WEB FORUMS CRAWLER FOR ANALYSING USER SENTIMENTS

Raghu.D[#1], Saravana Jothi.M[*2], Kaviarasu.N[#3], Syed Iejas.S[*4]

[#]*Department of Computer Science and Engineering, Anna University*
K.S.R.College of Engineering, Tiruchengode-637 215,
Namakkal district, Tamil Nadu
[1]raghud93@gmail.com
[3]kavi.n164@gmail.com

[*] *K.S.R.College of Engineering, Tiruchengode-637 215,*
Namakkal district, Tamil Nadu
[2]saravanajothi05@gmail.com
[4]ijaz.sms@gmail.com

Abstract— **Here, we present Web forum crawler based on user sentimental, a supervised web-scale forum crawler. The goal of web crawler is to only analyse suitable forum content from the web with minimal overhead. Mainly, 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. And each forum has a specific feedback analysis to learn the user sentiment. Based on this observation, we reduce the web forum crawling problem to a URL type recognition problem and base on that the user sentiment are analysed by means of the feedbacks given by them helps to learn it accurate and effective regular expression patterns of implicit navigation paths and user satisfaction from an automatically created training set using aggregated results to make impact on weak page type classifiers. Mostly forum crawlers are analysing forums based on their URL's it causes a major drawback over negotiation of contents of forums. To overcome that we go for User sentiment based web forum crawling.**

Keywords— **Forum crawling, ITF regex, page type, page URL, forum feedback, thread feedbacks, user sentiment, Post nature classification.**

I. Introduction

Mainly Internet forums are discussion boards about the products, current trends in technology, certain topics that are of different categorized and doubts about the things what have to be done. Researchers go for forums to gain knowledge and also sharing their knowledge. Mostly forums act as a database by means of collecting a large and variety of information. Existing technique is composed of mining the crawl logs and using clusters of same pages to extract transformation rules, which are used to identify URLs belonging to each cluster. URLs are identified based on three ways:

- Thread URL
- Index URL
- Page flip URL

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**. A crawler starting from the entry URL only needs to follow index URL, thread URL, and page-flipping URL to traverse EIT paths that lead to all thread pages. The challenge of forum crawling is then reduced to a URL type recognition problem. By which,

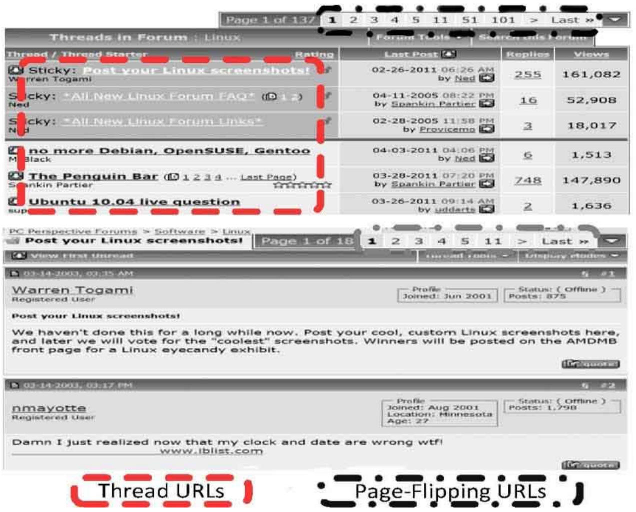
- They get reduce the forum crawling problem to a URL type recognition problem.
- Recognize the index URL, thread URL, and page-flipping URL using the page classifiers.
- Compare FoCUS with a baseline generic breadth first crawler, a structure-driven crawler.
- Design an effective forum entry URL discovery method.
- From its result, we can identify the URLs related to the post made by user over threads and post.
- Using this we cluster words from the user comments as positive, negative and average.
- Then based on number of comments, each threads and page gets ranked and listed in database.
- So that we can easily browse through our related contents and get best outcomes.

Our test results show that FoCUS achieved over 98% effectiveness and 97% coverage on a large set of test forums powered by over 150 different forum software packages. The rest of this paper is organized as follows.

Section 2 is a brief review of related work. In Section 3, we define analysis through URLs. We give our observations on forums and describe the detail of the proposed approach in Section 4. In Section 5, we report the results of our experiments. In the last section, we draw conclusions of our research work.

