EWFPC: Extraction of Web Forums Using Page Type Classifiesr

[1]Gotivada Swamy, [2]Srinivasa Rao Dalai, [3]A.Phani Sridhar
[1]M.Tech, (Computer Science)
[2]Asst. Professor, Dept. of CSE
[3]Head of the Department, CSE

Abstract

In this paper, we present EWFPC (EXTRACTION OF WEB FORUMS USING PAGE TYPE CLASSIFIESR), a supervised web-scale forum crawler. The goal of EWFPC is to only trawl relevant forum content from the web with minimal overhead. Forum threads contain information content that is the target of forum crawlers. Although forums have different layouts or styles and are powered by different forum software packages, they always have similar implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages.
Based on this observation, we reduce the web forum crawling problem to a URL type recognition problem and show how to learn accurate and effective regular expression patterns of implicit navigation paths from an automatically created training set using aggregated results from weak page type classifiers.

In this paper, we present EWFPC(EXTRACTION OF WEB FORUMS USING PAGE TYPE CLASSIFIESR), a supervised web-scale forum crawler, to address these challenges. The goal of EWFPC is to trawl relevant content, i.e. user posts, from forums with minimal overhead.

By applying Index URL Thread URL Detection,Page-Flipping URL Thread URL Detection,Entry URL Discovery algorithms.

Keywords

Index URL Thread URL Detection,Page-Flipping URL Thread URL Detection,Entry URL Discovery, Forum Crawlers, Web Forum Crawling Problem

I. Introduction

Internet forums are important platforms where users can request and exchange information with others. For example, the Trip Advisor Travel Board is a place where people can ask and share travel tips.

A Forum typically has many uninformative pages such as login control to protect users' privacy. Following these links, a crawler will trawl many uninformative pages. Though there are standard-based methods such as specifying the "rel" attribute with "no follow" value (i.e. "rel=no follow")2, Robots Exclusion Standard (robots.txt)3, and Sitemap4, for forum operators to instruct web crawlers on how to crawl a site effectively, we found that over a set of 9 test forums more than 47% of the pages trawled by a generic crawler following these protocols are duplicate or uninformative.

In this paper, we present EWFPC(EXTRACTION OF WEB FORUMS USING PAGE TYPE CLASSIFIESR), a supervised web-scale forum crawler, to address these challenges. The goal of EWFPC is to trawl relevant content, i.e. user posts, from forums with minimal overhead.

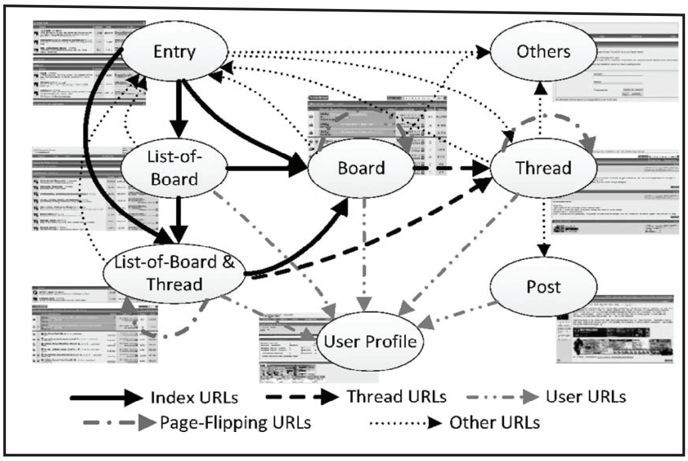


Fig. 1: Example Link Relations in a Forum

Fig. 1 illustrates a typical page and link structure in a forum. For example, a user can navigate from the entry page to a thread page through the following paths:
1. Entry ⇨ board ⇨ thread
2. Entry ⇨ list-of-board ⇨ board ⇨ thread
3. Entry ⇨ list-of-board & thread ⇨ 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
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. Links connecting multiple pages of a board and multiple pages of a thread are referred as page-flipping URLs.

The major contributions of this paper are as follows:
1. We reduce the forum crawling problem to a URL type recognition problem and implement a crawler, EWFPC, to demonstrate its applicability.
2. We show how to automatically learn regular expression patterns (ITF regexes) that recognize the index URL, thread URL, and page-flipping URL using the page classifiers built from as few as 5 annotated forums and show that EWFPC outperforms these crawlers in terms of effectiveness and coverage.
3. We evaluate EWFPC on a large set of 160 unseen forum packages that cover 668,683 forum sites. To the best of our knowledge, this is the largest scale evaluation of this type. In addition, we show that the patterns are effective and the resulting crawler is efficient.

