Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 8, August 2014, pg.79 – 84

RESEARCH ARTICLE

FoCUS – Forum Crawler Under Supervision

V.Rajapriya
School of Computer Science and Engineering, Bharathidasan University, Tamil Nadu, India
rajpriyavaradharajan@gmail.com

ABSTRACT - ***Forum Crawler Under Supervision (FoCUS) is a supervised web-scale forum crawler. The web contains large data and innumerable websites that are monitored by a tool or program known as crawler. The goal is to crawl relevant forum content from the web with minimal overhead. Forums have different layouts or styles and are powered by different forum software packages. They have similar implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages. It reduces the web forum crawling problem to a URL-type recognition problem. It also shows 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. These type classifiers can be trained and applied to large set of unseen forums. It produces the best effectiveness and addresses the scalability issue and includes the concept called sentimental analysis.***

Keywords: Page classification, URL pattern learning, Sentimental analysis

I. INTRODUCTION

Data mining helps us to extract new information and uncover hidden patterns out of the stored and streaming data and also the process of discovering hidden patterns in large sets. Classification, Regression, Clustering are data mining tasks. Data mining tools predict behavior and future trends allowing business to make proactive, knowledge driven decisions. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases and derives its name from the similarities between searching for valuable information in a large database. Web crawler is a system for bulk downloading of web pages. It crawls relevant content from web with minimal overhead. Web crawlers are used for indexing pages for search engines, archiving the web, analyzing the web etc. Web search engines and other sites use web crawling software to adopt their web content or indexing of other site's web content. It can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping.

II. RELATED WORK

Data mining is used in business field to discover patterns and relationships in the data in order to help make better business decisions. It automates the process of finding predictive information in large database and also uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Web crawler was originally a separate search engine with its own database, and displayed advertising results in separate area of the page and also provides a composite of separately identified filtered results from most popular search engines.

"IROBOT: AN INTELLIGENT CRAWLER FOR WEB FORUMS" by RUI CAI, JIANG-MING YANG, WEI LAI, LEI ZHANG, (2008)

It propose and build prototype of an intelligent forum crawler, iRobot, which has intelligence to understand the content and structure of a forum site, and then decide how to choose traversal paths among different kinds of pages. It's even hard for human being to understand how links are formulated and write down appropriate patterns to distinguish links with different functions. After investigating a substantial number of web forums, found out two observations: (1) The repetitive regions in a forum page can robustly characterize the content layout of that page. (2) The location of a link on a page is important.

"DE-DUPING URLS VIA REWRITE RULES" by ANIRBAN DASGUPTA, RAVI KUMAR, AMIT SASTURKAR (2008)

De-duping URLs is an important problem for search engines, since all the principal functions of search engine, including crawling, indexing, ranking, and presentation are adversely impacted by the presence of duplicate URLs. They provide an efficient algorithm for mining and learning URL rewrite rules and shows that it is complete. Their algorithm learns every URL rewrite rule that is correct, for an appropriate notion of correctness. These duplicate URLs adversely affect the performance of commercial search engine in various ways. They present a different approach to URL de-duping problem based on automatically generating URL rewrite rules by mining a given collection of URLs with content-similarity information. These rewrite rules can be applied to eliminate duplicates among URLs that are encountered for first time during crawling even without fetching their content.

"LEARNING URL PATTERNS FOR WEBPAGE DE-DUPLICATION" BY HEMA SWETHA KOPPULA, KRISHNA P. LEELA, AMIT AGARWAL (2010)

It present a set of techniques to mine rules from URLs and utilize these rules for de-duplication using just URL strings without fetching the content explicitly. The technique is composed of mining the crawl logs and utilizing clusters of similar pages to extract transformation rules, which are used to normalize URLs belonging to each cluster. They present a machine learning technique to generalize the set of rules which reduces the resource footprint to be usable at web-scale. The rule extraction techniques are robust against web-site specific URL conventions. It compares the precision and scalability of approach with recent efforts in using URLs for de-duplication. It includes false positives due to the approximate similarity measures and explore ways of handling these in robust fashion. Generalization is performed separately for source and target and explores the feasibility of generalizing both in an iterative fashion.

"EXTRACTING AND RANKING PRODUCT FEATURES IN OPINION DOCUMENTS" By L.ZHANG, B.LIU, S.H.LIM, E.O'BRIEN-STRAIN (2010)

An important task of opinion mining is to extract people's opinions on features of an entity. Double propagation is a state of the art technique for solving the problem. It works well for medium size corpora. The problem is formulated as a bipartite graph and well-known web page ranking algorithm HITS is used to find important features and rank them high. Feature ranking is applied to the extracted feature candidates to improve the precision of the top ranked candidates. It first uses part-whole and "no" patterns to increase recall. It then ranks the extracted feature candidates by feature importance, which is determined by feature relevance and frequency. Many phrases represent type of semantic relation. The web page ranking algorithm HITS was applying to compute feature relevance. They also plan to study the problem of extracting features that are verbs or verb phrases.

