Review on Intelligent Crawling Web Forum

Trupti D. Narkhede
PG Student
Computer Department,
MET BKC Adgaon, Nashik, Maharashtra

P.M. Yawalkar
Faculty
Department of Computer Engineering,
MET BKC Adgaon, Nashik
Maharashtra

ABSTRACT

Internet forum are important service where user can request and exchange information with other. The Focus(Forum crawler Under Supervision), it is web-scale forum crawler. A Crawler traverses the World Wide Web in a systematic manner with intention of gathering data or knowledge. The goal of Focus is to crawl applicable forum content from the web. Web crawlers following the hyperlinks in Web pages to automatically download a partial snapshot of the Web. Based on this observation, smaller the forum web crawling problem to a URL-type . Although forum have different layouts or styles and different forum software packages. They have always similar implicit navigation path connected by specific URL types to users from entry pages to thread pages.

Keywords

EIT path, forum crawling, ITF regex, URL Type page classification, page type.

1. INTRODUCTION

An Internet forum, or message board, is the online discussion site where people can discourse in the form of posted Messages. They are different from chat rooms in that messages are at least temporarily achived.[1]

A web is a process which can collect forum data automatically and store it in a database. The data can be used for big data analysis and web content mining.

The Internet forums[3] are important platforms where users can request and exchange information with others. The information in forums, researchers are more and more knowledge from them.[2] 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-forum.[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 duplicate links that point to the common pages but different URL forms. This is mainly two non crawler characteristics of forums 1) repeated links and 2) page-flipping links. A forum typically has many duplicate links that point to the common page but with different URLs generated. e.g., shortcut links pointing to the posts or URLs for user experience the normal of this “view by date” or “view by title.” A generic crawler are follows this links would crawl many duplicate pages, making it inefficient.[5]. A web crawler is an internet that systematically browser 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 have 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 link duplicate pages.[4]

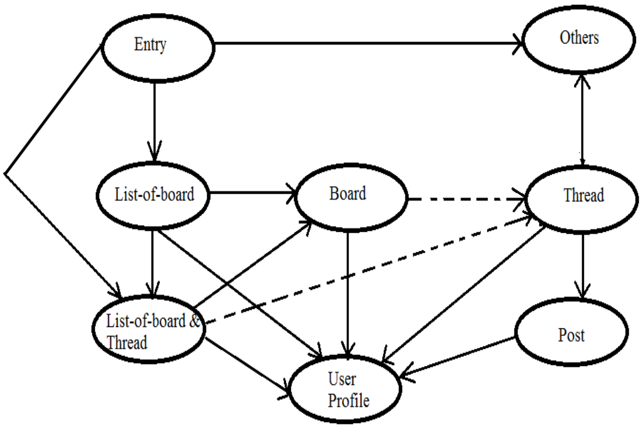

Fig 1: Link relation in a forum.

Fig 1.Illustrates a page and link structure in a forum. The user can navigate from the entry page to thread page

There are the different path for in forum

1.entry→board→thread

2.entry →list-of-board→ board→ thread

3.entry→list-of-board & thread→ thread

4.entry→list-of-board→thread

The pages between the entry page and thread page which are on a the index page. The represent implicit path is called as the EIT path. Forums exit the different layouts or styles and software packages.

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

Entry page →index page→ thread page

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 innumerable new posts created by millions of Internet users to talk about any conceivable topics and issues. To download forum data efficiently and effectively. The characteristics of most forum sites. In general, content of a forum is stored in a database.

2. RELATED WORK

Vidal et al [4] The method 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 [3],there is hundreds of different forum software packages used on the Internet.

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 [5]. iRobot purpose to automatically learn a forum crawler with minimum human mediation by sampling pages, clustering them, and to search the traversal path by a spanning tree algorithm. However, referred as page-flipping URLs. 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. For this ,prototypes of an intelligent forum crawler is proposed and build known as iRobot [5], which 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.

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

Near duplicate file or URL are removed in efficient manner and Finding Near-Duplicate Web Pages need less time to remove duplicates. Some problems may identified duplicate from near pages only not in whole forum page and URL paths removed on whole data pages, not removes in the each and every path.

Entry URL discovery is the trivial problem in existing crawler, since they assume that the URL at the home page is the entry URL and it may be varies for different forum packages and sites. Without entry URL the crawler is less effective.

Another important work is detection of duplicate links. 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 and more robust.

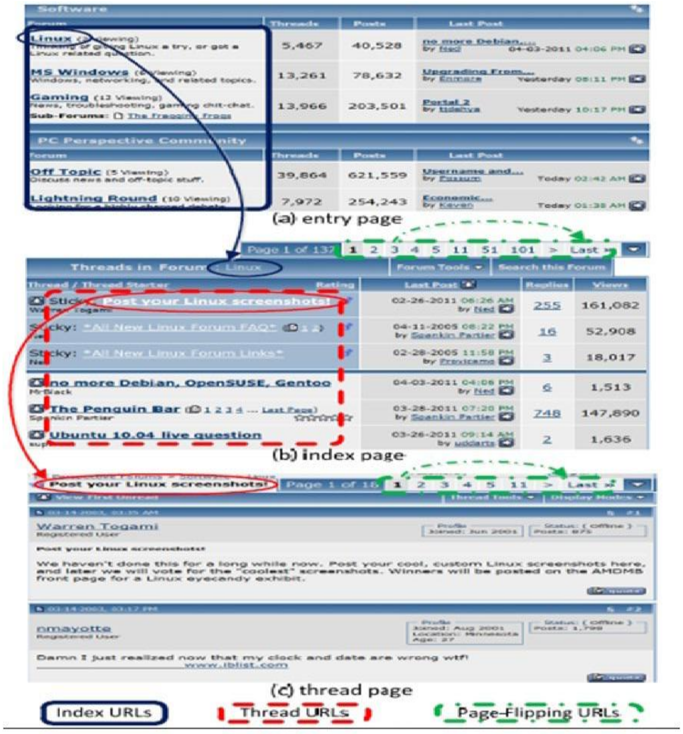

Fig 2. EIT paths: entry →board → thread.

