of the paper is orga-nized as follows. First we briefly describewebcrawler and its architecture and then the structure of forums is explained in section III. In section IV, each technique is discussed in detail. Section V provides the comparison between each technique. And in last section, we draw conclusion.

II. WEBCRAWLER

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 find the pages i.e. web pages that match the queries

.

Web Crawler Architecure

The architecture of web crawler is shown in the Fig.1below. It includes:

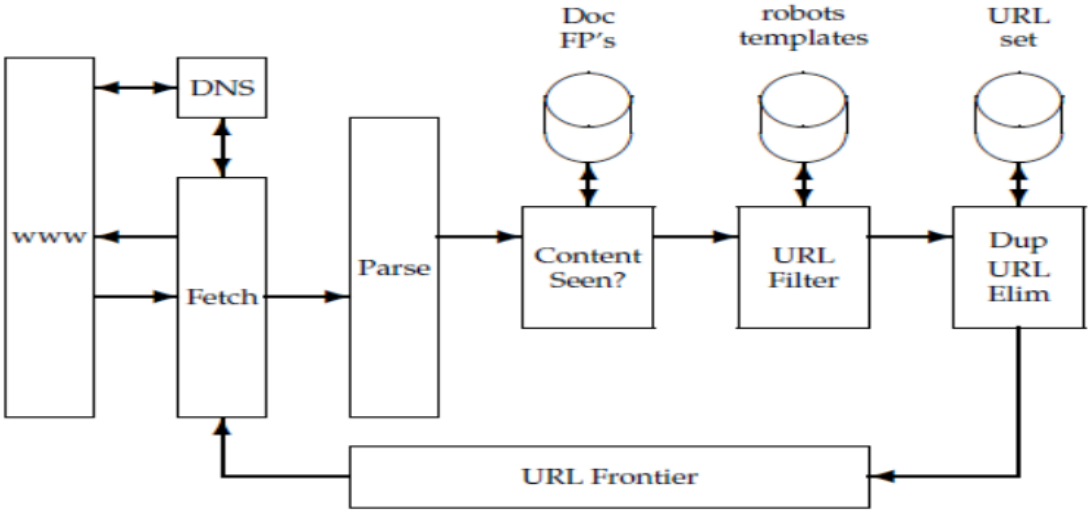

Fig. 1Architecture of WebCrawler

1. 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.
2. DNS is domain name service resolution and it look up IP address for domain names.
3. Fetch: It is used to fetch the URL for that it uses the http protocol.
4. Parse : It is used to parse the page. In this text, images, videos etc. and Links are extracted.
5. 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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

A crawler fetches the web page associated with a URL by taking the URL from the frontier using http protocol. Once fetched, the web pages are stored temporarily. Then the page is parsed and the text as well as the links in it is extracted. After the URL is parsed, the very next task is to check whether the web page with same content has already been seen at another URL. Next, a URL filter is used to determine whether the extracted URL should be excluded from the frontier based on one of several tests. The final process is to check for any duplicate URLs, if it is present in the frontier i.e. already crawled then we do not add it to the frontier.

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 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 and 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 semithreaded 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.

III. WEB FORUM CRAWLING TECHNIQUES

Board Forum Crawling (BFC)

Guo et al [6] proposed Board Forum Crawling, a Web forum technique which leverages the organized characteristics of the web forum sites. It simulates the user behaviour for the extraction of the web forum contents i.e. it starts crawling from the home page and then enters the board page and then crawls all the posts present in the site directly.
They define 3 kinds of pages in a Web forum site: homepage, post page and board page. Homepage is the entrance page of the site. Post Page contains exact information of the Web forum site and Board page contains a link index of some post pages in one board.

The method first extracts the board page seed from the homepage. Then for each board page seed, it selects a link queue of all subsequent board pages in the same board with the input seed. Then 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 list containing whole link index of all post pages in all board pages. And finally download all post pages which is in the list.
Board forum crawling technique is simple forum crawling technique, but it doesn't deal with entry URL discovery problem and it provides little support for duplicate link detection while comparing with other techniques.

iRobot: An Intelligent Web Forum Crawler

iRobot [4] automatically understands the content and structure of each forum sites and then decides how to traverse to different pages in forum sites. To find out such traversal paths, it automatically rebuild the sitemap of the target Webforum and then select an optimal navigation path which only traverses informative pages and skip invalid and duplicate ones.

