Volume 5, Issue 6, June 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Intelligent Crawling Web Forum

Trupti D. Narkhede, Prashant M. Yawalkar
Department of Computer Engineering
MET-BKC Savitribai Phule Pune University
Nasik, Maharashtra, India

Abstract*— Data mining on a very large data set is a complex task. ICWF (Intelligent Crawling Web Forums) is web-scale forum crawler. The objective of ICWF is to simply crawl related forum content from the web with minimum overhead. A Crawler traverses the World Wide Web in a systematic manner with intention of gathering data or knowledge. Web forums are used by large number of users to post and share their comments with other users of various websites. The forums consist of many lists of topics on their boards with a large list of threads in each board. The users can create many threads and share their views in posts as well. Forum threads have information contented that is the aim of the forum crawler. Although forums include various layouts or styles and they are powered by various software packages forum, they have similar complete navigation path linked by specific types URL to lead users from entry pages to thread pages. Based on this observation, there is a need to decrease the web crawling forum problem to a URL type recognition problem and how to show learn effective and accurate patterns of regular expression to complete navigation paths. These paths are obtained from an automatically generated training set using aggregated results from weak type page classifiers. Page type classifiers can be trained from as few as five forums annotated and applied to a big set of hidden forums. It is expected that effectiveness coverage on a big set of forums test can be achieved to a significant extent.

Keywords*— EIT path, ITF regex, forum crawling, URL type, page type, page classification, URL pattern learning.

I. INTRODUCTION

Internet forum or web forum or message board is an online discussion site where users hold conversations in the form of messages which are posted on the forum known as discussions or post [1]. 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 Internet forums [3] are important platforms where user can apply for and exchange information with others. The information in forums, researchers are more and more knowledge from them [2], [3] extracted structured data from forums. It uses tree like traversal strategy, which takes only one path from entry page to thread page. A forum can contain a number of sub-forums [2]. It doesn't maintain record for already crawled forum site since it is time consuming process. It can be used to multiple links or pages. A forum has many copy links that point to the common pages but different URL forms. There are two non crawler characteristics of forums 1) repeated links and 2) page-flipping links. A forum typically has many copy links that point to the common page but with different URLs generated. The example as shortcut links pointing to the posts or URLs for user experience the normal of this "view by date" or "view by title". Generic crawlers use breadth first traversal technique to traverse the contents of the forum in breadth first way. But this is not so effective due to various limitations such as the availability of duplicate links, a page flipping links and uninformative pages [5]. Generic forum crawlers can be used to crawl forum contents to identify the required information. The duplicate links are those links available in various pages of the forum that leads to the same page.

A web crawler is an internet that systematically browses World Wide Web purpose of web indexing. The forum can be divided into different categories for the applicable discussion. Under the categories are Sub-forum and these having more forums. Web search engines use to web crawling to update software their web indexes of other site. The download pages search easily for user more quickly.

In addition above two challenges, there is entry URL discovery. The entry URL of the forum point to home pages. A generic crawler that blindly follows the connection duplicate pages [5].

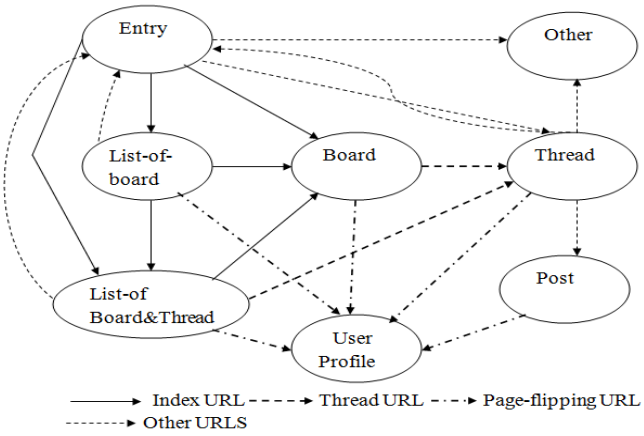

Fig 1: Link relation in a forum.