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 homepage of a forum, a list of boards and is also the lowest common ancestor of all threads. See Fig. 2a for an example.
    - Index Page: A page of a board in a forum, which usually contains a table-like structure; each row in it contains information of a board or a thread. See Figs. 2b for examples. In Fig.1, list of- Board page, list-of-board and thread page, and board page are all index pages.
    - Thread Page: A thread of page in a forum that contains a list of posts with user generated. SeeFigs.2c for examples.
- URL Type. There are four types of URL.
    - Index URL: A URL that is on an entry page or index page and points to an index page. It is anchor text shows the title of its destination board. Figs. 2a and 2b show an example.

- Thread URL: A URL that is on an index page and points to a thread page. It is anchor text is the title of its destination thread. Figs. 3b and 3c show an example.
- Page-flipping URL: A URL is leads users to one more page of the similar thread or same board. Correctly page-flipping URLs to make able a crawler to download all thread a large board or all posts of thread. See Figs. 2b, and 2c for examples.

- EIT Path: An entry-index-thread path is a navigation path from an entry page through a sequence of index pages to thread pages.
- ITF Regex: An index-thread-page-flipping regex is a regular expression that can be used to recognize index, thread, or page-flipping URLs. ITF regex is what FoCUS aims to learn and applies directly in 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 URLs.

3. SYSTEM ARCHITECURE

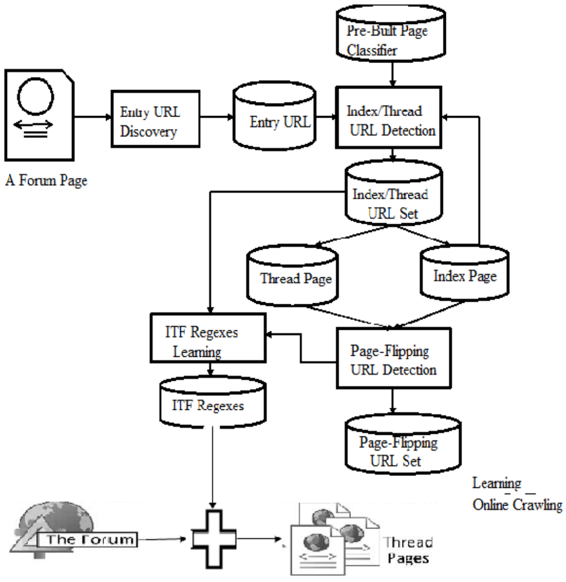

Fig 3. The Overall Architecture of Focus

In this shows the overall architecture of FoCUS. It 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. Given any page of a forum, FoCUS first finds its entry URL using the Entry URL Discovery module. Then, it uses the Index/Thread URL Detection module to detect index URLs and thread URLs on the entry page; the detected index URLs and thread URLs are saved to the URL training sets. Next, the destination pages of the detected index URLs are fed into this module again to detect more index and thread URLs until no more index URL is detected. After the Page-Flipping URL Detection module tries to find page flipping URLs from both index pages and thread pages and saves them to the training sets.

Finally, the ITF Regexes Learning module learns a set of ITF regexes from the URL training sets.

4. ALGORITHMIC STRATEGY

4.1 Algorithm for the Index URL and Thread URL Detection Algorithm

```
Step1 -  Enter data
Step2 - To collect all URL groups and longest
        anchor  text length
Step3 - Select URL group
Step4 - IF the pages are not Index or Thread page
        then  discarded
```

4.2 Algorithm for the Page-Flipping URL Detection Algorithm.

```
Step 1 - To detect the group page-flipping URLs  if
         it  fails
Step 2 - It enumerate all the outgoing URLs to
         detect the single page-flipping URLs
Step 3 -  set its URL type to page-flipping URL
```

4.3 The entry URL Discovery Algorithm

```
Step 1 - URL check in all forums
Step 2 - If keyword found, the path from URL host
Step 3 - Every page in a forum site contains a link to
         lead users back to its entry page.
Step 4 -  URL is detected as an index URL
Step 5-   An entry page have most index URLs Since
          it leads users to all forum threads.
```

5. CONCLUSION AND FUTURE SCOPE

In this paper to reduced the forum crawling problem to a URL type recognition problem and implicit navigation paths of forums, i.e., EIT path, and designed methods to learn ITF regexes explicitly. To smaller the forum crawling problem to a URL type recognition problem .To automatically collect the learn ITF regexes from the training sets and different URL page. A crawler is a program that is used to download and store Web pages, for the web search engine.

A Crawler the world wide web a systematic manner with the intention of gathering data or knowledge or for the aim of web indexing Web crawlers are used for a many purposes.

They are the main components of web search engines, system that assemble a corpus of websites, index them, and permit users to issue queries against the index and find the pages i.e. web pages that match the queries

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 and find the pages.

The forums which use JavaScript, include incremental crawling, and discover new threads and refresh crawled threads in a timely manner can be handled. The initial results of applying FoCUS-like crawler to other social media are very promising.

6. REFERENCES

[1] Jingtian Jiang, Xinying Song, NenghaiYu,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] "ForumMatrix,"http://www.forummatrix.org/index.php, 2012.

[4] M.L.AVidal,A.S.Silva,E.S.Moura,andJ.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.

[5] 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.

[6] 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,” Proc. 18th Int’l Conf. World Wide Web, pp. 181-190, 2009.

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