Fig 2a shows the architecture of iRobot.The system mainly consists of two parts: (I) offline sitemap reconstructing and traversal path selection; and (II) online crawling. The offline part mines useful knowledge based on a few sampled pages, and guide the following massive crawling. The sampling process actually adopts a combined strategy of breadth-first and depth-first inorder to keep the sampled pages as diverse as possible. Then, through the repetitive region-based clustering module it group forum pages with similar layout. Pages in each layout cluster are further grouped intosubsets according to their URL formats by the URL-based subclustering module. This module also constructs arcsamong various vertices, where each arc is characterized by both the URL pattern and link location of the corresponding

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

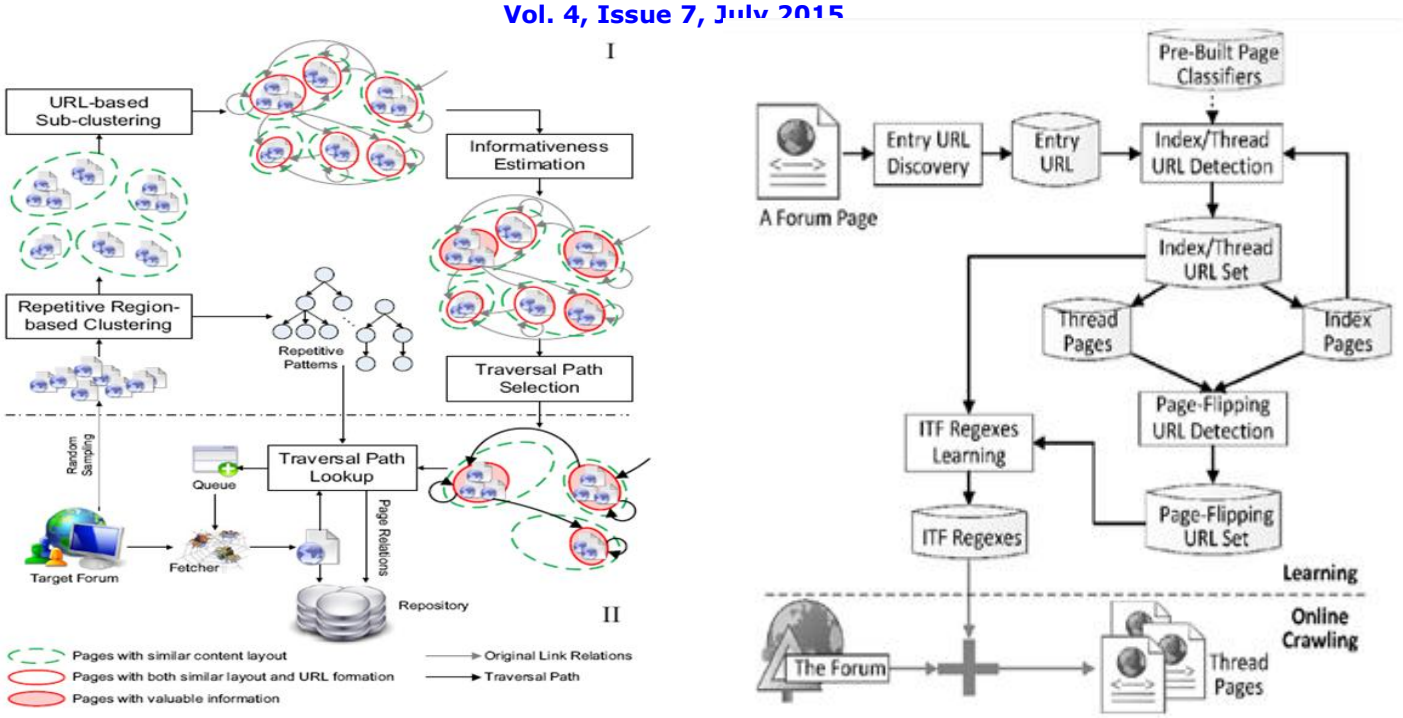

Fig.2 (a) The architecture of iRobot (b)The architecture of FoCUS

links. The informativeness estimation module then selects valuable vertices with informative pages on the sitemap removing the useless vertices with invalid or duplicate pages. Finally in the offline part, traversal path selection module will find an optimal traversal path to traverse the selected vertices and reject the discarded ones. In the online crawling part, the input page is sent to the traversal path lookup module. This module will align the input page with all the repetitive patterns and classified into one of the vertices on the sitemap; and corresponding link tables are simultaneously constructed. Then, for each link in each link table, the module decides whether it should be added to the crawling queue, by looking up the list of traversal paths.

The main advantage of iRobot system is that it requires less human intervention and it provides support for duplicate links and uninformative pages removal. But as it uses tree-like traversal path, it does not allow more than one path and its URL location might become invalid when the page structure changes. And it also does not deal with the frequent thread updating in forum.

FoCUS (FOrum Crawler Under Supervision)

FoCUS [7] is a supervised web-scale forum crawler. The main goal of FoCUS is to crawl relevant forum content with less overhead. It has reduced the forum crawling problem to a URL type recognition problem and showed how to leverage implicit navigation paths of forums.

Jiang classified forum pages into page types. Entry Page is the homepage of a forum. Index Page is a page of a board in a forum, which usually contains a table-like structure. Thread Page is a page of a thread in a forum that contains a list of posts with user generated content belonging to the same discussion. Other Page is a page that is not an entry page, index page, or thread page.

