Survey on Techniques used to Crawl Web Forums

Parina Shah[1] Ms. Gayatri Pandi(Jain)[2]
[1]M.E Student [2]Head of the Department
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering & Technology, Ahmedabad, Gujarat, India

Abstract— A Web Crawler is a computer program that browses the World Wide Web in automated manner, methodical or in an orderly fashion. The aim is to crawl relevant forum content from the web with minimal overhead. Forums have become very popular almost all over the world as they are open for discussions. There are innumerable new posts created by millions of Internet user's everyday upon various topics and issues. Forum crawling consists of various forum sites which are crawled depending upon the user search query of the user. Forums have similar navigation paths which are connected by specific URL types to lead users from entry pages to thread pages. Crawler reduces the web forum crawling problem by a URL type recognition problem. It also shows how to learn accurate and effective regular expression patterns of various navigation paths from automatically created training sets. Patterns of URLs are extracted and crawling is made more efficient.

Key words: web crawler, FoCUS, Forum Mining

I. INTRODUCTION

A web crawler is a system for downloading of web pages. Crawlers can be used for a variety of purposes. They are one of the main components of web search engines, systems that sequence a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries. A web crawler crawling upon the forums is a technique called forum mining. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases and derives its name from the similarities between searching for valuable information in a large database. Crawling is a system for bulk downloading of web pages from internet. It crawls relevant content from World Wide Web. Web crawlers are used for indexing pages for search engines, archiving the web, analysing the web etc. Web search engines and other sites use web crawling software to adopt their web content or indexing of other site's web content. It can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Crawlers are used for validation of hyperlinks and HTML code. Other use is called web scraping.

A search engine uses the web crawler. So, forums generally are topic based. User generally makes a search based on a particular topic called Query term. So, all the posts related to that topic are crawled and the user is provided with the output in particular time interval. A post is generally the reply posted by the users or members of the website. There exist certain URLs in the posts which may redirect the user to the same page. So, duplication of URLs may exist. This creates unnecessary crawling and indexing of the same page multiple times. It not only increases the search time but also reduces the efficiency. For example, www.gtu.ac.in website may exist several times in a forum related to result or online material download. Thus, same website is crawled multiple times by the crawler. De-duplicating the pages is an important concept to be resolved.

Forums exist in many different layouts or styles and powered by a variety of forum software packages, and they always have navigation paths to lead users from entry pages to thread pages. Thus, we need to find the shortest path which may lead us from entry page to thread page.

Patterns are extracted of the Index/Thread/Flipping (ITF) URLs. Thus, these patterns are then used to find the ITF regular expressions. This ITF regexes help us to provide best results in minimal time.

So, the goal of Crawler is to crawl relevant content such as user posts from forums with minimal overhead.

In our existing system, there exist several URLs which navigate to the same page on web forums. This increases the time of a crawler due to repetitive crawling of pages. Forums have classified all the records in the web and based upon user request the information is get searched and transmitted. The posts in the forums may get constantly updated and so the crawler needs to get added in the list. Duplicate pages exist in the web crawler list. So, it reduces the efficiency of the crawler. This concept is called Page Flipping and the URLs are called Page Flipping URLs.

So, it is very much important for any crawler to uniquely identify a page URL and navigate correctly from entry page to thread page. So, pattern evaluation and regular expressions for ITF are important aspects to be focused for better crawling.

The following shows the basic elements of Forum:

A. Web Forum Structure:

A web forum is a tree like or hierarchical structure. A forum can be divided into different categories for the relevant conversations that occur and then all of the posted messages under these categories are sub-forums and these sub-forums can be further divided into more sub-forums.

B. User groups:

A member or user of the forum can automatically get access to a more privileged user group based on conditions set by the administrator. All the anonymous users of the site are known as visitors. Visitors have privilege to be granted access to all functions that do not require break their privacy. A guest or visitor can usually view the contents and then the posted messages of the forum.

C. Posts:

A post is a conversation which user makes inside a forum upon a topic. User to writes a message which is enclosed into a block containing the user's details and the date and times it was posted. There is certain limit for submitting the post. The message should be usually of minimum 10

characters and upper limit to the length of a message is 10,000 to 20,000 characters.

The following literature survey shows different methods of crawling and techniques to make crawler efficient for forum mining.

II. LITERATURE SURVEY

A. Board forum crawling: a web crawling method for web forum [1]

In [1] Yan Guo et.al presented Board Forum Crawling (BFC) which is used to crawl Web forum. This method simulates human behaviour of visiting Web Forums and exploits the organized characteristics of the Web forum sites .Thus, BFC uses a method, which starts crawling from the homepage, followed by entering each board of the site, and then crawls all the included posts of the site directly. So, the Board Forum Crawling can crawl the most meaningful information of a Web forum site very efficiently and in a simple way. They experimentally evaluated the effectiveness of the method on real Web forum sites by comparing with the traditional breadth-first crawling.