Fig 1.Illustrates a page and relation structure in a forum. For Example, the client can navigate from the entry page to thread page.

There are different following path for the forum page.

1. entry→board→thread
2. entry →list-of-board→ board→ thread
3. entry→list-of-board & thread→ thread
4. entry→list-of-board&thread→board→thread
5. entry→ list-of-board → list-of-board & thread→thread
6. entry → list-of-board → list-of-board & thread →board →thread

The pages connecting the entry page and thread page which are on the index page represent implicit path which is called as the EIT path. These forums are available in many different styles and formats. The various pages in a forum can be classified into three important pages such as entry page, index page and thread page [4], [9].

EIT path is also known as the Entry-index-thread page.

Entry page →index page→ thread page

It Connect between two index pages are which referred as index URLs. Connections between an index page and a thread page are referred as thread URLs while Connections between various pages of a board and various pages of a thread are referred as page flipping URLs. A crawler starting from the entry URL, thread URL, index URL and page flipping URL to traverse EIT paths that all thread pages. They are generated different URLs i.e. Index URL, thread URL, User URL's. IRobot: An Intelligent Crawler for web forum [2]. Web forum is very important and popular world for the open discussions. Every day, there are in numerable new posts created by millions of Internet client to talk about any possible topics and issues. User can download discussion information efficiently and effectively. The structures of all forum sites are nearly same to one crawl method use in other forum site. In the end crawl fetch the data from forum and store it in the database.

II. RELATED WORK

Vidal et al. [5] proposed a method based on the structured-driven approach to generate web crawlers. It takes a page from any website as input and generates a structure-driven web crawler based on the various navigations available between the pages. This can be implemented only for learning regular expression of URL crawler from entry to the target pages. Target pages are found to the DOM tree pages with selected the target page.

Thread pages where found through pre sampled pages. It is used only for specific forum site since it is not applied for large forum crawling. Proposed system learns URL patterns from multiple sites so it can be used for large forum crawling according to Forum Matrix [7]. There are hundreds of different forum software packages which are used on the Internet. In wide-ranging web crawling, navigation patterns show the way to target pages. IRobot also adopted comparable thought but functional to page sampling and clustering method to discover target pages [6].

A web forum service receives to the user request, it generate the response page based on some predefined template. Forum sites generally have the following characteristics, duplicate pages with different Uniform Recourse Locators will be generated by the service for the different request most forum sites. In general, content of a forum is stored in a database. Due to these reasons, forum sites generally have the following common characteristics. First, duplicate pages with different Uniform Resource Locators will be generated by the service for different requests such as "view by date" or "view by title". Second, a long post divided into multiple pages usually results in a very deep navigation. Sometimes a user has to do tens of navigations if he/she wants to read the whole thread, and so does a crawler. Finally, due to privacy issue, some content such as user profiles is only available for registered users. A recent and more including all work on forum crawling is iRobot, an intelligent web forum crawler that can identify the navigation paths based on the learned intelligence from the structure and contents of the forum contents. Samples pages are downloaded from the forum website and provided as layout to the crawler [8]. IRobot idea to automatically learn a forum crawler with smallest amount human being mediation by sampling pages, clustering them, and to search the traversal path by referred to as

spanning tree algorithm. A crawler starting from the entry URL only needs to follow index URL, thread problem. It is an intelligent crawler for Web Forums. The fundamental step in many applications web is forum crawling problem, such as web data mining and search engine. Web forum crawling is not a trivial issue due to the in-depth link structure, the amount of duplicate pages and many invalid pages caused by login failure problems. Apart from these various other techniques such as the near-duplicate detection are also used for removing the duplicates when forum crawling. Content based duplicate detection methods are not efficient since they need sample data to be downloaded before implementing the crawling process. The URL based duplicate detection is also not efficient because it tries to extract and identify the various URLs by using the forum logs or results obtained from previous crawl. The proposed method to identifies the various URL patterns and by using a URL based de-duplication method, the duplicates can be easily avoided.