II. RELATED WORK

From [1] we learned about the usages and process takes throughout the forums. By which we gather information about their levels of contents such thread, posts, feedback, etc. This helps in analysing our technique over forums to maximize the efficiency of searching through its contents and user sentimental. Vidal et al. [24] proposed a 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 preselected sample target page. Guo et al. [15] and Li et al. [18] are similar to our work.
However, Guo et al. did not mention how to discover and traverse URLs. Li et al. developed some heuristic rules to discover URLs. However, their rules are too specific and can only be applied to specific forums powered by the particular software package in which the heuristics were conceived.



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

Threads in Forum : Linux

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 M Black		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 sup		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 »

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.

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

Fig.1 A sample forum board with various threads and pages. Each owns specific URLs which are shown using the colour patterns.

[19] Were it helps in learning to avoid duplication of URLs. This helps a lot while analysing the forums by means of URLs and to avoid the duplication of posts done by users and also multi-comments of users. By which the exact sentimental of user gets identified and also the efficiency over search process gets improvised. Fig.1 helps in identifying the difference between the thread URLs and the page flipping URLs. A perfect example for this is Google search engine. While we search through information we get many links per page (Thread URL) and also with a series of pages (Page-flip URL). References [13] and [14] acts a steeping guide to for analysing the user sentiment (sentiment mining) this acts on a major role over identifying the nature and satisfaction of users, Based on that the forums get ranked. And the efficiency over search is achieved. The process is maintained for periodic time so as the interest of users over threads gets varied by means of the current trends or process undergone.

III. URL DETECTION

It is an existing technique used for crawling web forums based on URLs. Based on the URLs are analysed by the web forum crawler which are used to rank forums. According to this method, it is highly efficient over Generic search methods as it takes lot of time and lacks in efficiency. In which it uses of ITF regexes, which adopts a two step supervised training procedure. The first step is training sets construction. The second step is regexes learning. Based on the regexes we can analyse the pattern for forums using the URLs. Already we have seen the types of URLs that are prioritized while crawling over forums. Web crawler crawls over forums based on their URLs, how much they get related with the content we given.

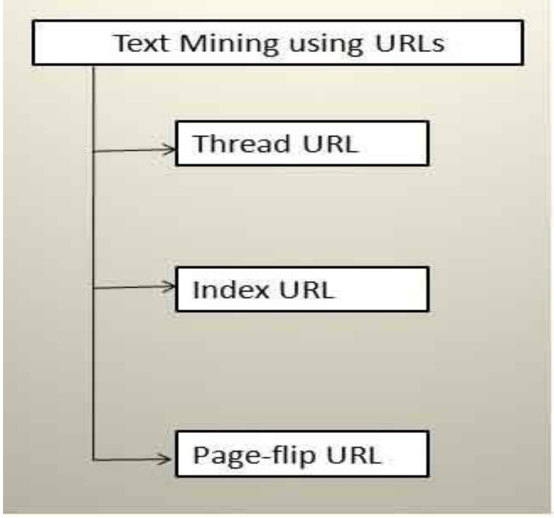


Fig.2 A Diagram for flow of control through the process execution by means of URLs.

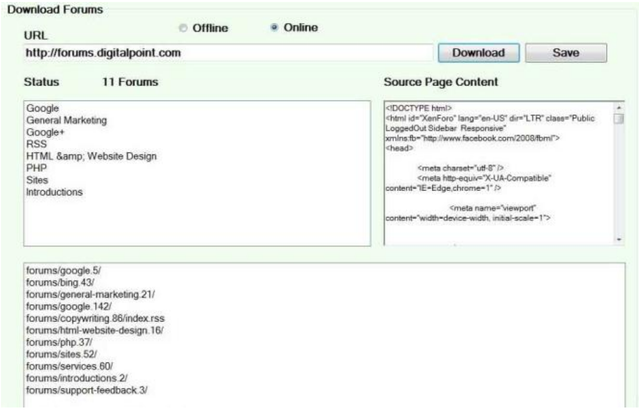
By which when a forum gets referred, each and every thread and its post are analysed. Fig.2 the URL training sets are:

A. *Index URL And Thread URL Training Sets:*

Recall that an index URL is a URL that is on an entry or index page; its destination page is another index page; its anchor text is the board title of its destination page. A thread URL is a URL that is on an index page; its destination page is a thread page; its anchor text is the thread title of its destination page.

