www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 8 Aug 2015, Page No. 13815-13818

A Review of Web Forum Crawling Techniques

Priyanka Bandagale[1], Dr.Lata Ragha[2], Atiya Kazi[3]

[1]Terna college of Engineering
Mumbai University
priyabandagle@gmail.com

[2]Terna college of Engineering,
Mumbai University,
Lata.ragha@gmail.com

[3]Finolex Academy of management and Technology,
Mumbai University
atiyakazi@gmail.com

Abstract: ***The world wide web contains huge amount of data and it contains numerous websites that is examined by a tool or a program known as Crawler. Due to the richness of the information contributed by lots of internet users every day, internet forum sites have become valuable deposits of material on the web. As a result, mining knowledge from forum sites has become more important and more significant. The main objective of this paper is to focus on the web forum crawling techniques. In this paper, the various techniques of web forum crawler and challenges of crawling are discussed. The paper also gives the outline of web crawling and web forums.***

Keywords: web crawler, web forum, URL pattern.

1. Introduction

Nowadays, there are numerous web forums dealing with diverse topics like news, monetary knowledge, software support, programming discussion, financial data, entertainment and technical discussion. Forums have a specific set of terminology associated with them; e.g., a single conversation is called a "thread", or topic[7]. A discussion forum is tiered or tree-like in structure: a forum can contain a number of sub forums, each of which may have several areas. Each new discussion in the forum's topic is called a thread, and can be replied to by as many people as so desire. While forum crawling is still a demanding task due to complicated in-site link structures and it can sometimes take a long time. The generic crawlers which uses a breadth-first strategy, usually downloads many duplicate and uninformative pages from the forums and they process each page in the page flipping links individually and ignores relationship between the pages. However, proper configuration of website download can speedup website forum scan. A web crawler is also known as web spider, this is program browses in World Wide Web in an automated manner[8]. Crawlers can also be used for specific type of information and then checking links or validating HTML code[5].

Forum Structure:

Each page in forum site may have its own layouts. Based on their layout structure[3], the pages in forum sites are classified into four categories:

- Entry page: The home page of the forum site which contains a list of boards.
- Index page: An index page contains table-like structure, where each row in the table contains information of a board or thread.
- Thread page: A thread page contains a list of users' posts.
- Other pages like login control, about us, user profile pages, etc.

Every forum site has similar navigation paths though they differs in layout and styles structure.

Entry page→ Index page→ Thread page.

The Forum structure is shown in Figure 1 [6].

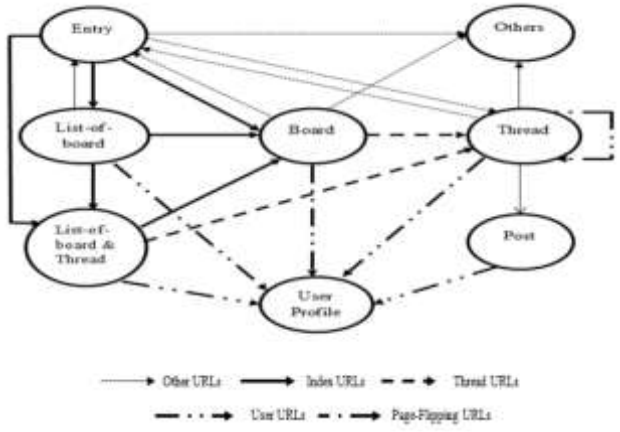

Figure 1: Typical Forum Structure

2. Related work

To make a balance between the "performance and cost", most of the generic Web crawlers implement the breadth-first strategy(BFS)[4] and limit depth of crawling. However, practically it is difficult to select an appropriate crawling depth for each site. A narrow crawling strategy does not give assurance to access all valued content, whereas a deep crawling