They have also used this method in a real project, and have crawled approx. 12000 Web forum sites successfully [1].

Traditional breadth-first crawling, which is called as TBFC in the paper, is popularly used in all kinds of cases.

To visit a post in a board, human usually starts from the homepage, and then it enters into a board, and then to find the post. This process says that the forum is organized in a structurally manner. Thus they have found that there exist mainly 3 kinds of pages in one Web forum site i.e. homepage and board page and post page.

The Precise of their BFC can reach up to 90%. It is much higher than that of the TBFC.[1]

The Recall of their BFC is also more than the TBFC in most of the sites. [1]

Experiments have shown BFC is an economical and efficient method. BFC has been used in a real world project, and approx.12000 Web forum sites have been crawled successfully.

B. Focus: Learning To Crawl Web Forums [2]

In [2] Jingtian Jiang, et.al presented FoCUS, a new approach towards web crawler. They have presented a crawler named FoCUS (Forum Crawler under Supervision).It is a supervised web-scale forum crawler. The aim of FoCUS is to crawl only relevant forum content from the web with minimal overhead. However, forums have different layouts or styles and are powered by different forum software packages. Forum threads contain information content which is the target of crawlers. They always have similar implicit navigation paths which are connected by specific URL types to lead users from entry pages to thread pages.

They call pages between the entry page and thread page which are on a breadth-first navigation path the *index page*. They have these implicit paths as the following navigation path (EIT path):

entry page -> index page -> thread page [2]

In their experiment [2] over 160 forum sites (10 pages each of index, thread, and other page) each powered by a different forum software package, our classifiers achieved 96% recall and 97%precision for index page and 97% recall and 98% precision for thread page with different amount of training data. It was unable to determine the type of the candidate group. So, we need to check the destination page type of the URLs in the candidate group. A voting method is adopted to determine the type since classification results on individual page might be full of errors. By utilizing aggregated classification results, FoCUS does not need strong page classifiers.

They have 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 different methods to learn ITF regexes explicitly. 160 forum sites have been experimented and the experimental results on each site powered by a different forum software package confirm that FoCUS crawler could effectively learn knowledge of EIT path and ITF regexes from minimum 5 annotated forums.

C. Learning URL Patterns for Webpage De-Duplication [3]

In [3] H.S. Koppula et.al have presented that the existence of duplicate documents in the World Wide Web adversely affects indexing ,relevance and crawling , which are the main building blocks of web search. Thus, In this paper, they have presented a set of techniques to mine rules from URLs and then utilize these rules for de-duplication using only the URL strings without fetching the content explicitly.

Their technique is mainly composed of mining the crawl logs and further utilizes these clusters of similar pages to extract transformation rules, which are used to normalize URLs which belong to each cluster. So, to preserve each mined rule for de-duplication is not efficient due to the huge number of such rules. So, they present a machine learning technique which generalizes the set of rules, which helps in reducing the resource footprint to be usable at web-scale. The rule extraction techniques are the best against the web-site specific URL conventions.

They have contributed the following things in the paper:

- They propose a technique for extracting host specific delimiters and tokens from URLs. Extension of the pairwise Rule generation is required to perform source and target URL selection. They have also introduced a machine learning based generalization technique for better precision of Rules. On the whole, these techniques form an effective solution to the de-duplication problem.[3]
- They have extended the representation of URL and the Rule presented in it. The se extensions result in better utilization of the information encoded in the URLs to generate precise Rules with more coverage.
- Since scale is a necessary dimension on the Web, they present a technique called Map- Reduce adaptation for the proposed techniques.
- They have demonstrated via experimental comparison that the proposed techniques produce twice more reduction in duplicates with half the number of Rules they are compared to. Concluding, large scale experimental evaluation upon a 3-

billion URL corpus, they show that the techniques used are scalable and robust.

D. Intelligent Crawler Web Forums Based On Improved Regular Expressions [4]

In this paper, named "Intelligent Crawler for Web forums based on improved regular expressions", the authors have presented a special crawler for Internet forums.

The main data collected from forum contents are the posts, and all relevant information that goes with them. A post always contains post body i.e name of the author of the post, text and the date when the post was created. It also contains additional information such as the link to the author's profile, gender, post number, date of joining the forum, anchor, frequency of activities, author's picture etc are also collected and can be of important use in later analysis. Thus, In most cases, users want to get the structured information or data directly instead of web pages [4].