4. We compare EWFPC with a baseline generic breadth-first crawler, a structure-driven crawler, and a state-of-the-art crawler iRobot.
5. We design an effective forum entry URL discovery method. Entry URLs need to be specified to start crawling to get higher recall. But entry page discovery is not a trivial task since entry pages vary from forums to forums.
6. We show that, though the proposed approach is targeted at forum crawling, the implicit EIT like path also apply to other User Generated Content (UGC) sites, such as community Q&A sites and blog sites.

The rest of this paper is organized as follows. we define terms used in this paper. We give our observations on forums and describe the detail of the proposed approach in we report the results of our experiments. In the last section, we draw conclusions and point out future directions of research.

A. Problem Statement

We evaluate EWFPC on a large set of 160 unseen forum packages that cover 668,683 forum sites. To the best of our knowledge, this is the largest scale evaluation of this type. In addition, we show that the patterns are effective and the resulting crawler is efficient.

II. 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.

Disadvantages

1. Consuming large amount of data’s.
2. Time wasting while crawl in the web.

III. Proposed System

We propose a new system for web crawl as **EWFPC: 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. 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.

A. Advantages

1. We show how to automatically learn regular expression patterns (ITF regexes) that recognize the index URL, thread URL, and page-flipping URL using the page classifiers built from as few as five annotated forums.
2. We evaluate EWFPC on a large set of 160 unseen forum packages that cover 668,683 forum sites. To the best of our knowledge, this is the largest evaluation of this type. In addition, we show that the learned patterns are effective and the resulting crawler is efficient.

B. Observations

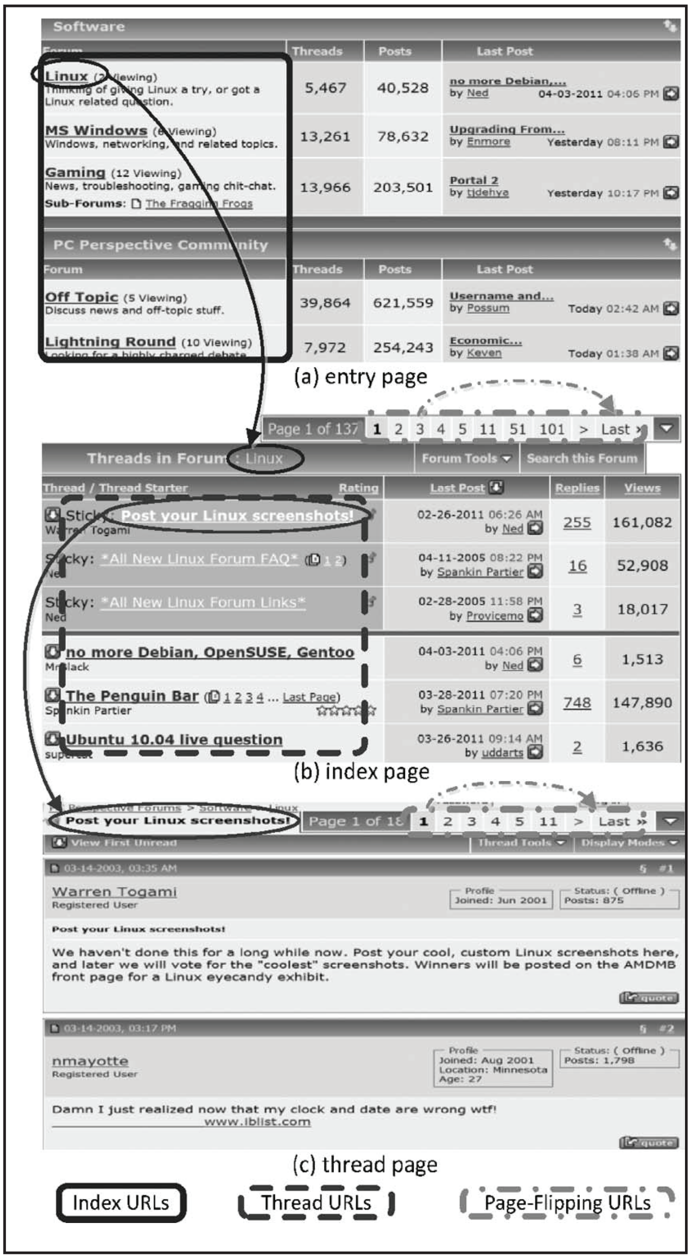
In order to crawl forum threads effectively and efficiently, we investigated about 40 forums (not used in testing) and found the following characteristics in almost all of them.