B. *Page-Flipping URL 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. [26] 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



URL http://forums.digitalpoint.com

Status 11 Forums

Google
General Marketing
Google+
RSS
HTML &amp; Website Design
PHP
Sites
Introductions

Source Page Content

```
<!DOCTYPE html>
<html id="XenForo" lang="en-US" dir="LTR" class="Public
LoggedOut Sidebar  Responsive"
xmlns:fb="http://www.facebook.com/2008/fbml">
<head>

        <meta charset="utf-8" />
        <meta http-equiv="X-UA-Compatible"
content="IE=Edge,chrome=1" />

                <meta name="viewport"
content="width=device-width, initial-scale=1">
```

forums/google.5/
forums/bing.43/
forums/general-marketing.21/
forums/google.142/
forums/copywriting.86/index.rss
forums/html-website-design.16/
forums/php.37/
forums/sites.52/
forums/services.60/
forums/introductions.2/
forums/support-feedback.3/

Fig.3 shows the extraction process of URLs in the forums by which the crawler performs.

By means of the training sets, the forums containing URLs are identified based on that the URLs are used for crawling over web forums.

Algorithm 5 Generalize Pairwise Rules

```
Input: Pair-wise Rules: R = {< c, t >}
Output: Generalized Rules: Rgen = {< cgen, tgen >}
Class ⇐ {t}; keySet ⇐ K; Nodes ⇐ Class
while keySet is not empty do
    ∀key ∈ keySet InfoGain(key) ⇐ Entropy(R) − Pv∈ {c(key)} ♯c(key)=v
    keysel ⇐ select key with max InfoGain
    new node set P = ∅
    for all nodes n ∈ Nodes do
        V = {ci(keysel)}; where < ci, tj >∈ {< c, t >} ∧ tj ∈ {Class(n)}
        if |V | > threshold|V | then
            P = P ∪ (< keysel, ∗ >)
        else
            P = P ∪ (< keysel, v >)
        end if
    end for
    merge nodes in P with the same value v
    Nodes ⇐ P
    remove keysel from keySet
end while
{< cgen, t >} ⇐ all paths in the DTree
```

Fig.4 algorithm learned from [16] to avoid the duplication over URLs

Algorithm [4] helps in learning the ways to overcome the URL duplication problem. As the duplication of URL caused a lot of problems, for analysis of the web forums. M. Henzinger [16] proposed the concept foe avoiding the duplication in URLs.

IV. User Sentiments

We describes about a system to overcome the existing crawl systems for web forums. In this method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. [5]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, we learn URL patterns across multiple sites and automatically finds forum entry page given a page from a forum. Experimental results show that web crawler using user sentiment 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 in formativeness measure, and finding a traversal path by a spanning tree algorithm. However, the traversal path selection procedure requires human inspection.

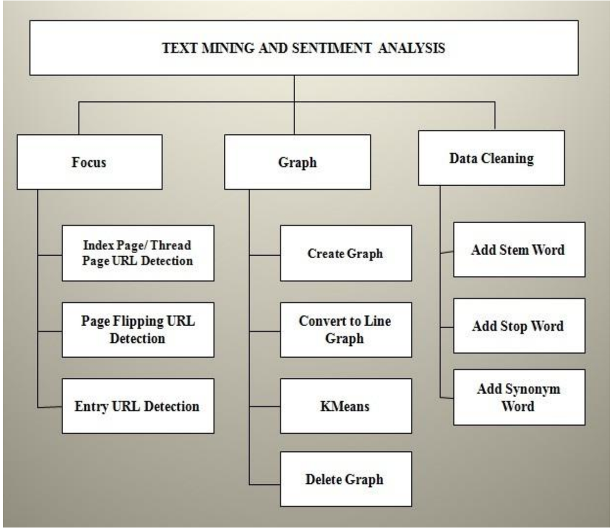


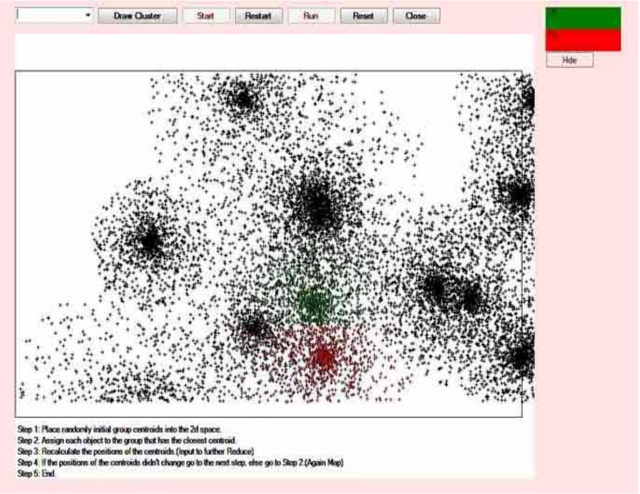
Fig.5 flow chart that represents the process of Web Crawler based on user sentiments.