They have proposed and implemented a specialized web-scale forum crawler. They have reduced the forum crawling problem to a URL type recognition problem and showed how to take implicit navigation paths of forums, and have designed methods to use this paths effectively. The authors have also proposed a model for storing the collected information that is suitable for further analysis. The test is carried out on 1,200 forum sites which are powered by different forum software packages and languages which confirm that web-scale forum crawler can effectively collect forum elements in much more effective way than generic crawler. [4]

E. iRobot: An Intelligent Crawler For Web Forums [7]

In [7] R. Cai et.al, presented iRobot crawler for web crawling. The authors have focused upon deep web crawling and near duplicate detection. Its main aim is to understand Content and structure of a forum site.

They found the following observation:

- The repetitive regions in a forum page can greatly characterize the content layout of that page.
- The location of a link on a page is important.

The main goal of iRobot is to automatically rebuild the graphical architecture representation,

i.e., the sitemap of the target Web forum and then select an optimal traversal path which only traverses informative pages and skip invalid and duplicate ones. They first randomly crawl a few pages from the source site. Then, based on the first observation, all the repetitive regions are discovered from all of these sampled pages, and are further employed as features to group these pages into clusters according to their content layouts. Thus, Each cluster can be considered as a vertex in the sitemap. [7] So, based on the next observation, each arc in the sitemap is characterized by both the URL pattern and the location of related links, to provide more effective and accurate discrimination between links with different functions.

iRobot crawler has some advantages in comparison with a generic crawler: 1.Effectiveness: It can intelligently skip most invalid and duplicate pages, keeping informative and unique ones. 2. iRobot can significantly reduce the ratios of invalidation and duplication 3.Efficiency: Relationship reserved Archiving

Thus, iRobot can significantly reduce duplicate and invalid pages, without losing the valuable ones. It only needs to pre-sample maximum of 500 pages for discovering necessary knowledge. iRobot can keep around 95% page relations in crawling, which is very useful for data mining tasks and further indexing [7]

III. PROPOSED WORK

The main disadvantage of our existing system is crawling of URLs which may lead users to same thread or page inside a forum. Thus, it increases the crawling time of the crawler. Page Flipping and Clone mining are very much required to make the crawler more efficient and increase the coverage of the crawler. Page Flipping URL's destination pages have similar layout as source pages. Clone mining is done to identify the URLs which have same structure but different data. Index-Thread-Flipping URLs are analyzed and patterns are extracted. These Regular expressions are used to eliminate Page Flipping URLs. This crawling technique thus will help to improve the precision and recall value of a crawler.

No loss of data can occur via this technique as only the URLs are processed. No tampering is done on the content of the pages. Forum sites have different software packages and resources, so some of the forum sites will be used for experimenting this approach.

IV. CONCLUSION

The present survey illustrates various methods and techniques used for making an efficient web crawler. Web Forums have made a tremendous place in the World Wide Web. Web Forums have different layouts, different structure and different types of pages. Any kind of web crawler needs to improve its efficiency, coverage and precision by understanding the structure of a web forum. Almost all the crawlers come across the issue of page flipping or duplicate web pages. Thus, by understanding the issues faced by crawler and the structure of a forum, one can achieve a highly efficient and faster web crawler. It must be able to eliminate irrelevant pages or any kind of unnecessary information to utilize the time in efficient way. Thus, Techniques used for Parsing and extraction of URLs play an important role in improving the crawler efficiency upon a forum site.

REFERENCES

[1] Yan Guo, Kui Li, Kai Zhang, Gang Zhang, "Board Forum Crawling: A Web Crawling Method for Web Forum", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[2] Jingtian Jiang, Nenghai Yu, Chin-Yew Lin, "FoCUS: Learning to Crawl Web Forums",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:6 YEAR 2013

[3] H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg, and A. Sasturkar, "Learning URL Patterns for Webpage DeDuplication," Proc. Third ACM Conf. Web Search and Data Mining, pp. 381-390, 2010.

[4] Miloš Pavković, Prof. Jelica Protić, "Intelligent Crawler for Web Forums based on Improved Regular Expressions",21st Telecommunications forum TELFOR 2013/IEEE/YEAR 2013
[5] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, "Study Of Web Crawler and Its Different Types", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05
[6] A. Bergholz, B. Chidlovskii,"Crawling for domain-specific Hidden Web resources", Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE'03). pp.125-133,IEEE Press, 2003
[7] R.Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An Intelligent Crawler for Web Forums", In Proc. of 17thWWW, pages 447-456, 2008
[8] http://www.slideshare.net/iamthevictory/web-crawler
[9] http://hunyadi.info.hu/levente/en/publications/6-crawlernet
[10] http://mias.uiuc.edu/files/tutorials/mercator.pdf
[11] http://en.wikipedia.org/wiki/CRAWLER.