Learning patterns of thread URLs, index URLs and page-flipping URLs and adopting a easy URL string de-duplication technique, Focus can avoid duplicates without duplicate detection.

Existing techniques like content based duplicate detection and URL based duplicate detection are used. Content based duplicate detection is less effective since it prone to less bandwidth. URL based duplicate detection is not efficient for URL with same text. By using the URL patterns in proposed system duplicate pages can be removed. It is efficient as well as more robust.

The method for use learning regular expression patterns of URLs.

There are different classified forum pages.

- Page Type
- URL Type
- EIT Path
- ITF Regex

- Page Type: The classified forum pages into page types.
  - Entry Page: The entry page is the homepage of the forum that contains various lists of topics. For example See Fig.2a.
  - Index Page: A page of a board in a forum, which normally contains the table like structure; each row in it contains information of a board or a thread. For examples See Fig. 2b. In Fig.1, list-of-Board page, list-of-board and thread page, and board page are everyone index pages.
  - Thread Page: The thread page contains list of threads posted by the users of the forums. For example see Fig.2c.
- URL Type: There are four types of URL.
  - Index URL: The index URLs are used to navigate to index pages and they are available in the homepage or other index pages. It is anchor text display the title of its destination board. For example see Figs. 2a and 2b.
  - Thread URL: The thread URLs can be used to navigate to the thread from the index pages. It is anchor text is the title of its destination thread. Figs. 2b and 2c show an example.
  - Page-flipping URL: The page flipping URLs lead to another page or thread within the same page. Correctly page-flipping URLs to make able a crawler to download all thread a large board or all posts of thread. For example see Figs. 2b, and 2c.

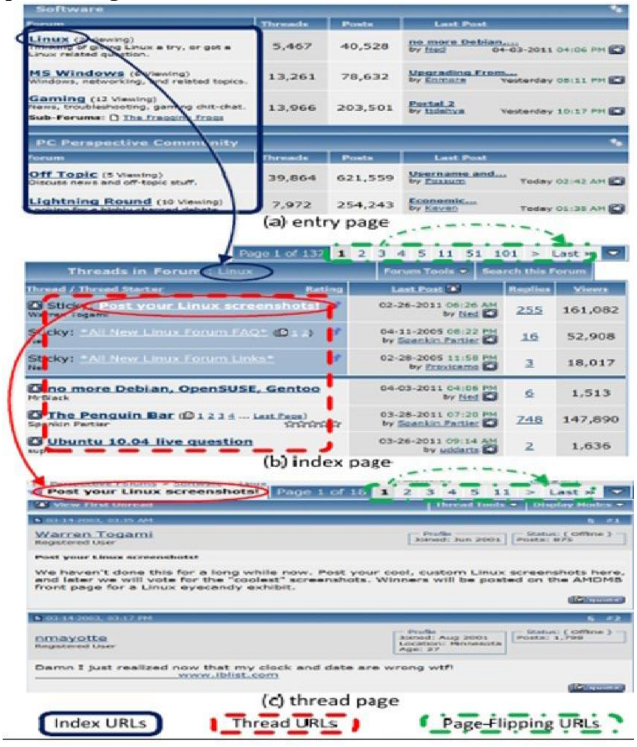

Fig 2. EIT paths : entry →board → thread.

- EIT Path: An entry-index-thread path is a navigation path from an entry page through the series of index pages to thread pages .
- ITF Regex: An index-thread-page-flipping regex is a regular expression. It can be utilized to recognize index, thread, or page-flipping URLs. ITF regex is the thing that ICWF mean to learn and applies directly in the online crawling.

The learned ITF regexes are site specific, and there are four ITF regexes in a site: one for recognizing index URLs, one for thread URLs, one for index page-flipping URLs, and one for thread page-flipping URL

III. PROPOSED SYSTEM

In this fig shows the overall architecture of ICWF. It consists of two main parts: the learning part and the online crawling part. The first learning part learns ITF regular expressions of a given forum from automatically constructed URL training examples.

