IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 01-05
www.iosrjournals.org

FoCUS: System to Learn Crawl Web Forum

S.S.Pophale[1], Andhale Asmita B[2], Auti Pallavi S[3], Gaikwad Akshata S[4] &Gavhane Pooja S[5].
(IT Dept)

Abstract: *FoCUS is Forum Crawler Under Supervision. The main aim of the system is to give only relevant content from forum and minimize the overhead. In that user can give request to forum and user can exchange information by using forum. Forum have implicit navigation paths using specific URL type from entry page to thread page. From observation system reduce forum crawling problem by using URL type recognition problem. System also use regular expression pattern. FoCUS design for learning ITF regexe explicitly. FoCUS learns EIT path and ITF regexe from forum. Forum automatically collect index URL, Thread URL and page flipping URL into training set.*

Keywords*– Forum crawler, ITF regexe, Page Type, URL Type.*

I. Introduction

A web crawler is one type software agent. Web crawler starts with the list of URL to visit page. As the crawler visit this URLs, it identifies all the hyper links in the page and adds them to the list of URLs is called crawler front tier. Forums are platforms where user can request for information .For example, Trip Advisor Travel Board where user can share travel tip. Forum consists of tree like directory structure forum divided into multiple categories for relevant discussions. Under the categories consist of sub forums and again this sub forum can further have more sub forum. The topics(commonly called thread) come under the lowest level of sub forums and these are the places under which members can start there discussion or posts. Logically forums are organize into a finite set of generic topics(usually with or main topic) driven and updated by a group known as a member and govern by a group as moderators. All message boards will use one of three possible display formats.

Each of tree basic message board display formats; non -threaded /semi-threaded /fully threaded, has its own advantages and disadvantages. A threaded is the collection of post usually displayed from oldest to latest, all though this s typically configurable: options for newest to oldest and for a threaded can be available. Thread define by title additional description that may some once the intended discussion, and opening or original post which opens whatever dialogue or makes whatever announcement the poster wished . a thread can contend any number of post from same member , even if thread are one after other. Sometimes called a bulletin board or message board a web forum one line center for ongoing, in depth discontinue of specific topic and issues. Forum is not a chat room. That is user not participate in a real time discussion however user can ask and respond to question , explain service technique and strategy. Zhai et al[1] extract structured data from forum generic crawler[5] are crawler which use a breadth first traversal strategy are ineffective and inefficient for crawling. Generic crawler processes each and every page and ignores relation between pages. Glance et al. [9] tried to give minimum business intelligence. Gao et al[8] describe forum in the form of answer and pair focus is forum crawler under supervision the main aim of focus is to trawl related information that is user post from minimum overhead of forum .

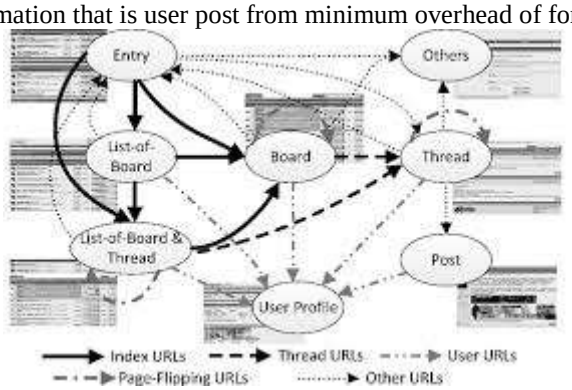

Fig 1 : Link between Forums

Forum consist of many different page layouts by verity of software packages. In forum user can travel from entry page to thread page using following path:

1. Entry -> board ->thread
2. Entry -> list-of- board-> board ->thread
3. Entry -> list-of- board & thread-> board ->thread
4. Entry -> list-of- board & thread-> board ->thread
5. Entry -> list-of- board-> list-of- board & thread-> board ->thread
6. Entry -> list-of- board-> list-of- board & thread->thread

Path between entry page to index page or between two entry pages is called as index URLs. Path between index pages to thread page is called as thread URLs link between multiple thread pages is called as page flipping URLs. The goal of the forum crawling to reduced URL type recognition problem. FoCUS also show how to learn regular expression patterns, i.e. ITF regexes. The collection of FoCUS is as follows:

1. FoCUS reduces the forum crawling problem to a URL type recognition problem.
2. FoCUS also show how to learning regular expression pattern using index URL, thread URL, and page flipping URL using page classification.
3. Using comparison of breadth-first crawler, structure-driven crawler, iRobot show that FoCUS is effective.

