Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 2, February 2014, pg.374 – 381

RESEARCH ARTICLE

A Supervised Method for Multi-keyword Web Crawling on Web Forums

A.Gowtham[1], Dr.K.Deepa[2]

[1]Department of CSE (PG), Sri Ramakrishna Engineering College, India

E-mail: gowslm@rediffmail.com

[2]Department of IT, Sri Ramakrishna Engineering College, India

E-mail: deepkarun@rediffmail.com

Abstract – 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. In this paper a supervised web forum multi-keyword crawler is proposed to crawl relevant contents from the forum pages by reducing the delay. All the forums in the web have navigation paths that lead to the forum threads and these paths are connected by specific types of URLs. Thus the proposed method needs to recognize the various URLs by using the regular expression patterns within the forum. Accurate page classifies trained by using other forums can be used to classify the regular expression patterns and detect the URLs. The obtained results show that the proposed method is more reliable and accurate comparing to other existing methods.

Keywords - Web crawler, page classification, forum crawler, URL based learning.

I. INTRODUCTION

The forums available in the internet are so important for various kinds of users where they can interact with many other users. The users can post and share information in the forums. Many such forums are available for various topics such as the web forum for gaming, commercial products, television programs, technology, companies, books etc. The major goal of the forums is to share information with other users who are interested in the same topics. The data available in forums are very large and valuable, so mining information from forums will be useful to the user and the organization. For example, a gaming company managing a forum can extract important information about the posts from various users that will be helpful in improving the features of the game.

In general to extract knowledge from a forum the contents of the forum should be downloaded and stored in a separate place. But this will consume a lot of storage capacity

and time. Generic forum crawlers can be used to crawl forum contents to identify the required information. 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, uninformative pages and page flipping links. The duplicate links are those links available in various pages of the forum that leads to the same page. Uninformative pages such as the user profile, long page and contacts does not have any needed information. The page flipping links are the links available in a forum page that is divided into many parts. These links lead to another part of the same page.

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. The entry page is the homepage of the forum that contains various lists of topics. The index page is the intermediate pages between the homepage and the thread pages. The thread page contains list of threads posted by the users of the forums. The URLs used to navigate between these pages can also be classified into three main types as index URL, thread URL and page flipping URL. The index URLs are used to navigate to index pages and they are available in the homepage or other index pages. The thread URLs can be used to navigate to the thread from the index pages. The page flipping URLs lead to another page or thread within the same page.

In this paper, a supervised method is proposed to crawl web pages by using relevant forum contents posted by the users. The proposed method reduces the delay in extracting the information from forum contents. The links available in the forums lead the users from the homepage to the thread pages. These paths to the forum threads are called as navigation paths and they can be formed by identifying the various paths in the forum that leads to the thread pages. A regular expression can be used to recognize the three types of URLs. The proposed method learns the regular expressions from the forums and uses them to recognize the different URLs available to navigate the paths to the threads.

The rest of the paper is as follows. Section 2 shows the various related works. The Section 3 explains the proposed method and techniques used. The results are displayed and discussed in Section 4 and the conclusion & future works are described in final Section 5.

II. RELATED WORKS

Vidal et al. [1] 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 that specific website from where the input sample page is selected.

Guo et al. [2] proposed a board forum crawling method that starts from the home page and crawls each board available in the forum till each post in the site is crawled. This method works by extracting the behavior of each user who visits the forum. This method can be used to find most useful in-depth information in the forum. But no technique was used to discover the URLs.

Cai et al. [3] proposed the 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. But the proposed method is much more reliable than the iRobot crawler techniques.

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 in this paper identifies the various URL patterns and by using a URL based de-duplication method, the duplicates can be easily avoided.

III. PROPOSED METHOD

Before going for the implementation of the proposed method certain sample forums were analyzed and certain observations were made. Observations included the various navigation paths, URLs and layouts of the forums. The forum pages can be classified based on the various layouts by using the proposed method and by using the layouts the various URLs can also be detected.

The features of the proposed method are as follows,

- It automatically learns about the various URLs available in the web forum.
- The proposed crawler is compared with existing crawlers and proved to be more effective.
- The reuse and accuracy is high compared to existing methods.
- The proposed method can also be applied to blogs and other community websites.