C. Navigation Path

Despite differences in layout and style, forums always have implicit navigation paths leading users from their entry pages to thread pages. In general crawling, Vidal et al. learned “navigation patterns” leading to target pages (thread pages in our case).

D. URL Layout

URL layout information such as the location of a URL on a page and its anchor text length is an important indicator of its function. URLs of the same function usually appear at the same location.



Software

Forum	Threads	Posts	Last Post
Linux (2 Viewing) Thinking of giving Linux a try, or got a Linux related question.	5,467	40,528	no more Debian,... by Ned 04-03-2011 04:06 PM
MS Windows (6 Viewing) Windows, networking, and related topics.	13,261	78,632	Upgrading From... by Enmore Yesterday 08:11 PM
Gaming (12 Viewing) News, troubleshooting, gaming chit-chat. Sub-Forums: The Fraggin Frogs	13,966	203,501	Portal 2 by tjdehye Yesterday 10:17 PM

PC Perspective Community

Forum	Threads	Posts	Last Post
Off Topic (5 Viewing) Discuss news and off-topic stuff.	39,864	621,559	Username and... by Possum Today 02:42 AM
Lightning Round (10 Viewing) Looking for a highly charged debate	7,972	254,243	Economic... by Keven Today 01:38 AM

Page 1 of 137 1 2 3 4 5 11 51 101 > Last »

Threads in Forum : Linux Forum Tools Search this Forum

Thread / Thread Starter	Rating	Last Post	Replies	Views
Sticky: Post your Linux screenshots! Warren Togami		02-26-2011 06:26 AM by Ned	255	161,082
Sticky: *All New Linux Forum FAQ* (1 2) Ned		04-11-2005 08:22 PM by Spankin Partier	16	52,908
Sticky: *All New Linux Forum Links* Ned		02-28-2005 11:58 PM by Provicemo	3	18,017
no more Debian, OpenSUSE, Gentoo MrBlack		04-03-2011 04:06 PM by Ned	6	1,513
The Penguin Bar (1 2 3 4 ... Last Page) Spankin Partier	☆☆☆☆☆	03-28-2011 07:20 PM by Spankin Partier	748	147,890
Ubuntu 10.04 live question supercat		03-26-2011 09:14 AM by uddarts	2	1,636

PC Perspective Forums > Software > Linux

Post your Linux screenshots! Page 1 of 18 1 2 3 4 5 11 > Last »

View First Unread Thread Tools Display Modes

03-14-2003, 03:35 AM #1

Warren Togami
Registered User
Profile Joined: Jun 2001
Status: (Offline) Posts: 875

Post your Linux screenshots!

We haven't done this for a long while now. Post your cool, custom Linux screenshots here, and later we will vote for the "coolest" screenshots. Winners will be posted on the AMDMB front page for a Linux eyecandy exhibit.

quote

03-14-2003, 03:17 PM #2

nmayotte
Registered User
Profile Joined: Aug 2001 Location: Minnesota Age: 27
Status: (Offline) Posts: 1,798

Damn I just realized now that my clock and date are wrong wtf!
www.iblist.com

quote

Fig. 2:

E. Page Layout

Index pages from different forums share a similar layout. The same applies to thread pages. For example, the index pages from two different forums have the similar page layout. However, an index page usually has a very different page layout from a thread page.

This is the only step where manual annotation is required for EWFPC. Inspired by these observations, we developed EWFPC. The main idea behind EWFPC is that index URL, thread URL, and page-flipping URL can be detected based on their layout characteristics and destination pages; and forum pages can be classified by their layouts. This knowledge about URLs and pages and forum structures can be learned from a few annotated forums and then applied to unseen forums.

The architectural design for this research work has been also conducted which is shown below.

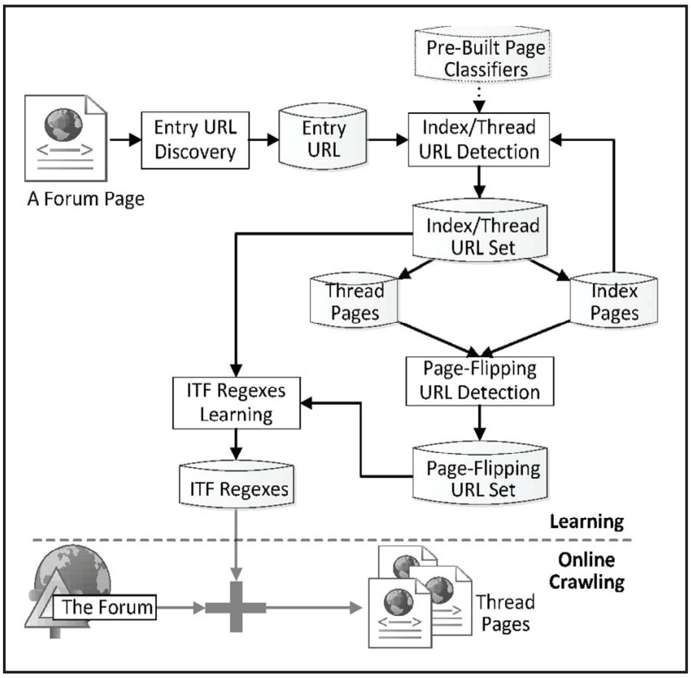


Fig. 3:

F. System Overview

Fig. 4 shows the overall architecture of EWFPC. 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.

G. Page-Flippingurl Training Set

Page-flipping URLs point to index pages or thread pages but they are very different from index URLs or thread URLs. Wang et al. proposed "connectivity" metric to distinguish page flipping URLs from other loop-back URLs. However, the metric only works well on the "grouped" page-flipping URLs, i.e., more than one page-flipping URL in one page.

Algorithm: Index URL Thread URL Detection

Input: sp:an entry or index page

Output: it_group:a group of index/thread URLs

```
letit_group be p:data
url groups=Collect URL groups by aligning HTML DOM tree of sp;
for eachug in url_groupsdo
    ug.anchor_len=Total anchor text length in ug;
end foreach
if_group=argmax(ug.anchor_len)in url_groups;
if_group.DstPageType=Majority page type of the destination pages of URLs in ug;
if_group.DstPageType is INDEX_PAGE
    if_group.Urltype=INDEX_URL;
else ifif_group.DstPageType is THREAD_PAGE
    if_group.Urltype=Thread_URL;
else
    if_group=p;
end if
return if_group;
```

In particular, the grouped page-flipping URLs have the following properties:

1. Their anchor text is either a sequence of digits such as 1, 2, 3, or special text such as "last."
2. They appear at the same location on the DOM tree of their source page and the DOM trees of their destination pages.
3. Their destination pages have similar layout with their source pages. We use tree similarity to determine whether the layouts of two pages are similar or not. As to single page-flipping URLs, they do not have the property 1, but they have another specialproperty.
4. The single page-flipping URLs appearing in their source pages and their destination pages have the same anchor text but different URL strings.

H. Learning ITF Regexes

We have shown how to create index URL, thread URL, and page-flipping URL string training sets; next we explain how to learn ITF regexes from these training sets.

Algorithm: Page-Flipping URL Thread URL Detection

Input: sp:an index page or thread page

Output: if_group:a group of page_flipping URLs

```
letpf_group be
url groups=Collect URL groups by aligning HTML DOM tree of sp;
for eachug in url_groupsdo
    if the anchor texts of ug are digit strings
        pages=Download(URLs in ug);
        if pages=have the similar layout to spand ug appears at same location of pages as in sp
            pf_group=ug;
            break;
        end if
    end if
endforeach
ifpf_group is
    foreachurl in outgoing URLs in sp
        P=Download(url);
        pf_url=Extract URL in p at the same location as url in sp;
        ifpf_url exists andpf_url.anchor==url.anchorandpf_url.UrlString|=url.UrlString
            Addurl and cand_url into pf_group;
            break;
        end if
    endforeach
end if
pf_groupUrlType=PAGE_FLIPPING_URL;
return pf_group
```

Each pattern matches a subset of URLs. These patternsare refined recursively until no more specific patterns canbe generated. These three patterns are the final output as they cannot be refined further.

I. Online Crawling

Given a forum, EWFPC first learns a set of ITF regexesfollowing the procedure described in previous sections. Thenit performs online crawling using a breadth-first strategy (actually, it is easy to adopt other strategies).

EWFPC firstpushes the entry URL into a URL queue; next it fetches a URLfrom the URL queue and downloads its page;

and then itpushes the outgoing URLs that are matched with any learnedregex into the URL queue.

EWFPC repeats this step until theURL queue is empty or other conditions are satisfied. What makes EWFPC efficient in online crawling is that itonly needs to apply the learned ITF regexes on new outgoing URLs in newly downloaded pages.

Algorithm:Entry URL Discovery

Input:url: a aurl pointing to a page from a forum

Output: Entry_url:entryurl from a forum

```
b_url=GetNaiveEntryUrl(url);//baseline
p=Download(url);
urls=extract outgoing URLs in p that start with b_url
samp_urls=randomly sample a few URLs from urls;
Add the host of url into samp_urls; // observations(2)
foreach u in samp_urlsdo
    p=Download(u);
    urls=urls {outgoing urls in p strting with b_url;// observation(1)
each foreach
let entry_url be b_url,index_urls be ,count be 0;
foreach u in urls do
    if u is in index_urlscontinue;//observation (3)
    p=Download(u);
    i_urls=detect index URLs in p;
    inde_urls=index_urls U i_urls;
    if count<i_urls //observation (4)
        Count= i_urls;
        entry_url=u;
    end if
end foreach
returnentry_url;
```

1. Almost every page in a forum site contains a link tolead users back to its entry page. Note that thesepages are from a forum site.Aforum site might not bethe site hosting this forum. For example, http://www. englishforums.com/English/ is a forum sitebut http://www.englishforums.com/ is not a forumsite.
2. The home page of the site hosting a forum mustcontain the entry URL of this forum.
3. If a URL is detected as an index URL, it should notbe an entry URL.
4. An entry page have most index URLs since it leadsusers to all forum threads. Based on the aboveobservations and the index URL detection module described the "Index URL and Thread URL Training Sets".

J. Evaluations of EWFPC Modules

To further check how many annotated pages EWFPC needs to achieve good performance. We conducted similar experiments but with more training forums (10, 20, 30, and 40) and applied cross validation.

We find that our page classifiers achieved over 96 percent recall and precision at all cases with tight standard deviation. It is particularly encouraging to see that EWFPC can achieve over 98 percent precision and recall in index/thread URL detection with only as few as five annotated forums.

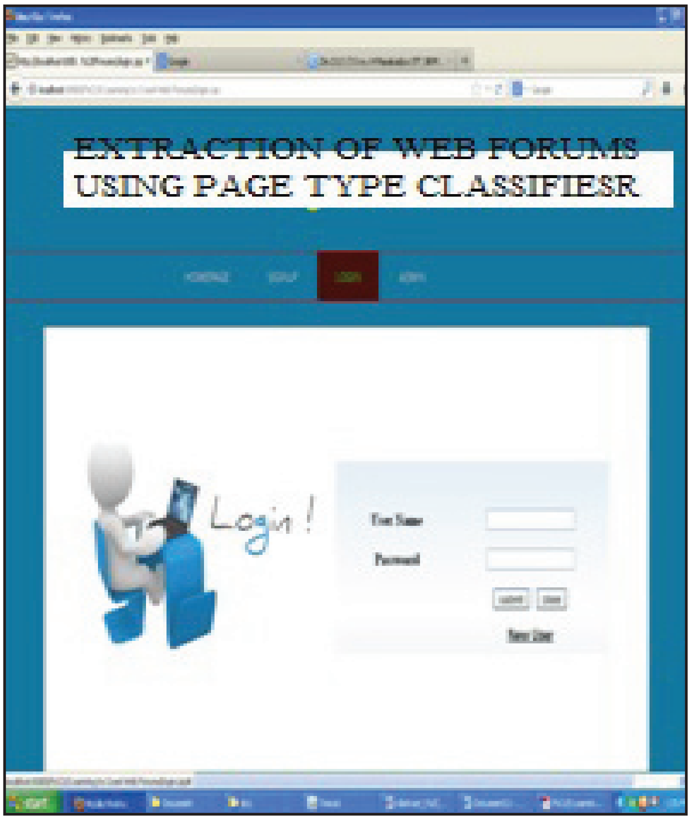
K. Evaluation of Page-Flipping URL Detection

To test page-flipping URL detection, we applied the module\ described "Page-Flipping URL Training Set"on the 10-Page/160 test set and manually checked whether it found the correct URLs. The method achieved 99 percent precision and 95 percent recall. The failure is mainly due to JavaScript-based page-flipping URLs or HTML DOM tree alignment error.

IV. Future Work

In the future, we would like to handle forums which use JavaScript, include incremental crawling, and discover new threads and refresh crawled threads in a timely manner.

V. Sample Results



EXTRACTION OF WEB FORUMS USING PAGE TYPE CLASSIFIESR

Fig. 4: User Login

This is the Login Page for the User.He gives his user name and password.

References

[1] T. Taleb, K. Hashimoto,"MS2: A Novel Multi-Source Mobile-Streaming Architecture", In IEEE Transaction on Broadcasting, Vol. 57, No. 3, pp. 662–673, 2011.

[2] X. Wang, S. Kim, T. Kwon, H. Kim, Y. Choi,"Unveiling the BitTorrent Performance in Mobile WiMAX Networks", In Passive and Active Measurement Conference, 2011.