ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Mining Internet Web Forums through intelligent Crawler employing implicit navigation paths approach

Jidugu Charishma
M.Tech Student,
Department Of Computer Science Engineering,
NRI Institute Of Technology, Guntur.

S.Venkat Narayana
Professor,
Department Of Computer Science Engineering,
NRI Institute Of Technology, Guntur.

Abstract:

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. They differ from chat rooms in that messages are often longer than one line of text, and are at least temporarily archived. Also, depending on the access level of a user or the forum set-up, a posted message might need to be approved by a moderator before it becomes visible. A focused crawler is a web crawler that collects Web pages that satisfy some specific property, by carefully prioritizing the crawl frontier and managing the hyperlink exploration process.

Some predicates may be based on simple, deterministic and surface properties. For example, a crawler's mission may be to crawl pages from only the .in domain. Other predicates may be softer or comparative, e.g., "crawl pages with large PageRank", or "crawl pages about baseball". An important page property pertains to topics, leading to topical crawlers. For example, a topical crawler may be deployed to collect pages about solar power, or swine flu, while minimizing resources spent fetching pages on other topics.

Crawl frontier management may not be the only device used by focused crawlers; they may use a Web directory, an Web text index, backlinks, or any other Web artifact. In this paper, we present a novel Forum Crawler under Supervision (FoCUS) method, which supervised internet-scale forum crawler.

The intention of FoCUS is to crawl relevant forum information from the internet with minimal overhead, this crawler is to selectively seek out pages that are pertinent to a predefined set of topics, rather than collecting and indexing all accessible web documents to be capable to answer all possible ad-hoc questions.

Keywords:Web Forums, Crawlers, Data mining, page classification, page type, URL pattern learning, URL type

Introduction:

Forums have a specific set of jargon associated with them; e.g. a single conversation is called a "thread", or topic. A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of subforums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish. Depending on the forum's settings, users can be anonymous or have to register with the forum and then subsequently log in in order to post messages. On most forums, users do not have to log in to read existing messages.

A crawler is a program that is used to download and store Web pages, often for a Web search engine. A Crawler traverses the World Wide Web in a systematic manner with the intention of gathering data or knowledge or for the aim of web indexing. Web crawler is also referred as robot or a spider. A web crawler could be a system for the bulk downloading of websites. A crawler starts off by placing an initial set of URLs, in a queue, where all URLs to be retrieved are kept and prioritized. The crawler gets a URL in some order from this queue, downloads the page, extracts any URLs within the downloaded page, and then in the queue it puts the new URLs. This whole process is continued. Finally the collected pages are used later for other applications, like for Web search engine or a Web cache. 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 i.e. web pages that match the queries.

Uses Of Web Crawling:

Web Archiving :

It is a service where large sets of web pages are collected periodically and archived for posterity, provided by the Internet archive.

Web Data Mining :

In Web data mining web pages are analyzed for some statistical properties or where different data analytics is performed on them. Attributor, a company that monitors the web for copyright and trademark infringements can be the example.

A focused crawler must predict the probability that an unvisited page will be relevant before actually downloading the page. A possible predictor is the anchor text of links; this was the approach taken by Pinkerton in a crawler developed in the early days of the Web. Topical crawling was first introduced by Filippo Menczer Chakrabarti et al. coined the term focused crawler and used a text classifier to prioritize the crawl frontier.

Andrew McCallum and co-authors also used reinforcement learning to focus crawlers. Diligenti 'et al. traced the context graph leading up to relevant pages, and their text content, to train classifiers. A form of online reinforcement learning has been used along with features extracted from the DOM tree and text of linking pages, to continually train classifiers that guide the crawl. In a review of topical crawling algorithms, Menczer et al. show that such simple strategies are very effective for short crawls, while more sophisticated techniques such as reinforcement learning and evolutionary adaptation can give the best performance over longer crawls.

Crawlers are also focused on page properties other than topics. Cho et al. study a variety of crawl prioritization policies and their effects on the link popularity of fetched pages. Najork and Weiner show that breadth-first crawling, starting from popular seed pages, leads to collecting large-PageRank pages early in the crawl. Refinements involving detection of stale (poorly maintained) pages have been reported by Eiron et al.

The performance of a focused crawler depends on the richness of links in the specific topic being searched, and focused crawling usually relies on a general web search engine for providing starting points.

Davison presented studies on Web links and text that explain why focused crawling succeeds on broad topics; similar studies were presented by Chakrabarti et al. Seed selection can be important for focused crawlers and significantly influence the crawling efficiency.

A whitelist strategy is to start the focus crawl from a list of high quality seed URLs and limit the crawling scope to the domains of these URLs. These high quality seeds should be selected based on a list of URL candidates which are accumulated over a sufficient long period of general web crawling. The whitelist should be updated periodically after it is created.

Existing System:

The existing system is a manual or semi automated system, i.e. The Textile Management System is the system that can directly sent to the shop and will purchase clothes whatever you wanted.