A. *Learning of collective behaviour*

- The input is network data, labels of some nodes and number of social dimensions; output is labels of unlabeled nodes.
- The following steps are worked out.

1. Convert network into edge-centric view.
2. Edge clustering over the nodes is performed.
3. Construction of social dimensions based on edge partition. A node belongs to one community as long as any of its neighboring edges is in that community.
4. Apply regularization to social dimensions.
5. Construction of classifier based on social dimensions of labeled nodes.
6. Use the classifier to predict labels of unlabeled ones based on their social dimensions.

[6]Based on this the various collective behavior of the user sentimental in viewing the forum is analyzed, which are used for ranking the web forums. Each web forums are provided with an ID so that easy crawling of information's from the forums.



Step 1: Place randomly initial group centroids into the 2d space.
Step 2: Assign each object to the group that has the closest centroid.
Step 3: Recalculate the positions of the centroids.(Input to further Reduce)
Step 4: If the positions of the centroids didn't change go to the next step, else go to Step 2.(Again Map)
Step 5: End.

Fig.6 shows the cluster formation over the view of users in the forums.

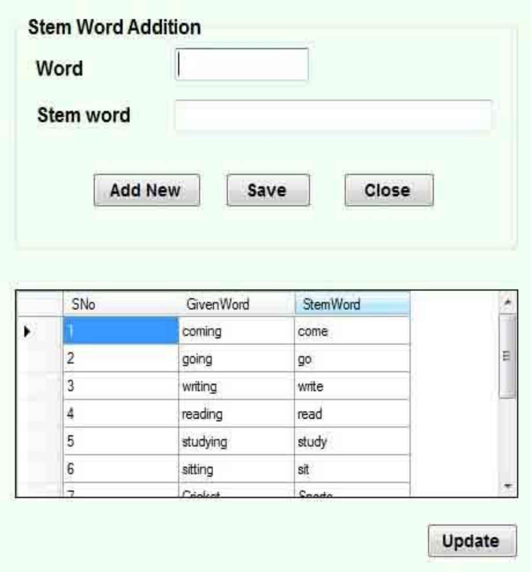
B. Sentiment Analysis:

Mainly, the sentiment analysis over forums depends on the following methods:

1. *Forum Topic Download:*
   Here the source forum page gets analyzed and all its contents are downloaded. The HTML content is displayed in a rich text box control. After the necessary details about forums are gathered, the URLs are parsed and checked for forum links. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control.

2. *Forum Sub Topic Download:*
   All the forum links pages in the source web page are downloaded. The HTML content is displayed in a rich text box control during each page download and parsing of forum sub topic URLs. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control.

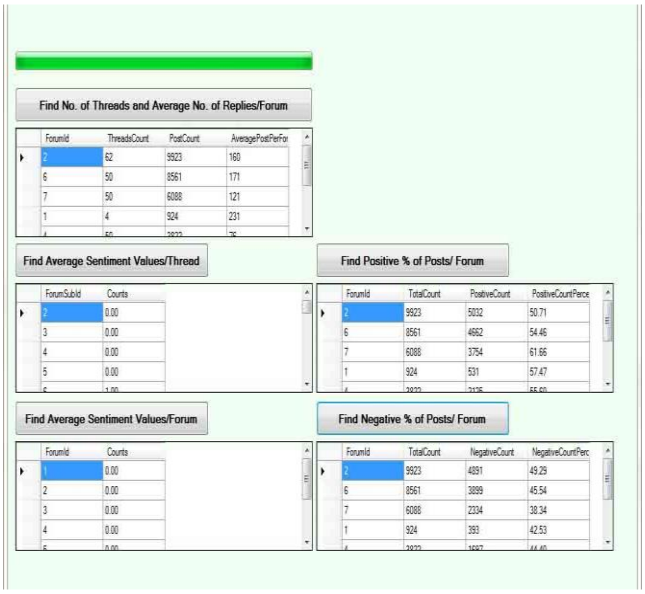
Base on which all the URLs of forums are parsed and the user sentimental are analyzed. This process are done with the help of K-means clustering, used to group of the web forums and threads. This are done by means of Stem word and synonym word comparison over the parsed text.[7]



SNo	GivenWord	StemWord
1	coming	come
2	going	go
3	writing	write
4	reading	read
5	studying	study
6	sitting	sit

Fig.7 It describes the addition of stem word and synonym word which are used for the parsing of user comments.