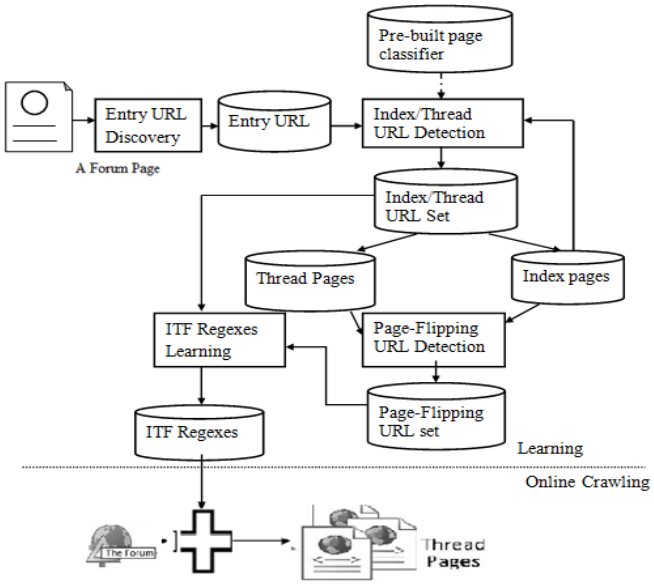

Fig 3. The overall architecture of ICWF.

After the URLs and the regular expression patterns are recognized the second online crawling part i.e. the crawling phase is executed on the forum website. First the given any forum page is selected entry URL's; ICWF first thinks its entry URL using the URL Entry Discovery module. In the next step the index and the thread URLs are detected by using the page classifier module. Then, it uses the Index/Thread URL Detection module to detect index and thread URLs on the entry page; the detected index URLs and thread URLs are saved to the URL training sets. All the thread URLs are identified here and the index pages are fed to this module again to detect more thread and index URLs within the pages. After the Page-Flipping URL Detection module tries to locate page flipping URLs from both index pages and thread pages and saves them to the training sets.

Finally, using these regular expressions the proposed method crawl the webpage forum starting from the entry URL and travelling to all the URLs satisfying the regular expressions.

IV. ALGORITHMS STRATEGY

A. The Index Url And Thread Url Detection

Step1 - Enter data
Step2 - Call View Topic by category to start any thread.
Step3. Then select thread.
Step4. Check whether any topic is available in thread.
Step5. If available then call List (Topics) in selected Thread.
Step6. else List count is 0.
Step7. then crawl for topic which is selected to be search one by one.
Step8. end.

B. URL Landing

Ct: Crawling thread.
At: Available thread.
Rt: Recent thread
Ot: Other running thread
Fu: Forum Users.

Basic operation algo for Landing URL

Step1 - start
Step2 - get value for Ct
Step3 - check Ct status with Available working thread At
Step4 - compare (Ct == At)
Step5 - If is it true then crawl on the Rt page instead of home index page.
Step6 - Else SHOW home index URL.
Step7 - check the post whether it is right.
Step8 - check Ot simultaneously which are updated by Fu.
Step9- end

C. The page-flipping URL detection

Step1 - Start
Step2- Select the view all topics then it shows.
Step3- if it is selected available topic then it is done.
Step4- else topic is not available then go to next option to get difference of five.
Step5- crawl page up to topic id.
Step6- end.