III. PROPOSED WORK

In this paper, propose sparse social dimensions which can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods. It includes online forums hotspot detection and forecast using sentiment analysis and text mining approaches. This is developed in two stages: emotional polarity computation and integrated sentiment analysis based on K-means clustering. The unsupervised text-mining approach is used to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span. Index URL, Thread URL, Page flipping URL algorithms is used.

$$Effectivenes = \frac{\text{crawled threads}}{\text{crawled threads} + \text{other pages} * 100\%}$$

Index URL, Thread URL and Page flipping URL algorithm are used. It consists of two major 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. Given any page of a forum, FoCUS first finds its entry URL using the Entry URL Discovery module.
Once the learning is finished, FoCUS performs online crawling starting from the entry URL, FoCUS follows all URLs matched with any learned ITF regex. FoCUS continues to crawl until no page could be retrieved or other condition is satisfied. To learn ITF regexes, FoCUS adopts a two-step supervised training procedure. The first step is training sets construction. The second step is regexes learning.

1. INDEX URL AND THREAD URL TRAINING SETS

The homepage of a forum which contains a list of boards is also the lowest common ancestor of all threads. A page of a board in a forum, which usually contains a table-like structure, each row in it contains information of a board or a thread. The 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. The only way to distinguish index URLs from thread URLs is the type of their destination pages. Therefore user needs a method to decide the page type of a destination page.

```
Algorithm IndexUrlAndThreadUrlDetection
Input:   sp: an entry page or index page
Output:  it_group: a group of index/thread URLs
1:  let it_group be φ;data
2:  url_groups = Collect URL groups by aligning HTML
    DOM tree of sp;
3:  foreach ug in url_groups do
4:    ug.anchor_len = Total anchor text length in ug;
5:  end foreach
6:  it_group = arg max( ug.anchor_len ) in url_groups;
7:  it_group.DstPageType = Majority page type of the des-
    tination pages of URLs in ug;
8:  if it_group.DstPageType is INDEX_PAGE
9:    it_group.UrlType = INDEX_URL;
10: else if it_group.DstPageType is THREAD_PAGE
11:   it_group.UrlType = THREAD_URL;
12: else
13:   it_group = φ;
14: end if
15: return it_group;
```

2. 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. The proposed metric is used to distinguish page-flipping URLs from other loop-back URLs. However, the metric only works well on the "grouped" page-flipping URLs more than one page-flipping URL in one page.


```
Algorithm PageFlippingUrlDetection
Input:    sp: an index page or thread page
Output:   pf_group: a group of page-flipping URLs
 1: let pf_group be φ;
 2: url_groups = Collect URL groups by aligning HTML
    DOM tree of sp;
 3: foreach ug 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: end foreach
12: if pf_group is φ
13:    foreach url in outging URLs in sp
14:       p = Download( url );
15:       pf_url = Extract URL in p at the same location as
                   url in sp;
16:       if pf_url exists and pf_url.anchor == url.anchor
            and pf_url.UrlString != url.UrlString
17:          Add url and cand_url into pf_group;
18:          break;
19:       end if
20:    end foreach
21: end if
22: pf_group.UrlType = PAGE_FLIPPING_URL;
23: return pf_group;
```

3. ALGORITHM OF SCALABLE K-MEANS VARIANT

The data instances are given as input along with number of clusters, and clusters are retrieved as output. First it is required to construct a mapping from features to instances. Then cluster centroids are initialized. Then maximum similarity is given and looping is worked out. When the change is objective value falls above the 'Epsilon' value then the loop is terminated.

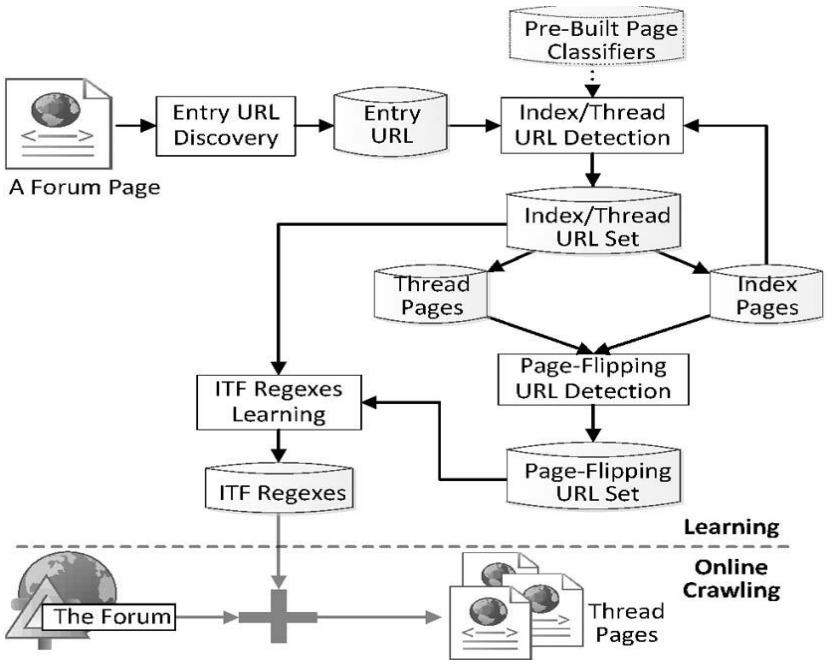
IV. IMPLEMENTATION

In this paper, Sentimental analysis method is used. It determines the hotspots based on the comments and divides into positive, neutral and negative. These keywords also divides themselves based on comments. Hotspots means referring maximum number of counts in URL link values. Hotspots are selected based on the keywords. To find out the hotspots these steps followed:

