International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 3 Issue 3 ǁ March 2014 ǁ PP.47-50

Dynamic Strategy for Web Forum Crawling

[1]P.M..Balaganesh, [2]R.Sureshkumar, [3]A.Hemalatha
[1] M.E., (P.HD) DEAN, CSE, Sembodai Rukmani Varatharajan Engineering College, Sembodai
[2]ME., (P.HD) AP CSE Sembodai Rukmani Varatharajan Engineering College, Sembodai
[3]ME CSE. Sembodai Rukmani Varatharajan Engineering College, Sembodai.

ABSTRACT: *Forum Crawler under Supervision (FoCUS), a supervised web-scale forum crawler. The Main goal of this project is to crawl 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 we show how to learn accurate and effective regular expression patterns of implicit navigation paths from automatically created training sets using aggregated results from weak page type classifiers.*

INDEX TERMS: Forum Crawler, URL Types, Page Classification, URL Discovery.

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Also is referred as knowledge discovery in databases. It is a process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

A Web crawler is an Internet boot that systematically browses the World Wide Web, typically for the purpose of Web indexing. A Web crawler may also be called a Web spider or a Web scutter. A forum consists of a tree like directory structure. The top end is "Categories". A forum can be divided into categories for the relevant discussions. Under the categories are sub-forums and these sub-forums can further have more sub forums. Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites web content. Crawlers can validate hyperlinks and HTML code. Web crawlers can copy all the forum pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

II. PROPOSED SYSTEM

Data Mining Algorithms Used to evaluate the web pages. The extension of crawler part i.e. Content mining is done. In this, the presents of Parallel Crawler approach to improve the crawler performance that means Multi- threaded Downloader Supported. Crawler with multi-threaded downloader is responsible for starting threads and obtaining the information about the website being fetched. Multiple processes are run in parallel to perform the above task. Select limited pages to analyze crawler performance.

Classification

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels.
The model is used to classify new objects. Goal of classification is to build structures from examples of past decisions that can be used to make decisions for unseen cases. Predicts the cluster in which a new case fits in the characteristics of the groups can be defined by an expert or fed from historic data.

Page Classification

Based on the web terminology First the forum pages classified into page type as follows, Entry Page: The homepage of a forum, which contains the lowest common ancestor of all threads. Index Page: Major link on the home page. Thread Page: A page of a thread in a forum that contains a list of posts with user generated content belonging to the same discussion. Other Page: A page that is not an entry page, index page, or thread page.

URL Classification

URL type is classified for each page type as follows, Index URL: A URL that is on an entry page or index page and points to an index page. Its anchor text shows the title of its destination board. Thread URL: A URL that is on an index page and points to a thread page. Its anchor text is the title of its destination thread. Page-flipping URL: A URL that leads users to another page of the same board or the same thread. Correctly dealing with page-flipping URLs enables a crawler to download all threads in a large board or all posts in a long thread. Other URL: A URL that is not an index URL, thread URL, or page-flipping URL.

Clustering

Clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the

Classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Pattern Clustering

Content crawling which is the major work of this paper, URL relevant page is downloading to refer their content search from the web source using pattern clustering approach. Grouping the URL's of similar sites and thread URL has the thread i.e. user Posted Content from that stored pages. Check the similarity of web page URL with content using the sequential pattern clustering algorithm.

Advantages of Proposed System

Download rate is maximized and downloading time is minimized while using multi threaded downloader. URL Type is recognized easily by the algorithm specified path. Algorithm Specific Path is EIT-> URL Crawling for Index, Page flipping, Thread Detection. URL crawling by EIT path Specification means, DOM tree Structure. i.e. Anchor text length and their depth value detection (link, hyperlink analysis). So the Page Structure not affected. Static pages are stored in user desired folder to make the local search engine. Any difficulties encountered can be viewed separately in the Error's view. Search Content gives the results faster.

III. RELATED WORK

Architecture Diagram

The crawler working function is explained from the different kind of layers in the system architecture. In Fig 1 The top level layer is for retrieving details of web sources such as WWW or Internet. Which has URL and their data i.e. URI of web pages. The middle layer provides the interaction with both user and web. Download all needed details to user by that crawler launched by the system server.

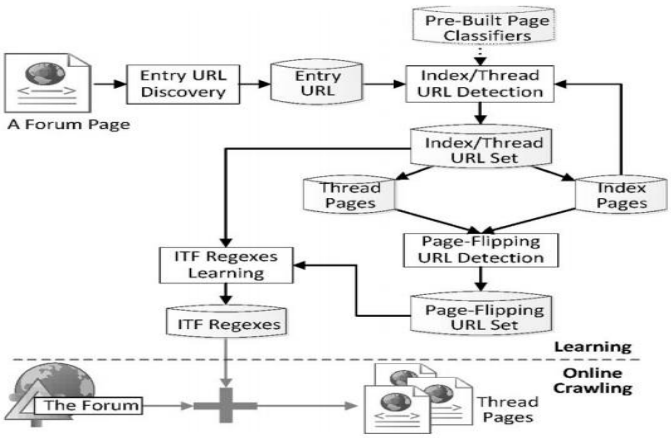

Fig 1 Web Crawler with Multithreaded Downloader

The bottom layer performs the database functions to user as well as server. The user interacts with database to search and retrieve the information and the search engine update their database index based on the crawled details. The following resultant diagram gives the overview of the final working system input, output process and their flow. The dotted circles represent URL's and the straight lines for representing internal process of the system. Finally the external function flow is shown by the dotted lines between search user, database and search engine server.

INTERNET <□ □ CRAWLER < □ □ SERVER< □ □ DB < □ □ USER

The search engine user gives their needed data as thread the resultant URL be the type of thread URL. User content automatically entered that content form the query string behind the URL of a site.