The basic architecture of the proposed method is shown in Fig 1. The figure shows the working flow of the proposed method. It consists of two phases as the learning phase and the crawling phase. The first phase is the learning phase where the crawler learns and identifies the various navigation paths and URL links by using the page classifier that is pre-build by using other sample forums. After the URLs and the regular expression patterns are identified the second phase i.e. the crawling phase is executed on the forum website.

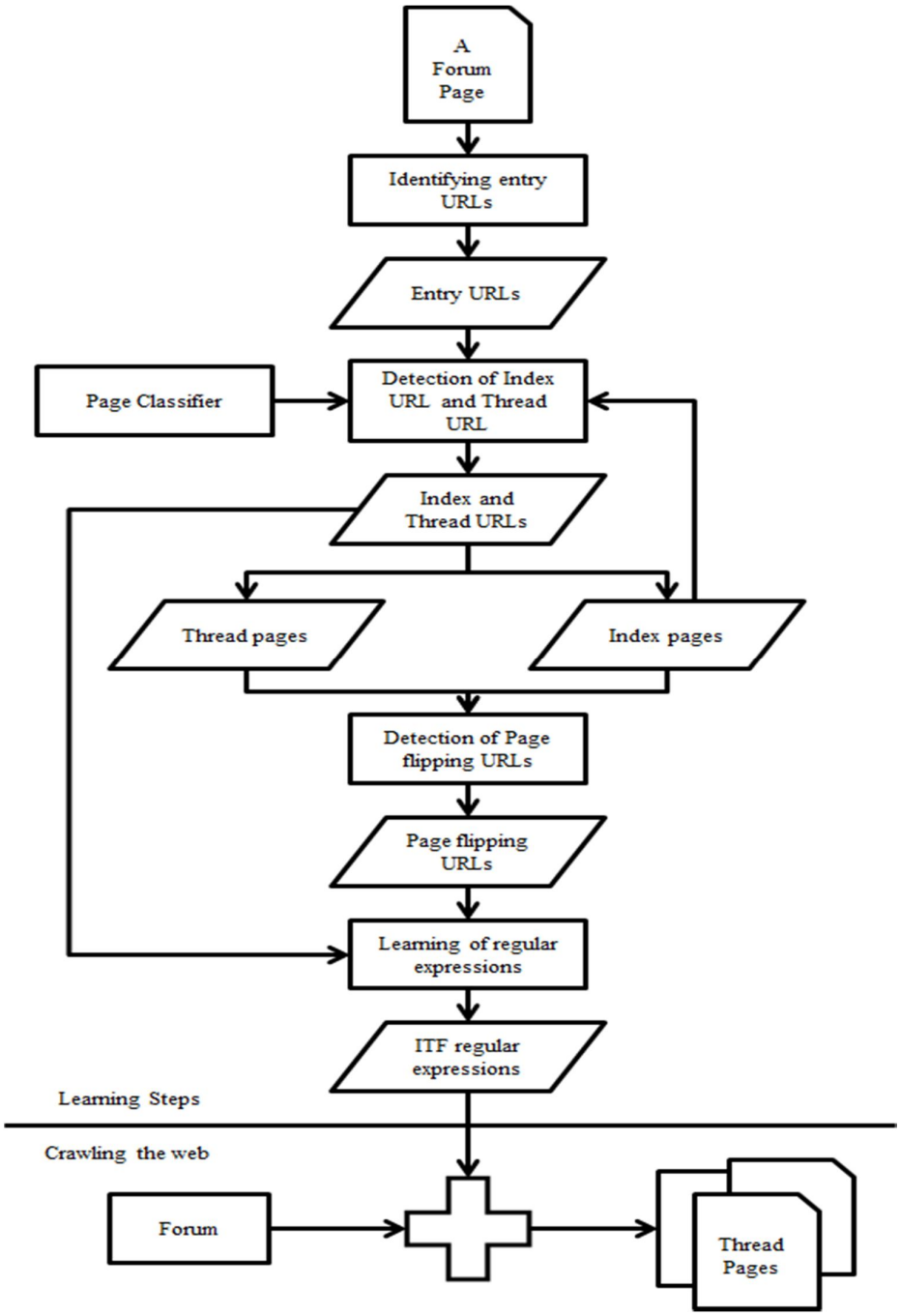

Fig 1. Architecture of the proposed crawler technique with each step

First the given forum page is selected and the entry URLs URL is identified in the next step. In the next step the index and the thread URLs are detected by using the page classifier module. The classifier Support Vector Machine (SVM) is being used here. 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. This continues till all the index and thread URLs are identified. The identified URLs are provided to the URL training set. In the next step the page flipping URLs need to be identified and also added to the URL training set.