II. 2. Related Work

Vidal et al.[7] describe method of learning regular expression of pattern from entry page to target page it is very effective but working on specific side from sample page. Target page found using DOM tree .The same process has repeated every time for every new page. Yang et al [2] proposed a method for extracting structured data from varies unstructured web forum pages

M. Yang[2] can be use the template independent approach for structure data extraction on pages. They can be providing a robust and accurate extraction. It can be incorporate both page level information and site level information. It can be use template independent approach therefore its work on limited number of websites. In contrast, FOCUS system to learn URL pattern across multiple sites and automatically find target page. Bar Yossef et al. [4] did mention how to discover URLs. Li et al developed some heuristic rules to discover URLs buts alphabetically. In case the URLs list is produce from web server, document sketches are not available. it can be eliminate redundancies in a collection of websites. Recent work on forum crawling is iRobot by Cai et al.[6]. The aim of iRobot is to automatically learn crawler with minimum intervention. First sampling forum pages then clustering them. After that selecting informative clusters and finding traversal path by using spanning tree algorithm. They also introduced skeleton page-flipping link. Skeleton links are supporting to structure of forum site, page flipping links are connectivity of metric. Another related work is to remove duplicates. Content based duplicate detection is not bandwidth efficient cause of it can only used download WebPages. URLs based detection is not helpful so it tries to use rules of different URLs with similar text

III. 3. Terminology

Page type and URL Type:

Entry page: Entry page is a page where arrives at your site some other domain. It is the lowest ancestor of all thread pages in forum.

Index page: It is the page that contains a table. Each row contains information about post. List of thread, list of board pages are index pages.

Thread page: Thread page contains a list of post with user content. Thread page contain actual post information.

Index URL: The URL which is present on entry and index pages.

Thread URL: The URL which is present on index and thread pages. It show title of its destination thread.

Page-flipping URL: The URL that is on same board and same thread pages.

EIT Path: EIT is entry-index-thread path. The navigation path from entry page to thread page using index URL, thread URL, and page-flipping URL.

ITF Regex: An ITF is index-thread-page flipping regular expression. ITF Regex is used to recognize index, thread, or page-flipping URLs on EIT path.

IV. 4. Observation

4.1 Navigation Path

Because of different layout and style, forum has implicit navigation path that leading entry page to thread page. FoCUS describe EIT path to specify type of URL and page to reach thread page.

4.2 URL Layout

The location of URL on page and text length is important to find URL layout information. The anchor text is long in index page and thread page.

4.3 Page Type

In different forum, index page share similar layout. The index layout is different from thread layout. An index page has no of records and thread has records that contain user posts.