strategy may cause in numerous duplicate and invalid pages. Few research works tried to uncover more effective crawling strategy than the BFS to improve the quality of content. A new method is Board Forum Crawling (BFC) to crawl Web forum. This method utilizes the structured characteristics of the Web forum sites and simulates human behavior of visiting Web Forums. This method starts its crawl from the homepage, and then enters each board of the site, and then eventually crawls all the posts of the site directly. The Board Forum Crawling (BFC) can crawl most considerable information of a Web forum site efficiently and in a simple way. A recent and more comprehensive work on forum crawling is iRobot by Cai et al. [1]. The goal of this method is to automatically learn a forum crawler with minimum human interference by sampling forum pages, clustering them, selecting informative clusters via an in formativeness measure. This method uses spanning tree algorithm for finding a traversal path and these selection procedure requires human inspection. Follow up work by Wang et al. [2] proposed an algorithm to address the traversal path selection problem. They introduced the concept of skeleton link and page-flipping link. The most important link underneath the structure or architecture of a forum site known as skeleton link. Forum Crawler under Supervision (FoCUS) [3] is a supervised web-scale forum crawler. FoCUS crawls suitable forum content from the web with least overhead. FoCUS learns uniform resource locator patterns across various sites and automatically locates a forum's entry page given a page from the forum. FoCUS is effective for large-scale forum crawling. FoCUS defines EIT path which permit over one path and URL patterns would not be affected by a change in page structure.

3. Types of Web Forum Crawling Techniques

Web forum crawlers can be categorized as follows:

3.1 Board forum crawling

Typically to visit a post in a forum site, humans visit the home page, then find the post across multiple links. This process proves the structural organisation of forums. Board forum crawling [4] method achieves organized characteristics of forum sites and it simulates the human behaviour of visiting web pages for extracting data form forum sites. This method starts its crawling process from home page, then it extracts the links of board pages from board page seeds, then the subsequent board page seeds are extracted from each board, finally user posts are extracted from each subsequent board page seed. The crawling process for typical home page is as follows:

- Begin mining board page seeds from homepage.
- For every board page seed, a link queue of all successive board pages within the same board is created.
- In the next step for each queue, each page in the queue is downloaded and identified whether belongs to the category of board page. Later, extraction of links of later pages from the board page will be performed.
- This is followed by creation of the whole link index of all post pages in all board pages.
- In the last step, post pages which are linked by the whole index obtained from previous stages are downloaded.

Board forum crawling technique doesn't deal with entry URL discovery problem and it doesn't support duplicate link detection while comparing with other techniques.

3.2 iRobot:

This approach [1] automatically understands the content and structure of each forum sites and then decides how to traverse to different pages in forum sites. To find out such traversal paths, it first automatically re-builds the sitemap of the target web forum and then it selects the best possible traversal paths which only traverses informative pages and skips invalid and duplicate pages.
iRobot system consist of two major parts:

1) offline sitemap reconstructing and traversal path selection
2) online crawling.

The offline sitemap reconstructing and traversal path selection part involves four major steps:

1) Repetitive region-based clustering
2) URL-based sub-clustering
3) Informativeness estimation
4) Traversal path selection

The offline part starts by randomly sampling some few pages form the target forum site. The sampled pages are then given as input to repetitive region-based clustering step. Then the following steps are carried out:

- First, the repetitive pattern (an abstract representation of all records in repetitive region) is generated for each repetitive region in every page.
- Feature description is then created for every page by recording the number of times the repetitive pattern occurs in that page.

ii) URL-Based Sub-Clustering
In this step, each layout clusters that were found in previous step are further split into subsets by grouping those pages with similar URL formats. The similarity between any two URL addresses are found out based on the following two assumptions:

1) URLs having the same number and the same order of the paths are said to be similar.
2) URLs are said to be similar, if both have same parameters of keys.

After the subset clustering, each cluster is finally represented with a URL pattern i.e a sequence of regular expressions generated for every segment of paths and parameters in URL. Each obtained subset cluster is then taken as a vertex of the sitemap. Finally, this step connects various vertices of the sitemap with each other.

iii) Informativeness Estimation
The pages that are having more valuable information are found out in this step. The pages of forum sites are said to be informative if it should satisfy following assumptions:
1) Pages in the large cluster with high probability are valuable.
2) A valuable page usually has relatively large file size.
3) The semantic diversity of each page in cluster is also used to find informativeness of that page. Based on the above three assumptions, the infomativeness measure is calculated for each page. The pages having high informativeness value are

selected and the other remaining pages are discarded in this step.