By using the training set that contains the entire URL (index URL, thread URL and page flipping URL) some regular expressions are learned. And 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.

A. Index and thread URL detection

The layout and characteristics of the entry URL and thread URL are different and they have the similar properties. The only way to differentiate between the thread and index URL is the anchor text and the destination page. The thread URL has some long forum posts and the anchor text is short. But in index URL the anchor texts are long and have narrow records.

The proposed method uses various characteristics and layouts of the page to construct a page classifier that can be used to classify the various pages. A DOM tree is constructed for the page layout. Based on the characteristics and the information extracted from the DOM tree the URL can be identified. The various characteristics of the URL include anchor length, text length, link, timestamp, tree similarity, text similarity, profile links, groups, etc.

The proposed method is implemented in a total of 7 forum websites containing more than 150 pages. The DOM tree is constructed for the identified index and thread pages by using the identified characteristics. A sample DOM tree is shown in Fig. 2 that has the various parts of a URL.

B. Page flipping URL detection

The next step is to detect the page flipping URL in the forum pages. A page flipping URL leads to an index page or thread pages within the same page and so

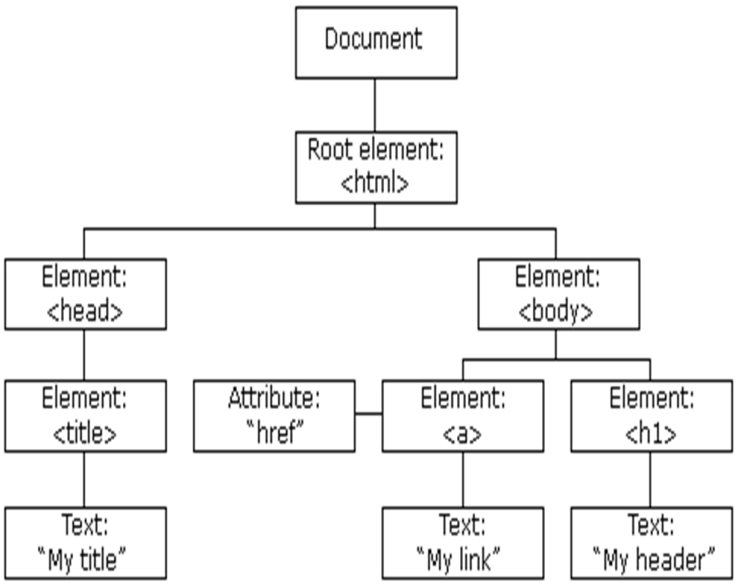

Fig 2 Sample DOM tree

they have different characteristics compared to that of index URL and thread URL. The existing method for detecting page flipping URL works only if the page has more than one page flipping URLs. In the proposed method certain characteristics of the page flipping URL were identified for detecting them. By constructing and aligning the DOM tree, page flipping URL that is available can be identified.

The proposed method provides on overall more than 95 percent of precision and recall in identifying the thread URL, index URL and page flipping URL. This is applied on all the 7 forums containing more than 150 pages and the identified URLs are saved as the initial training set for the crawling algorithm.

C. Learning ITF regular expressions

The proposed method used supervised way of learning to construct the training set and to learn the regular expressions pattern. The ITF regular expressions can be learned by using the identified thread, index and page flipping URL training sets.