The users are purchase dresses for festivals or by their need. They can spend time to purchase this by their choice like color, size, and designs, rate and so on.

They But now in the world everyone is busy. They don't need time to spend for this. Because they can spend whole the day to purchase for their whole family. So we proposed the new system for web crawling.

Disadvantages:

1. Consuming large amount of data's.

2. Time wasting while crawl in the web.

Proposed System:

We propose a new system for web crawl as FoCUS: Learning to Crawl Web Forums. It is a system overcome by existing crawl systems. In this method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Target pages were found through comparing DOM trees of pages with a pre-selected sample target page. It is very effective but it only works for the specific site from which the sample page is drawn. The same process has to be repeated every time for a new site. Therefore, it is not suitable to large- scale crawling.

In contrast, FoCUS learns URL patterns across multiple sites and automatically finds forum entry page given a page from a forum. Experimental results show that FoCUS is effective in large scale forum crawling by leveraging crawling knowledge learned from a few annotated forum sites. A recent and more comprehensive work on forum crawling is iRobot. iRobot aims to automatically learn a forum crawler with minimum human intervention by sampling forum pages, clustering them, selecting informative clusters via an informativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection.

Advantages:

FoCUS learns URL patterns across multiple sites and automatically finds forum entry page given a page from a forum. Experimental results show that FoCUS is effective in large scale forum crawling by leveraging crawling knowledge learned from a few annotated forum sites.

Implementation:

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

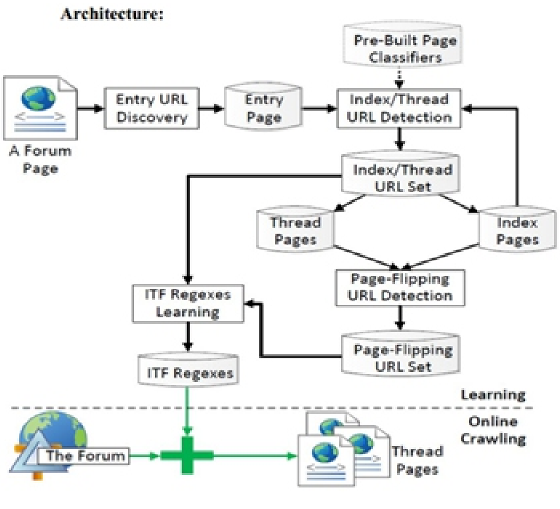
Problem Statement:

A crawler starting from the entry page of a forum only needs to follow index URLs, thread URLs, and page-flipping URLs to traverse EIT path and achieve all thread pages. The challenge of forum crawling is then reduced to a URL type recognition problem. In this paper, we show how to learn regular expression patterns, i.e. ITF regexes, recognizing these three types of URLs from as few as 5 annotated forum packages and apply them to a large set of 160 unseen forums packages. Note that we specifically refer to "forum package" rather than "forum site".

Scope:

The goal of FoCUS is to trawl relevant content, i.e. user posts, from forums with minimal overhead. Forums exist in many different layouts or styles and powered by a variety of forum software packages, but they always have implicit navigation paths to lead users from entry pages to thread pages. We call pages between the entry page and thread page which are on a breadth-first navigation path the index page. We represent these implicit paths as the following navigation path (EIT path):

entry page index page thread page Links between an entry page and an index page or between two index pages are referred as index URLs. Links between an index page and a thread page are referred as thread URLs.



MODULES :

1.Signup & Login
2.Upload New Files
3.Crawl On Web
4. Forum Pages

MODULE DESCRIPTION:

1.Signup & Login:

In this module, we have two sub modules. They are, User signup & login: In this module user can create account with our site by filling details. And then they can login with our site using this user name and password Admin login: The owner of this system have a own user name and password for login with the page.

2.Upload File:

In this module the owner of the site have to upload a new file for crawl in this site. The user of the page wants to crawl in the site. So the admin should upload a maximum of files for the users need. Also the admin can view the user details those are having account in his page. And they can view files which they are already uploaded in database.

3.Crawl in Web:

The goal of this paper is crawl on the web. So the user can view files in this site which they are uploaded by admin. The users can search a files what they need to know about that. Also they can view the related searches based on their search. The search contains additional links of its contents also. This web crawling proposed like tree search. And then user can view their own details which they already gave while signup with this site. They also can change / modify the details.

4. Forum Pages :

We classified forum pages into four page types:

Entry Page: A page that is the lowest common ancestor of all threads pages in a forum.

Index Page: A page that contains a table-like structure; each row in it contains information on URLs pointing to a board or a thread.

Thread Page: A page that contains a list of posts with user generated content (UGC).

Other Page: A page that is not an entry, index, or thread page.