iv) Traversal Path Selection
The traversal path selection consists of two major parts:
1) cleaning the sitemap
2) optimal traversal path selection.
The sitemap is automatically cleaned by removing most useless vertices and arcs in it, by following the heuristics below:
1) Vertices with low quantity of information are dropped.
2) For a layout cluster, containing several vertices, reserving one representative is enough, as the others are prone to be duplicates.
3) Arcs of self-linking are removed for every vertex, except those whose anchor text is a digital string or some particular strings such as "next". Finally, the optimum traversal path that traverses all the survived vertices with minimum cost is found out. At last, in online crawling step, a downloaded page is first classified into one of the vertex on the sitemap. Then for an out link from that page, the traversal path lookup step further find out its URL pattern and location on that page and finally decide how to follow it by looking up in the lookup table. And for each link in the link table, it decides whether that link should be added to the crawling queue based on the list of traversal paths. The main advantage of iRobot system is that it requires less human intervention and it provides support for duplicate links and uninformative pages removal. But its sampling strategy and informative estimation is not robust.

3.3 ISLK

ISLK stands for Incorporation of Site-Level Knowledge to Extract Structured Data from Forum Sites. This approach [2] focuses on extraction of structured data from forum sites such as users' post title, post content, post time, and post author. It incorporates both page-level knowledge and site-level knowledge and employs markov logic networks to draw the joint inference from both the knowledge's to extract the structured data. For this, the page-level knowledge such as link to user profile, timestamp existence, order of timestamp (ascending or descending) etc are learnt as features from individual pages of forum sites. The site level knowledge representing linkages among different pages in forum sites and the interrelationships of pages belonging to same page type are obtained by reconstructing the sitemap of forum sites. It consists of three main parts:
1) offline sitemap recovering,
2) feature Extraction
3) joint inference of pages having same template.
This step is similar to the offline sitemap recovering step of iRobot technique.

ii) Features Extraction
In this part, the DOM tree is first constructed from the HTML content of the forum page. Once the DOM tree is constructed, the following three kinds of features are extracted:

1)Inner-page features includes the features that leverage the relations among the elements inside each page such as inclusion relation among elements, timestamp existence, order of timestamp, size and location of each elements in page.

2) *Inter-template features* that are generated based on the site-level knowledge includes existence of link to user profile, existence of link to users' posts etc.

3) *Inter-page features* includes existence of text elements having similar DOM path and tag attributes, existence of hyperlink elements with similar DOM path and tag attributes, and existence of inner elements with similar DOM path and tag attributes are extracted.

iii) Joint Inference of Pages having Same Template
In this step, the Markov Logic Networks are used to combine the features extracted in previous step efficiently to draw the joint inference to extract data such as post title, post author, etc. The rules / formulas are defined for extracting data from list pages and post pages separately. Based on those rules the data like list records, list title are extracted from list pages and data like post records, post author, post time, post content are extracted from post pages respectively.

3.4 FOCUS

Given any page of a forum, FOCUS[3] first finds its entry URL using 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 set. Next, the destination pages of the detected index URLs are feed to this module again to detect more index URLs and thread URLs until no more index URL detected. After that, the Page-Flipping URL Detection module tries to find page-flipping URLs in both index pages and thread pages and saves them to the training set. Finally, the ITF Regexes Learning module learns a set of ITF regexes from the URL training set. FoCUS performs online crawling as follows: it first pushes the entry URL into a URL queue; next it fetches a URL from the queue and downloads its page, and then pushes the outgoing URLs that are matched with any learned ITF regex into the URL queue. This step is repeated until the URL queue is empty. To Forum Crawler under Supervision (FoCUS) [3] is a supervised web-scale forum crawler. FoCUS finds pertinent forum content from the web with nominal overhead. FoCUS learns uniform resource locator patterns across multiple sites and automatically finds a forum's entry page given a page from the forum. FoCUS is effective for large-scale forum crawling. FoCUS defines EIT path which permit over one path and URL patterns would not be affected by a change in page structure. It shows way to learn regular expression patterns (ITF regexes) that recognize the index uniform resource locator (URL), thread uniform resource locator (URL) and page-flipping uniform resource locator (URL) using the page classifiers. FoCUS adopts a simple URL string de-duplication technique. The main advantage of FoCUS is that it can avoid duplicates without duplicate detection. This technique uses EIT path to traverse from entry pages through a sequence of index pages to thread pages. EIT means Entry – Index – Thread path. Index URLs are the links between an entry page and an index page or between two index pages. Thread URLs are the links between an index page and a thread page. Page-flipping URLs are the