1. FORUM TOPIC DOWNLOAD: The source web page is keyed in the website and the content is being downloaded. The HTML content is displayed in a rich text box control.
2. PARSE FORUM TOPIC TEXT AND URL'S: The downloaded source page web content is parsed and checked for forum links. The links are extracted and displayed in a list box control. The link text are extracted and displayed in another list box control.
3. 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.
4. PARSE FORUM SUB TOPIC TEXT AND URLS: The downloaded forum pages web content are parsed and checked for sub forum links. These links are extracted and displayed in a list box control. The link text are extracted and displayed in another list box control.

The proposed system automatically analyzes the emotional polarity of a text, based on a value for each piece of text is obtained. The absolute value of the text represents the influential power and the sign of the text denotes its sentiment values. The K-means clustering is applied to group the posts and its replies. One hundred and forty two forums are extracted from forums.digitialpoint.com Six thousand six hundred and fifty four threads are spread among those forums Posts are found to be valid only if they contain minimum three words. Posts are grouped and

average value taken for each thread, likewise threads average sentiment values are taken and forum's average value is derived.



The architecture of FoCUS

It is found that more of the forums sentiments values are having more positive percent and less negative percent. The proposed methodology efficiently analyzes their sentiments. An incomparable advantage of the proposed model is that it easily scales to handle networks with millions of posts. Since the proposed model is sensitive to the number of social dimensions as shown in the experiment, further research is needed to determine a suitable dimensionality automatically.

V. CONCLUSION

The algorithms are developed to automatically analyze the emotional polarity of the text, based on which a value for each piece of text is obtained. The absolute value of the text represents the influential power and sign of the text denotes its emotional polarity. Using the hotspot predicting approaches can help the education institutions understand what their specific customer timely concerns. Results generated from the approach can also be combined to competitor analysis to yield comprehensive decision support information. The new system is designed such that enhancements can be integrated with current modules easily with less integration work. This becomes useful if the above enhancements are made in future. It will be very useful for people to know multiple comments simultaneously and can share the views. The application if developed as website can be used from anywhere. It is designed such that those enhancements can be integrated with current modules easily with less integration work.

Advantages:

- Index page, Thread page and Page Flipping URL are identified as well as forum post contents are also extracted.
- Sparsifying social dimensions can be effective in eliminating the scalability bottleneck.
- K-Means clustering approach classifies the forums in to related groups.
- Posts are also clustered to find the number of items belongs to the individual clusters.

REFERENCES

[1]. "iRobot: An Intelligent Crawler for Web Forums" by R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, Proc. 17th Int'l Conf. World Wide Web, pp. 447-456, 2008.
[2]. "De-Duping URLs via Rewrite Rules" by A. Dasgupta, R. Kumar, and A. Sasturkar, Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 186-194, 2008.
[3]. "Learning URL Patterns for Webpage De-Duplication" by H.S. Koppula, K.P. Leela, A. Agarwal, K.P. Chitrapura, S. Garg and A. Sasturkar, Proc. Third ACM Conf. Web Search and Data Mining, pp. 381-390, 2010.
[4]. "Extracting and Ranking Product Features in Opinion Documents" by L. Zhang, B. Liu, S.H. Lim, and E. O'Brien-Strain, Proc. 23rd Int'l Conf. Computational Linguistics, pp. 1462-1470, 2010.
[5]. "Automatic Extraction of Web Data Records Containing User-Generated Content" by X.Y. Song, J. Liu, Y.B. Cao, and C.-Y. Lin. In *Proc. of 19th CIKM*, pages 39-48, 2010.
[6]."FoCUS: Learning to Crawl Web Forums" by Jingtian Jiang, Xinying Song, Nenghai Yu and Chin-Yew Lin, 2013 "Proc. IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6.
[7]." Web Crawler" by Raja Iswary, Keshab Nath, October 2013, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2.
[8]."Automated Path Ascend Locomotion – A Prescribed Topical Crawl Method" by P.Senthil, S.Pothumani, T.Nalini, IJARCSSE, Issue 2, Vol 3, february 2013.
[9]."Web Forum Crawling Techniques" by Namrata H.S Bamrah, B.S Satpute, Pramod Patil, International journal of computer Applications, No 17, Vol 85, January 2014.