Based on this the parsing has been performed. And the number of posts are analysed. From which the valid post, positive post, negative post and average post are identified. Using this we can easily rank of the forums with its specific ID.[8]


Find No. of Threads and Average No. of Replies/Forum
Find Average Sentiment Values/Thread
Find Positive % of Posts/ Forum
Find Average Sentiment Values/Forum
Find Negative % of Posts/ Forum

ForumId	ThreadsCount	PostCount	AveragePostPerFo
2	62	9923	160
6	50	8561	171
7	50	6088	121
1	4	924	231

ForumSubId	Counts
2	0.00
3	0.00
4	0.00
5	0.00

ForumId	TotalCount	PositiveCount	PositiveCountPerc
2	9923	5032	50.71
6	8561	4662	54.46
7	6088	3754	61.66
1	924	531	57.47

ForumId	Counts
1	0.00
2	0.00
3	0.00
4	0.00

ForumId	TotalCount	NegativeCount	NegativeCountPerc
2	9923	4891	49.29
6	8561	3899	45.54
7	6088	2334	38.34
1	924	393	42.53

Fig.8 this shows the result of analyzing the positive, negative and average posts in the forums.

V. CONCLUSIONS

Proposed and implemented web crawlers, a supervised forum crawler. We reduced the forum crawling problem to a URL type recognition problem and showed how to leverage implicit navigation paths of forums, i.e. entry-index-thread (EIT) path, and designed methods to learn ITF regexes explicitly. Experimental results on various forum sites each powered by a different forum software package confirm that this web crawler based on user sentimental could effectively learn knowledge of EIT path and ITF regexes. We also showed that web crawler based on user sentimental can effectively apply learned forum crawling knowledge on unseen forums to automatically collect index URL, thread URL, and page-flipping URL string training sets and learn the ITF regexes from the training sets. These learned regexes could be applied directly in online crawling. Training and testing on the basis of forum package makes our experiments manageable and our results applicable to many forum sites. Moreover, it can start from any page of a forum, while all previous works expect an entry page is given. The system uses both qualitative and quantitative accounts of features derived from online posts done by user. The crawler can easily learn methods to analyse the web forums, which helps in searching a variety of information in an efficient manner.

References

[1] Internet Forum, http://en.wikipedia.org/wiki/Internet_forum, 2012.

[2] "Message Boards Statistics," http://www.big-boards.com/ statistics/, 2012.

[3] nofollow, http://en.wikipedia.org/wiki/Nofollow, 2012.

[4] "RFC 1738—Uniform Resource Locators (URL)," http://www.
ietf.org/rfc/rfc1738.txt, 2012.

[5] Session ID, http://en.wikipedia.org/wiki/Session_ID, 2012.

[6] "The Sitemap Protocol," http://sitemaps.org/protocol.php, 2012.

[7] "The Web Robots Pages," http://www.robotstxt.org/, 2012.

[8] "WeblogMatrix," http://www.weblogmatrix.org/, 2012.

[9] S. Brin and L. Page, "The Anatomy of a Large-Scale Hyper textual

[10] Web Search Engine." Computer Networks and ISDN Systems, vol. 30, nos. 1-7, pp. 107-117, 1998.

[11] 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.

[12] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-Duping URLs via Rewrite Rules," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.

[13] C. Gao, L. Wang, C.-Y. Lin, and Y.-I. Song, "Finding Question-Answer Pairs from Online Forums," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 467-474, 2008.

[14] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T.Tomokiyo, "Deriving Marketing Intelligence from Online Discussion," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 419-428, 2005.

[15] 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.

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

[17] 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.

[18] 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.

[19] G.S. Manku, A. Jain, and A.D. Sarma, "Detecting Near-Duplicates for Web Crawling," Proc. 16th Int'l Conf. World Wide Web, pp. 141-150, 2007.

[20] U. Schonfeld and N. Shivakumar, "Sitemaps: Above and Beyond

[21] the Crawl of Duty," Proc. 18th Int'l Conf. World Wide Web, pp. 991-1000, 2009.

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

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

[24] M.L.A. Vidal, A.S. Silva, E.S. Moura, and J.M.B. Cavalcanti, "Structure-Driven Crawler Generation by Example," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 292-299, 2006.

[25] 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. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 459-466, 2008.

[26] 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.

[27] Y. Zhai and B. Liu, "Structured Data Extraction from the Web based on Partial Tree Alignment," IEEE Trans. Knowledge Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.