The existing methods uses string generalization techniques that can be applied to the URL training set to generate the regular expressions but this is not to efficient. Consider there are 4 types of URL in a forum page then a regular expression matching the entire 4 URL should be used. In the proposed method general patterns such as * and + are used for the regular expressions. Consider the following URL from the forum site,www.afterdawn.com.

http://www.forums.afterdawn.com/forum_view.cfm/163/
http://www.forums.afterdawn.com/forum_view.cfm/216/
http://forums.afterdawn.com/t.cfm/f-163/unix_network_discovery_issues_limited_by_tools_lack_of_data-970991/
http://forums.afterdawn.com/t.cfm/f-216/the_official_pc_building_thread-4th_edition-867265/

The regular expression can be formed in the first 2 URL as given below;

http://www.forums.afterdawn.com/forum_view.cfm/\d+/

Where the expression \d+ can be used to represent the forum number. A general regular expression for all the 4 URL can be given as follows;

http://www.forums.afterdawn.com/f-\d+/\e+

Where the expression \d+ represents the forum page number and the expression \e+ represent the string for the thread page.

IV. ENTRY URL AND ONLINE CRAWLING

A. Online Crawling

This crawling process on the above URLs are being done by using Breadth-First search strategy and the ITF are being queued for as a group and others are dequeued as Entry URLs. The general procedure of the BFS strategy is used here and more than one keyword are provided as input for searching. That is the crawling process is carried out by using multiple keywords.

B. Entry URL discovery

The URLs other than in the queue are checked for its entry pages. The entry URL can be available in other pages also that can lead to the home page of the forum. A forum can have one entry URL or more depending on the layout of the forum. In the proposed method certain rules were considered for detecting the entry URL using multiple keywords. The entry URL differs in every forum since they have different home pages. Many keywords such as forum, community, board, etc. are considered. If any of the keywords are identified then the path

from the URL till the keyword is considered as the entry URL else if no keyword is found then the host URL is taken as the entry URL.

V. PERFORMANCE ANALYSIS

The performance of the proposed web crawler is measured by calculating the precision, recall and accuracy values using the formulas given in Eq. (1), Eq. (2) and Eq. (3).

$$Precision = \frac{TP}{TP+TN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad (3)$$

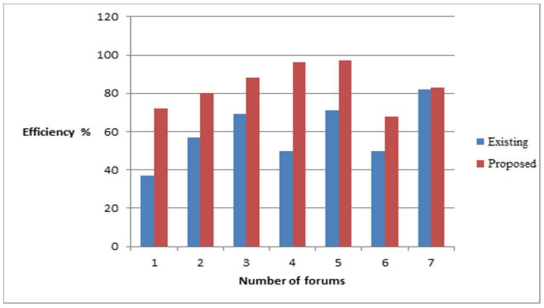

Fig 4. Efficiency comparison with existing method iRobot

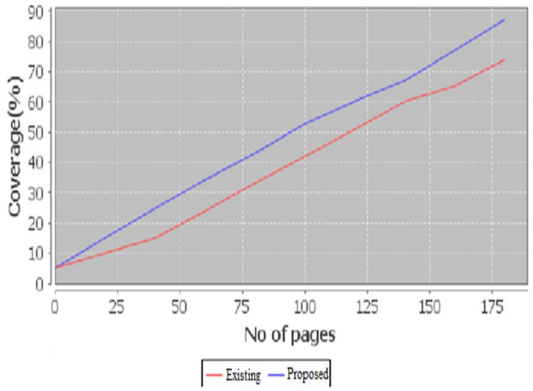

Fig 5. Comparison of coverage in URL detection and online crawling

The calculated performance values are plotted in 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 and Fig 5.

VI. CONCLUSION

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. The learned regular expressions can be applied directly in the proposed online crawling process. Since the crawling process is carried out by using the testing and training sets from the forum package, it can be managed easily and can be applied to many forum sites. Moreover, it can start from any page within the forum site, where as the existing methods needs an entry URL to initiate the process.

REFERENCES

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

[2] 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.

[3] 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.

[4] A. Dasgupta, R. Kumar, and A.Sasturkar, "De-Duping URLs via Rewrite Rules," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.

[5] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song, "Finding Question-Answer Pairs from Online Forums," Proc. 31st Ann. Int'l ACM
SIGIR Conf. Research and Development in Information Retrieval,pp. 467-474, 2008.

[6] N. Glance, M. Hurst, K. Nigam , M. Siegler , R. Stockton, and T.Tomokiyo, "Deriving Marketing Intelligence from Online Discussion,"Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 419-428, 2005.

[7] 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.