V. EXPERIMENTAL RESULTS

Different forum software packages are selected from forum Matrix [7]. We have selected forums as our training set. The obtained results were satisfying in the advanced crawling with respect of relevancy, effectiveness and time of crawling the forum package. There are three technique used in crawling for 1)Index/Thread URL Detection 2)page flipping URL Detection 3) URL Landing 4) Intelligent crawling

TABLE I

Types	Existing	Proposed
URL Landing	4	1
Thread URL crawling	3	2
Page Crawling	4	2
Intelligent Crawling	11	5

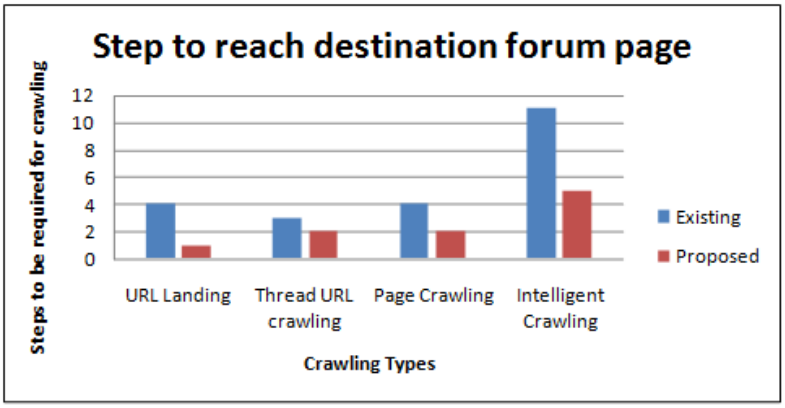

Fig 4. The step to reach destination forum page

The calculated performance values are plotted in bar graphs and a comparison is made with existing web crawler technique. It proves that the proposed method has a higher performance and efficiency than the existing methods. The comparison graph is shown in Fig 4 In URL Landing , proposed system is avoid the no of pages in less time than existing system .Similar to Index URL ,Thread URL and Page flipping URL.

VI. CONCLUSIONS

Web mining is the application of data mining technique to discover patterns from the web. Improving the structure of the web mining process is an important activity relevant to the user's or group of users' requirement. The web forum page crawling system is a supervised forum crawler. It reduced the forum crawling problem to a URL type recognition problem and showed how to leverage implicit navigation paths of forums, such as EIT path, and designed methods to learn ITF regexes explicitly. It can effectively learn knowledge of EIT path from as few as five annotated forums. It proved that it can effectively apply learned forum crawling knowledge on numerous unseen forums to automatically detect the various index, thread, and page-flipping URLs training sets and learn ITF regexes from the training sets. Further, it is expected that use of URL Landing algorithm will reduce the navigation time to a considerable extent.

ACKNOWLEDGMENT

The author wish to thank MET's Institute of Engineering Bhujbal Knowledge City Nasik, HOD of computer department, guide and parents for supporting and motivating for this work because without their blessing this was not possible.

REFERENCES

[1] Jingtian Jiang, Xinying Song, Nenghai Yu, and Chin-Yew Lin, "*FoCUS Learning to Crawl Web Forums*," Proc IEEE Trans. Knowledge Data Eng., vol. 25,no. 6 ,June 2013.

[2] Y. Zhai and B. Liu, "*Structured Data Extraction from the Web based on Partial Tree Alignment*," Proc .IEEE Trans. Knowledge Data Eng., vol.18, no.12, pp. 1614- 1628, Dec. 2006.

[3] J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W.Ma, "*Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums*", Proc.18th Intl Conf. World Wide Web, pp. 181-190, 2009.

[4] Message Boards Statistics, http://www.big-boards.com/ statistics/, 2012.

[5] M.L.A. Vidal, A.S. Silva, E.S. Moura, and J.M.B. Cavalcanti, , "Structure *Driven Crawler Generation by Example* ", Proc.29th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp.292-299, 2006.

[6] K. Li, X.Q. Cheng, Y. Guo, and K. Zhang, "Crawling Dynamic Web Pages in WWW Forums", Proc Computer Eng., vol. 33, no. 6,pp. 80-82, 2007.

[7] Forum Matrix, http://www.forummatrix.org/index.php, 2012.

[8] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang,, " iRobot: An Intelligent Crawler for Web Forums", Proc. 17th Intl Conf. World Wide Web, pp. 447-456,2008.