Fig.2b shows the architecture of FoCUS.The architecture of FoCUS consists of two major parts: the learning part and the online crawling part. The learning part first learns ITF regexes of a given forum from automatically constructed URL training examples. The online crawling part then applies learned ITF regexes to crawl all threads efficiently.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

FoCUS first finds the entry URL of the input page using the Entry URL Discovery module. Then it detects index URLs and thread URLs on the entry page using the Index/Thread URL Detection module. The detected index URLs and thread URLs are saved to URL training sets. Inorder to make sure that the page contains no index URLs, the destination pages of thedetected index URLs are fed again to this module. Then both index pages and thread pages are fed to the Page-Flipping URL Detection module to find pageflipping URLs and save them to the training sets. Finally, from the URL training sets the ITF Regexes Learning module learns a set of ITF regexes which is a regular expression that can be used to recognize index, thread, or page-flipping URLs. When the learning is finished, FoCUS performs online crawling. It starts from the entry URL and follows all URLs matched with any learned ITF regex. FoCUS continues to crawl until no page could be gathered or other condition is satisfied.

IV. COMPARISON

A Web crawler is an Internet boot that systematically browses the World Wide Web, typically for the purpose of Web indexing.The three main Web forum crawling techniques that has been described in this paper includes Board Forum Crawling,iRobot and FoCUS.

PROPERTIES	BOARD FORUM CRAWLING	IROBOT	FOCUS
Definition	Exploits the organized characteristics of the Web forum sites and simulates human behavior of visiting Web Forums	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	A supervised web-scale forum crawler which crawl relevant forum content from the web with minimal overhead
Duplicate Detection	Requires to avoid duplicate	Not Required	Not Required
Navigation Path	Simulates Human behaviour	Applied sampling and clustering	EIT path is used
URL Traversal	Did not mentioned how to discover and traverse URLs	Used Tree-like Traversal paths	Used EIT paths
Path Selection	Doesnot allow more than one path	Doesnot allow more than one path	Allow all the path.
Discovering new URLs	None	Learns URL location Information	Learns URL patterns
Page Structure Change	None	URL Location becomes Invalid	Not affected
Learning Efficiency	None	Need to sample more page to achieve a stable performance	Needs less page
Crawling Efficiency	None	Less;Random Sampling strategy which samples many useless and noisy pages	More;Learns EIT path and ITF regexes directly. So not affected by noisy pages
Crawling Coverage	None	Low;Suffered from changed page flippping URL location across pages , Tree-like traversal path	More;Index/thread URL and page Flipping URL detection are effective

Table 1:Comparison of forum Crawling Techniques

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Board Forum Crawling technique Exploits the organized characteristics of the Web forum sites and simulates human behaviour of visiting Web Forums. iRobot aims to automatically learn a forum crawler with minimum human intervention by sampling pages, clustering them, selecting informative clusters via an informativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection. Forum Crawler Under Supervision (FoCUS) is a supervised web-scale forum crawler. The goal of this technique is to crawl relevant forum content from the web with minimal overhead. Forum threads contain information content that is the target of forum crawlers.

Board Forum Crawling cannot avoid duplicates without duplicate detection. Whereas iRobot and FoCUS doesn't require duplicate detection module to avoid duplicates.Board Forum Crawling Technique exploits the organized characteristics of the Web forum sites and simulates human behavior of visiting Web Forums.In iRobot, sampling and clustering is applied to find out Navigational Path.For FoCUS, an entry-index-thread path is a navigation path from an entry page through a sequence of index pages (viaindex URLs and index page-flipping URLs) to thread pages (via thread URLs and thread page-flipping URLs).

iRobot learns URL location information to discover new URLs in crawling, but a URL location might become invalid when the page structure changes. As opposed to iRobot, we explicitly define entry-index-thread paths and leverage page layouts to identify index pages and thread pages. FoCUS also learns URL patterns instead of URL locations to discover new URLs. Thus, it does not need to classify new pages in crawling and would not be affected by a change in page structures.

Table 1 compares the three techniques based on Duplicate Detection, Navigation Path, URL Traversal, Path Selection, Discovering new URLs, Page Structure Change, Learning Efficiency, Crawling Efficiency, Crawling Coverage.

V. CONCLUSION

Web forum is a Web application for holding discussion and posting user-created content. However generic crawlers are inefficient in crawling forum sites due to complex in-site structures and login controls of most forum sites. The tech-niques like Board Forum Crawling, iRobot and FoCUS that is used to extract structured data from forum sites are discussed in brief and a comparative study of those techniques has been performed in this paper. Comparing with other techniques of web forum crawling, FoCUS outperforms these crawlers in terms of effectiveness and coverage by providing good support for entry URL discovery, duplicate links detection, support to deal with template variation.

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. Baeza-Yates and C. Castillo," Crawling the infinite Web: five levels are enough", Proc. 3rd Workshop on Algorithms and Models for the Web-Graph, LNCS, volume 3243, pages 156-167, Rome, Italy, Oct. 16, 2004.
[3] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez," Crawling a Country: better strategies than breadth-first for Web page ordering", Proc. 14th WWW, pages 864-872, Chiba, Japan, May 10-14, 2005.
[4] 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.
[5] 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.
[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] Jingtian Jiang, Xinying Song, Nenghai Yu and Chin-Yew Lin, "FoCUS: Learning to Crawl Web Forums," Proc. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6.2013