Crawler Design

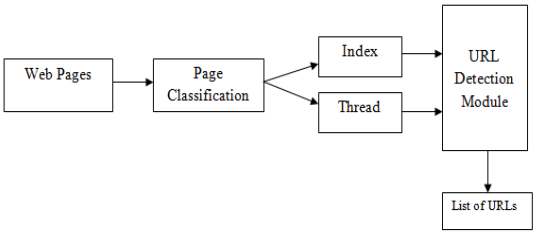

Fig 2 Crawler Design

In Fig 2 Tracking the URI's from URL. Collect varies pages from Internet. Classify the forum page and make Training Set. Store the URLs as single list.

URL Detection using Page-Flipping

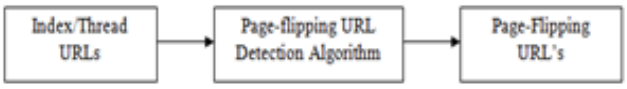

Fig 3 URL Detection using Page-Flipping

Generate the URLs from Index URLs. Select the Group of Page flipping URLs based on the Condition List as follows.

1) MIME Types

In this will set all kinds of data we need to extract from the particular URI like weather we need storing data, Boolean data and images (.jpeg, .png, .gif) information or not.

2) Advanced Settings

These are the settings made by the user in order to restrict some kind of website like with domain name as .NET, .ORG like this.

3) Output Settings

In this we mention the output folder name where we need to store the content about the website fetched.
Two algorithms will be used here.

1. The Index URL and Thread URL Detection Algorithm

let it_group be ψ; data url_group= Collect URL groups by alining HTML DOM tree of sp; foreach ug in url_groups dp ug.anchor_len= total anchor text length in ug; end foreach it_group = arg max (ug.anchor_len) in url_groups; it_group.DstPageType= Majority page type of the destination pages of URLs in ug; if it_group.DstPageType is INDEX_PAGE it_group.UriType = INDEX_URL; else if it_group Dst Page Type is THREAD_PAGE it_group.UrlType= THREAD_URL else it_group= ψ; end if return it_group;

2. The Page Flipping URL Detection Algorithm

Let pf_group be ψ; url-groups=collect URL groups by alining HTML DOM tree sp; foreach ug in url-groups do if the anchor texts of ug are digit strings pages= Download(URL in ug); if pages have the similar layout to sp and ug appears at same location of pages as in sp pf_group=ug; break; end if end if end foreach if pf_group is ψ foreach url in outgoing URLs in sp p=Download (url); pf_url=Extract URL in p at the same location as url in sp; if pf_url exists and pf_url.anchor== url.anchor and pf-url.UrlString!=url.UrlString Add url and cond_url into pf_group; break; end if end foreach end if pf_group.UrlType=PAGE_FLIPPING_URL; return pf_group;

URL Discovery

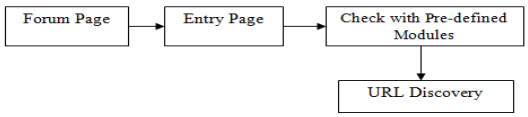

Fig 4 URL Discovery

Fig 4 shows all prior works in forum crawling assume that an entry URL is given. Compare our entry URL with pre-defined module. Manually checked if the output was indeed its entry page. Get all the URI's information and stores them in a queue.

Store the Page as Static

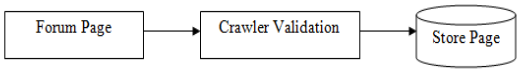

Fig 5 Store the Page as Static

In Fig 5 Check with many annotated pages and find the crawler performance. After the validations were completed then the discovered static pages should be stored in preferred location.

IV. CONCLUSION

In this system no need to consider page score and weights for analyzing the web source pages. The simple design of back end component to the search engine developed in the first part of module implementation. This i the process of URL crawling. The search engine module and their function are discussed. Their detailed functional part is showed from the resultant diagram. This is the process of content crawling.

V. ACKNOWLEDGEMENT

I would like to thank Mr.SureshKumar for give excellent insights and suggestions.

REFERENCES

[1] Agarwal, Chitrapura,K.P. Garg,S. Sasturkar,A. Koppula,H.S. Leela,K.P.(2010)
[2] "Learning URL Patterns for Webpage De- Duplication," Proc. Third ACM Conf. Web Search and Data Mining.
[3] Brin,S. and Page,L.(1998), "The Anatomy of a Large-Scale Hypertextual Web Search Engine." Computer Networks and ISDN Systems.
[4] Cai,R. Lai,W.Wang.Y.Yang,J-M. and Zhang,L. (2008) "iRobot: An Intelligent Crawler for Web Forums," Proc. 17th Int'l Conf. World Wide Web.
[5] Cheng,X.Q. Guo,Y. Li,K. and Zhang,K.(2007) "Crawling Dynamic Web Pages in WWW Forums," Computer Eng.
[6] Cao,Y.B and Lin,C.-Y. Liu,J. Song,X.Y.(2010) "Automatic Extraction of Web Data Records Containing User-Generated Content," Proc. 19th Int'l Conf. Information and Knowledge Management.
[7] Dasgupta,A. Kumar,R and Sasturkar,A.(2008)"De-Duping URLs via Rewrite Rules," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.
[8] Gao,C.Lin,C-Y. and Song,Y-I. Wang,L.(2008) "Finding Question- Answer Pairs from Online Forums," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval.
[9] Glance,N. Hurst,M. Nigam,K. Siegler,M. Stockton,R. And Tomokiyo,T.(2005) "Deriving Marketing Intelligence from Online Discussion," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.
[10] Guo,Y. Li,K. Zhang,K. and Zhan,G.(2006) "Board Forum Crawling: A Web Crawling Method for Web Forum," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence.