International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 3 ǁ March. 2013 ǁ PP.32-36

An Archetype for Web Mining with Enhanced Topical Crawler and Link-Spam Trapper

* N.Kabilan

**Student, BE Computer Science & Engineering*
Coimbatore Institute Of Engineering And Technology (C.I.E.T)

ABSTRACT: *Search engines are the entrance for today's web. Probably most of us get through the information's with the help of it. Seeking and rendering information from web through the search engine gives more pressure to retain the credibility and reliability of the search engine. Spam was one that creates a severe pain for the reliability of the search engines. Spam's occurrence in the search engine deceives the ranking algorithm to attain higher ranks rather for commercial purposes or to seek attention. Crawlers are the programs which facilitate the search engine's work by following the hyperlinks in Web pages to automatically download a partial snapshot of the Web. Crawling is the initial and also the most important step during the web searching procedure. With the rapid growth of information and the explosion of web pages in the World Wide Web, it gets harder for traditional crawlers to retrieve the information relevant to a user. Here in this paper a prototype is explored which would pay attention on focused results to imply the credibility of the results through two fold validation. Initially result is retrieved by paying attention to focused topical crawling. This was done by focused crawler and the result set will have higher relevancy and after that the spam validator to verify the link spams is applied to avoid the spam sites. Hope this strategy will overcome the Link farms and give the higher relevant document which user wish to look ahead.*

Keywords: *Crawler, Mining, Topical crawler, Search Engine, Link Spam Trapper*

I. INTRODUCTION

A Search engine Spam could be defined as the use of some purposely designed mechanisms to raise the ranking of web sites or pages in the search engine results. Crawler is the heart of a search engine and it is a program that retrieves Web pages or a Web cache. A web crawler, which is a central component of a search engine, impacts not only on the recall ratio and precision of a search engine, but also on the capacity of the data storage and efficiency of a search engine. The data present in the web exceeds day by day. The most popular search engine google provides only the index to 40% of the web. So in order to improve the result reliability and prominence, first focus should be implied on topical crawler to restrict the domain concepts much focused and to get reliable results as a core. After that the algorithm to detect the link spam occurance is applied to the result set obtained from topical crawler.

II. TOPICAL CRAWLER

A web crawler is a program, which browses the World Wide Web in some procedure implied on it. Web crawlers are mainly used for automating maintenance tasks by scratching information automatically. Typically, a crawler begins with a set of given Web pages, called seeds, and follows all the hyperlinks it encounters along the way, General crawlers insert the URLs into a tree diagram and visit them in a breadth-first manner. There has been some recent academic interest in new types of crawling techniques, such as focused crawling based on semantic web to eventually traverse the entire Web

A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics. Topical crawling generally assumes that only the topic is given, while focused crawling also assumes that some labeled examples of relevant and not relevant pages are available. Topical crawling was first introduced by Menczer [2]. Focused crawling was first introduced by Chakrabarti et al [7]. Topical crawler only downloads the links which were pertinent to the given query so some sort of computational predictions is required to decide whether the page is relevant or not. Significance may also be given to anchor texts or contents or the meta data to assess that relevancy. Topical crawling depends on search engine for the initiatives.

III. SPAM TAXONOMY

Spamming techniques can be divided into three categories: content based spam, link based spam and page-hiding based spam.

• Content based spam changes the textual content of web pages to be ranked higher.Some popular content based spam techniques include repeating keywords, using unrelated keywords and adding an entire dictionary at the bottom of a web page.

• link based spam changes the link structure of web pages to attack search engines using link based ranking algorithms, such as PageRank and HITS. The most well-known link based spamming techniques include building link farms, link exchanges, link bombs and adding comment spam in blogs.

• Page-hiding based spam hides the whole or part of a web page to search engines to achieve better ranking for the page. Cloaking and deliberate use of redirect are well-known examples of page-hiding based spamming techniques

IV. GETTING INTO THE CORE – LINK SPAM

Link spam is the manipulation of the link structure or anchor text among pages to get a higher rank. Some iinstances of link spams are

Link farms:

Each page has a collection of links that point to almost every other page. Some ranking algorithms may give pages a higher rank if there are many other pages linking to them. So, by building the link farm, each page will have a higher rank and a new targeted page pointed to by some pages from the link farms will get a higher value too.

Link exchange:

Web site owners promise to add a link to your site as long as you put a link to their sites, regardless of whether these sites are on a similar topic. Usually the owners will explicitly show this intention on their web pages, or they may send emails to other site owners requesting a link exchange.

Link bombing spam:

The anchor text of a link misdescribes the target page. Some search engine will judge the content of a page by the anchor text of the links that point to the page. So because of link bombing spam, the target page will be ranked high to some unrelated query words that just appear on the anchor text of links pointing to it. This link bombing spam is often called Google bombing.

Comment or blog spam:

Spammers add trash links in blogs or wiki systems. Because some blogs or wiki web sites have good reputation, spammers want to get benefit from generating this comment spam.

Combating Link spam:

Link spam contains a collection of links that point to almost every other page. Some ranking algorithms may give pages a higher rank when they possess many other sites link to it. Gy¨ongyi and Garcia-Molina discovered the optimal link structure to boost a certain target page when building a link farm . Wu and Davison show that link farms can form bipartite subgraphs in the document-link matrix when considering anchor texts as part of the link. Some statistical methods predict the unusual link structure of the websites in accordance with that we may find the odd man out of the series.

V. OPERATIONAL SCENARIO OF CRAWLER

Crawler fetches the content in the following scenario. We don't have any catalog of all accessible URLs on the web. The only way to collect URLs is to scan collected pages for hyperlinks to other pages that have not been collected yet. This is the basic principle of crawlers. They start from a given set of URLs, progressively fetch and scan them for new URLs (out links), and then fetch these pages in turn, in an end less cycle. New URLs found thus represent potentially pending work for the crawler. The set of pending work expands quickly as the crawl proceeds, and implementers prefer to write this data to disk to relieve main memory as well as guard against data loss in the event of a crawler crash. There is no guarantee that all accessible web pages will be located in this fashion; indeed, the crawler may never halt, as pages will be added continually even as it is running. Apart from out links, pages contain text; this is submitted to a text indexing system to enable information retrieval using keyword searches.

VI. DESIRED PROTOTYPE

According to prediction in this paper, a crawler which was facilitated with the facilities of crawling over a topic should be created and after that a spam validator is to be attached. Now a variation of the algorithm proposed by Tan Su Tung, Nor Adnan Yahaya[1] to detect link farm and Li Wei-jiang,Ru Hua-suo, Zhao Tie-

jun,Zang Wen-mao [2] for focusing on topical crawling is integrated here for the two fold validation. Infusion of these two algorithms to create a new prototype which may offer more credible results is prescribed.

Algorithm 1 : Link_Spam Trapper (Spam Validator)

Let x denote the URL of the web page and d[x] the domain name of x.

IN(d[x]) denotes the set of incoming links to the domain x.

OUT_Temp denotes the outgoing links from a temporary node temp.

Specify the parsing depth Par_Dep it may be predetermined number of levels whichour algorithm has to parse.

Specify one more parameter Thr_indicator, to define x if it is bad. Perform

Step 1: For each URL I in IN(d[x]),

Check if d[i]!=d[x] and d[i] is not in domain list of x, then add d[i] to the set of Incoming nodes, INLIST(x)

Step 2: Set p as temp and set current level LEV to 0

Step 3: If level<Par_Dep then

For each URL k in OUT_Temp and execute following steps

a. If d[k] != d[p] and d[k] is not in OUTLIST(x), then add d[k] to the set of OUTLIST(x).

b. Else If d[k] == d[x], then set LEV ++, set k as temp and repeat step 3.

Step 4: Calculate the intersection of INLIST(x) and OUTLIST(x), If the number of elements in the intersection set is equal to or bigger than the threshold Thr_indicator, mark x as a bad page.

Step 5: Repeat 1 ,2, 3 and 4 for every search result URL,x.

Algorithm 2 : Topical_Crawling Enabled with Link_Spam Detection

After omitting the spam pages from our search index by using the above algorithm, we were now focusing on the relevant results which may increase the credibility of the search results .

Let threshold of the web page content be T_Main and threshold of relevant of text of linkage in that page be T_Link and threshold of count of crawling pages be T_Crawl

Step 1: while (queue of linkage is not null) ∧ (amount of crawling pages < T_Crawl) do

Step 2: Get the linkage at the head of queue and downloading web page P the linkage linked and calculate the relevant to topic T;

relevance(P) = similarity(P, T_Main) if relevance(P) < T_Link then goto step 3

Step 3: Dismiss Page P and all of linkages in this page; Stop the process

Step 4: for each linkages a in the page P do

score a as follow:

a. relevance(a) = similarity(a, T)
b. if relevance(a) < T2 then
c. dismiss a;

Goto step 4.

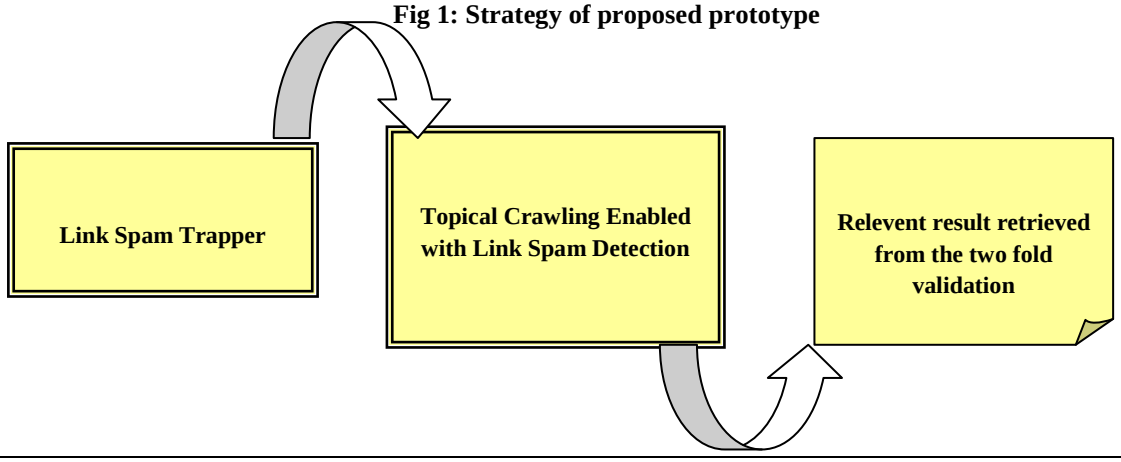
Step 5: if the linkage a has not been crawled then add linkage a into the queue of linkage

Step 6: Stop the process.

These two algorithms has to be integrated. The working strategy of the combined algorithms has been given in figure 2. In that initially algorithm 1 will get evaluated and after that algorithm 2 will be evaluated.

In figure 2 the working scenario with the sequence of steps has been given and this will imply the places **where the algorithms have been utilized.**

Fig 1: Strategy of proposed prototype



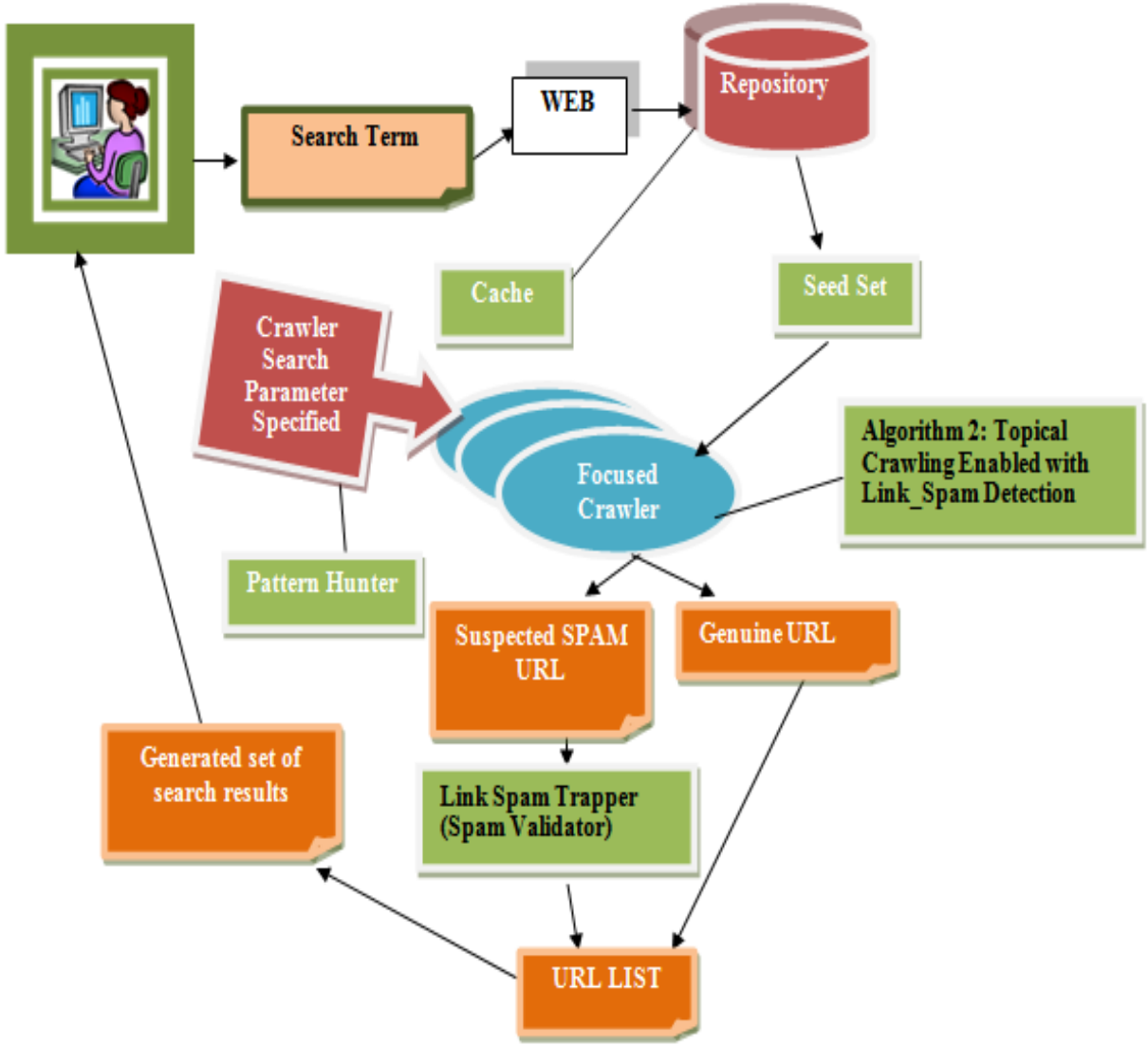


Figure 2: Working scenario of the proposed prototype

V. EVALUATION METRICS

To test the performance, different evaluation metrics have been proposed by researchers for their anti-spam approaches.

Usually two metrics are used to judge the performance of anti-spam approaches.

• Use precision and/or recall when pinpointing spam to demonstrate the performance.

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

• Use the improvement in search quality to demonstrate the performance.

Since prototype is now devised, the implementation of this scenario is in progress, to incorporate the information retrieval with a maximum reliability and thus the result and comparison analysis cannot be revealed.

VI. CONCLUSION

The war between spammers and hammers has been going for a long while. Since to make the information retrieval more reliable this proposal is one among the journers to develop the criteria to overcome the link spam and relevant result.

One more feature that is to included in this prototype is to have Multiclass Support vector machine incorporated to classify various taxonomies of the spammies. Hope soon a better warrior to give reliable information on the web may arise.

REFERENCES

[1]. Tan Su Tung, Nor Adnan Yahaya, " Multi-level Link Structure Analysis Technique for Detecting Link Farm Spam Pages", Proceedings of the 2006 IEEE/WIC/ACM International Conference on
[2]. Web Intelligence and Intelligent Agent Technology
[3]. Li Wei-jiang,Ru Hua-suo, Zhao Tie-jun,Zang Wen-mao ," A New Algorithm of Topical Crawler", in the proc 2009 Second International Workshop on Computer Science and Engineering
[4]. X. Ni, G.-R. Xue, X. Ling, Y. Yu, and Q. Yang, "Exploring in the weblog space by detecting informative and affective articles," in WWW '07: Proceedings of the 16th international conference on
[5]. World Wide Web.
[6]. Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen, "Model bloggers' interests based on forgetting mechanism," in WWW '08: Proceeding of the 17th international conference on World
[7]. Wide Web.
[8]. K. Ishida, "Extracting spam blogs with co-citation clusters," in WWW '08: Proceeding of the 17th international conference on World
[9]. Wide Web.
[10]. Dongfei Liu Jia Liu, "PPSpider: Towards an Efficient and Robust Topic-Specific Crawler Based on Peer-to- Peer Network", in : Proceeding of the 2009 Second International Workshop on Computer Science and Engineering.
[11]. S. Chakrabarti , M. Vandenberg , B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," Proc. of the eighth international conference on World Wide WebMay 1999, Toronto, Canada.
[12]. H. Liu, E. Milios, J. Janssen. "Focused crawling by learning HMM from user's topic-specific browsing." In Proc. of the 2004, IEEE/WIC/ACM International Conference on Web Intelligence.