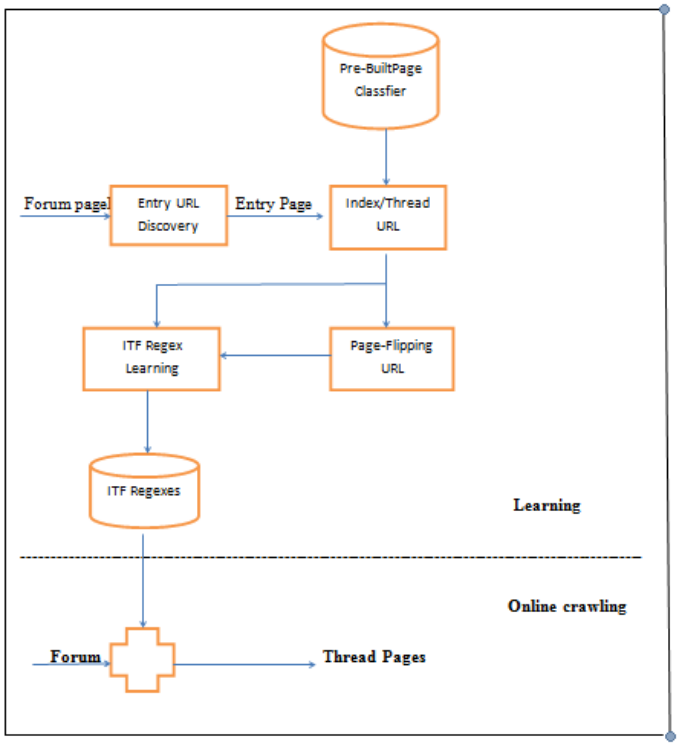

Figure 2: FoCUS Architecture

links connecting multiple pages of a board and multiple pages of a thread. The system consist of two major parts as shown in Figure 2.

- LEARNING PART
- ONLINE CRAWLING PART

The learning part learns ITF regexes of a given forum from automatically constructed URL examples. The online crawling part applies learned ITF regexes to crawl all threads efficiently. The online crawling part then tries to crawl all thread pages that equals the learned ITF regexes. To traverse EIT paths that lead to all thread pages a crawler should starts from the entry URL and needs to follow index URL, thread URL and page-flipping URL. EIT paths and URL patterns are more robust than the traversal path and URL location feature in iRobot.

4. Conclusion

The web crawler collects detail information about the website and the websites links. It includes the website URL, the web page title, the meta tag information, the web page content, the links on the page. In this paper the basic of web crawling is discussed and the survey of different web forum crawling techniques is discussed. FoCUS automatically crawl the forum data and it clean up the unwanted data. After cleaning the unwanted data, FoCUS allocates that space to new queries posted by the user. Comparing with other techniques of web forum crawling, FoCUS outperforms these crawlers in terms of effectiveness and coverage. It shows that the learned patterns are effective and the resulting crawler is efficient.

References

[1] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang. iRobot: An Intelligent Crawler for Web Forums. In *Proc. of 17th WWW*, pages 447-456, 2008.

[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 18th WWW*, pages 181-190, 2009.

[3] J. Jiang, X. Song, and N. Yu, "FoCUS: Learning to Crawl Web Forums," *IEEE Trans. Knowledge and Data Engg*, pp. 1293-1306,2013.

[4] 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.

[5] Gautam Pant, Padmini Srinivasan, and Filippo Menczer, "Crawling the Web," Department of Management Sciences.

[6] Namrata H.S Bamrah , B.S. Satpute, Pramod Patil " Web Forum Crawling Techniques", International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.

[7] InternetForum http://en.m.wikipedia.org/wiki/Internet_forum,2015.

[8] WebCrawler, http://en.m.wikipedia.org/wiki/Web_Crawler, 2015.