V. 5. Architecture Of System

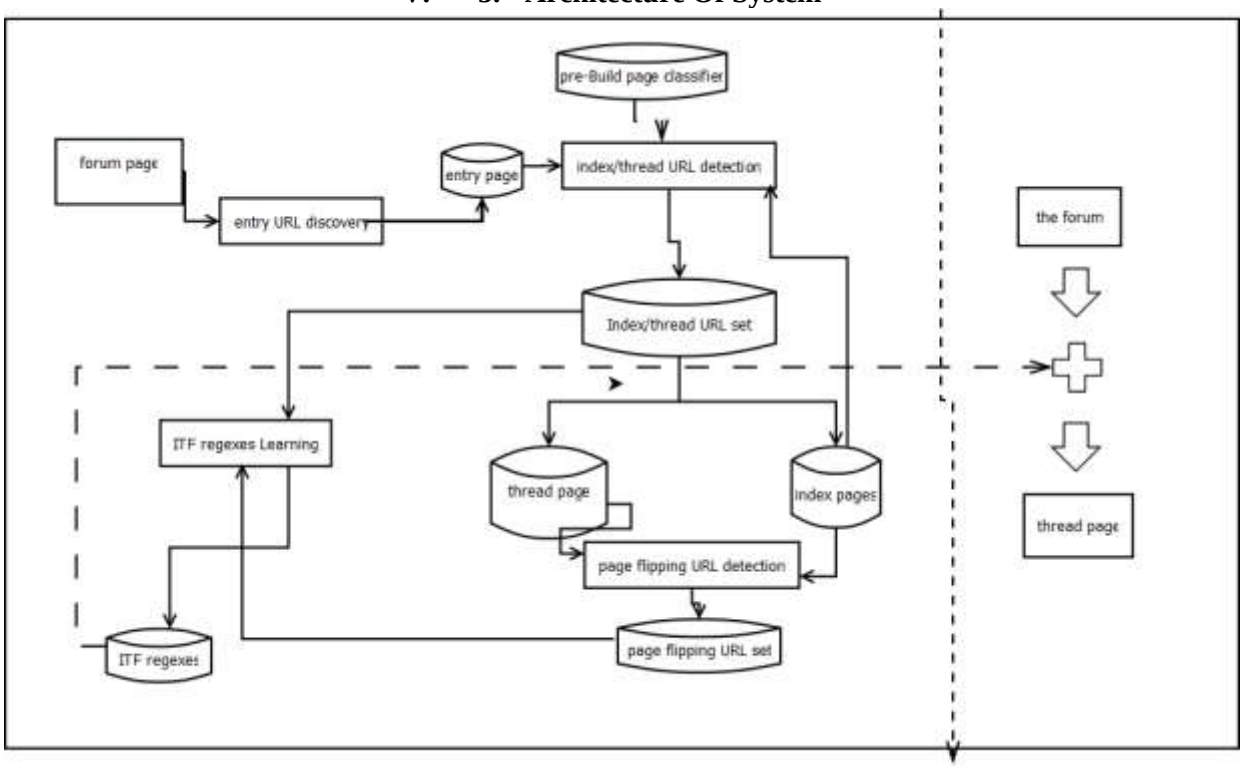

Fig 2: Architecture of FoCUS

Fig. shows the architecture of FoCUS. Architecture of FoCUS consists of Learning & Online crawling. The learning part learns ITF regexes from constructed URL set. The online crawling uses learn ITF regex to crawl. FoCUS consists of number of page types. Using Entry URL Discovery algorithm, FoCUS finds Entry URL with the help of entry URLs. After that using index/thread URL detection algorithm, FoCUS finds index URL, and thread URL on entry page. Detected URLs are store in training set. Again this algorithm detect index and thread URL for destination page. Next, page flipping URL detection algorithm finds page flipping URLs in both page, and saves them into training set. At last ITF regexes learning module learns regular expression from URL training set. Architecture of system performs online crawling as follows: First push entry URL into queue; next it fetches it from queue and downloads its page. Then outgoing URLs are match with learned ITF regex into queue. This step is repeated till queue becomes empty.

5.1 Learning ITF Regexes

5.1.1 Training set construction

The main aim of training set construction is to automatically create set of index URL, thread URL and page-flipping URL string.

a) Training set of index and thread URL:

The URL which present on entry page or index page is called index URL. Its destination page is another index page. The thread URL is present on index page. Its destination is thread page. Note that only using type of destination page we can separate index page and thread page. Every page has its own typical page layout. Index page has no of records and short plain text. Thread page has actual user post. The timestamps are descending order in index page while ascending order in thread page. FoCUS represent page layout and build classifier by using Support Vector Machine (SVM).

b) Training set of page flipping URL

Page-flipping URL is detected using page-flipping URL detection algorithm. The page-flipping URLs are different from index and thread URLs. Detected URL is save in training set.

5.1.2 Learning regex

After creation of index URL, thread URL and page-flipping URL training set, next is to learn regexes from training sets.

5.2 Online crawling

FoCUS first find training set or learn regular expression is described as above. The working of online crawling is based on breath-first strategy. Firstly entry URL push into URL queue and then it fetch from URL queue and download its page. The fetched URL is match to any learned regex into queue. FoCUS repeats this step till queue become empty. Time consuming operations may be performing during its learning phase.

VI. Algorithm

Algorithm 1: Index/Thread URL Detection
Input: entry page or index page.
1) Consider G,b=align DOM tree of entry page or index page and collect group of URL
2) For each URL group in b do
3) C= length of total anchor in url_group
4) End for each
5) G = arg max(c) in group of URL& G.destination page type=Majority type of destination pages
6) If G.destination page type =index page
7) G.type of url = Index_url
8) Else if G.destination page type=thread. page
9) G.type of URL=thread_URL
10) else G
11) end of if condition
12) return G
Algorithm 2: Page-flipping URL Detection
Input: thread page or index page
1) Consider a, b=align DOM tree of thread page or index page and collect URL group
2) fore churl group in b do
3) if anchor text of URL group consist of string of numbers
4) pages =download the url which is present in URL group
5) if URL group occurs at same place in pages as in index or thread page and pages have similar layout to thread or index page
6) b=URL group ,b.type of url=page flipping url
7) break
8) end of if condition
9) end of if condition
10) end for each condition
11) return b

VII. Conclusion

Proposed FoCUS system reduced the crawling problem using URL recognition problem. FoCUS also show implicit navigation path like entry-index-thread (EIT) path. FoCUS is actual learning part so it collect index, thread, and page-flipping URL training set. In FoCUS system, no any time consuming operations and it is effective system.

REFERENCES

[1] Y. Zhai and B. Liu. *Structured Data Extraction from the Web based on Partial Tree Alignment. IEEE Trans. Knowl. Data Eng., 7(12):314−328, 206.*
[2] J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W. –Y.Ma. *Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums. In Proc. of 7th WWW, pages 71-190, 209.*
[3] X.Y. Song, J. Liu, Y.B. Cao, and C.-Y. Lin. *Automatic Extraction of Web Data Records Containing User-Generated Content. In Proc. of 19th CIKM, pages 39-48, 210.*
[4] *Z. Bar-Yossef, I. Keidar, and U. Schonfeld. Do not crawl In the DUST: different URLs with similar text. In Proc. of 3thWWW,* pages 111-12, 207.
[5] S. Brin and L. Page. *The Anatomy of a Large-Scale Hyper textual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.*
[6] 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, 208.*

[7] M. L. A. Vidal, A. S. Silva, E. S. Moura, and J. M. B. Cavalcanti. *Structure-driven Crawler Generation by Example. In Proc. of 29th SIGIR, pages 292-299, 206.*

[8] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song. *FindingQuestion-Answer Pairs from Online Forums. In Proc. of 31stSIGIR, pages 467-474, 208.*

[9] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, andT. Tomokiyo. *Deriving Marketing Intelligence from Online Discussion. In Proc. 11th SIGKDD, pages 419-428, 205.*