ALGORITHMS:

Index URL and Thread URL Detection

```
Algorithm: INdexUrlThreadUrlDetection
Input: sp:an entry or index page
Output: it_group:a group of index/thread URLs
1:letit_group be p:data
2:url groups=Collect URL groups by aligning HTML DOM tree of sp;
3.foreachug in url_groups do
4.uganchor_len=Total anchor text length in ug;
5. endforeach
6:if_group=argmax(ug.anchor_len)in url_groups;
7:if_group.DstPageType=Majority page type of the destination pages of URLs in ug;
8.if_group.DstPageType is INDEX_PAGE
9. if_group.Urltype=INDEX_URL;
10. else if if_group.DstPageType is THREAD_PAGE
11.if_group.Urltype=Thread_URL;
12. else
13. if_group=;
14. end if
15:returnif_group;
```

Page-Flipping URL Detection

```
Algorithm: Page-FlippingUrlThreadUrlDetection
Input: sp:an index page or thread page
Output: if_group:a group of page_flipping URLs
1:letpf_group be
2:url groups=Collect URL groups by aligning HTML DOM tree of sp;
3:foreachug in url_groups do
4:if the anchor texts of ug are digit strings
5:pages=Download(URLs in ug);
6:if pages=have the similar layout to sp and ug appears at same location of pages as in sp
7:pf_group=ug;
8:break;
9:end if
10:end if
11:endforeach
12:ifpf_group is
13:foreachurl in outgoing URLs in sp
14:P=Download(url);
15:pf_url=Extract URL in p at the same location as url in sp;
```

```
ifpf_url exists and pf_url.anchor==url.anchor and pf_url.UrlString|=url.UrlString
Addurl and cand_url into pf_group;
break;
end if
endforeach
end if
pf_group UrlType=PAGE_FLIPPING_URL;
returnpf_group;
```

Conclusion:

In the paper we present tale method crawler which downloads and stores web pages, frequently for a web search engine. The rapid growth of internet poses more challenges to search for suitable link. We also symbolize the technique of FOCUS which are developed to extract only the relevant web pages of interested topic from the Internet. The design of FOCUS is capable to evaluate the text which found on a link with the input text file.

The crawler uses pattern recognition and generates the number of times the input text exists in the text establish on a link. The information so generated gives an imminent in the efficiency of the pattern-matching. FoCUS constantly keeps on crawling the internet and finds any new internet pages that have been added to the web, pages that have been detached from the web.

Due to growing and vibrant activity of the internet; it has become a confront to traverse the URLs in the web documents and to handle these URLs. We will take the seed URL as input and search with a keyword, the searching result is based on keyword and it will obtain the web pages where it will find that keyword.

References:

[1] FoCUS: Learning to Crawl Web Forums Jingtian Jiang, Nenghai Yu, Chin-Yew Lin, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 6, JUNE 2013.

[2] Forum Software Reviews. http://www.forumsoftware.org/forum-reviews.

[3] Message Boards Statistics. http://www.bigboards.com/statistics/[4] Z. Bar-Yossef, I. Keidar, and U. Schonfeld. Do not crawl in the DUST: different URLs with similar text. In Proc. of 16th WWW, pages 111-120, 2007.

[5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual 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 17th WWW, pages 447-456, 2008.

[7] A. Dasgupta, R. Kumar, and A. Sasturkar. De-duping URLs via rewrite rules. In Proc. of 14th KDD, pages 186-194, 2008.

[8] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song. Finding Question-Answer Pairs from Online Forums. In Proc. of 31st SIGIR, pages 467-474, 2008.

[9] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving Marketing Intelligence from Online Discussion. In Proc. 11th SIGKDD, pages 419-428, 2005.

[10] Y. Guo, K. Li, K. Zhang, and G. Zhang. Board Forum Crawling: a Web Crawling Method for Web Forum. In Proc. of 2006 IEEE/WIC/ACM WI, pages 475-478, 2006.

[11] M. Henzinger. Finding near-duplicate Web pages: a large scale evaluation of algorithms. In Proc. of 29th SIGIR, pages 284-291, 2006.

[12]X.Y. Song, J. Liu, Y.B. Cao, and C.-Y. Lin, “Automatic Extraction of Web Data Records Containing User-Generated Content,” Proc. 19th Intl Conf. Information and Knowledge Management, pp. 39-48, 2010.

[13]V.N. Vapnik,The Nature of Statistical Learning Theory. Springer, 1995.

[14]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.

[15]Y. Wang, J.-M. Yang, W. Lai, R. Cai, L. Zhang, and W.-Y. Ma, "Exploring Traversal Strategy for Web Forum Crawling," Proc. 31st Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 459-466, 2008.

[16]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 Intl Conf. World Wide Web,pp. 181-190, 2009.

[17]Y. Guo, K. Li, K. Zhang, and G. Zhang, "Board Forum Crawling: A Web Crawling Method for Web Forum," Proc. IEEE/WIC/AC M Intl Conf. Web Intelligence, pp.475-478, 2006.

[18]M. Henzinger, "Finding Near-Duplicate Web Pages: A Large- Scale Evaluation of Algorithms," Proc. 29th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 284-291,2006.

[19]H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg, and A. Sasturkar, "Learning URL Patterns for Webpage De- Duplication," Proc. Third ACM Conf. Web Search and Data Mining, pp. 381-390, 2010.

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