Design and Implementation of Focused Web Crawler Using Genetic Algorithm

Pranali Kale  
Student  
Computer Science and Engineering  
Rajiv Gandhi College of Engineering and Research  
Nagpur, India  
pranalikalea1@gmail.com

Prashant Dahiwale  
Associate Professor  
Computer Science and Engineering  
Rajiv Gandhi College of Engineering and Research  
Nagpur, India  
prashant.dahiwale@gmail.com

Abstract-- In recent years search engine contains a big data on the internet. And the rapid growth of the internet makes it very difficult for general purpose search engines, like google, bing etc., to extract most of the similar results in response to the client queries, To attempt this issue the Focused web crawlers are emerging. The crawler is kept focused to the user interests toward the topic. This paper proposes to design the Focused web crawling using genetic algorithm. The genetic algorithm is manage to optimize Web crawling and to choose more proper Web pages to be obtained by the crawler. In this project the overall working of the focused web crawling using genetic algorithm will be implementing. In genetic algorithm uses the jaccard, and data function technique, which is use to determine the relevancy of the web pages.

Keywords-- Crawling, focused crawling, Genetic Algorithm, web crawler

I. INTRODUCTION

In these days of the spirited world, where each and the every second is measured the important backed up by the information. Due to the huge amount of data on the world wide web and various user different requirements, so information retrieval becomes a challenge. When the data is searched, then the number of results are obtained. The user do not have the perseverance and extend to go through the each and the every page which is listed. Therefore search engines have the biggest job of give the results categorization, in the arrange of interestingness of the user inside it look the first page and the fast summary of the information provided on the page. For the Web crawler, the most significant fragment of the search engines, penetrating through several records and to pick the matching ones is a boring task. Furthermore, the network, which hold more than 11 million pages even still rising and varying quickly. The Crawler is the most vital part in the search engine. It can negotiate the Web space by following Web page’s hyperlinks and storing the downloaded Web documents in local repositories that will later be indexed and used to respond to the client’s queries efficiently [4]. Focused crawling was first introduced by S. Chakrabarti Focused Crawler is kind of crawlers try to find high-quality information on a specific subject as soon as possible and try to avoid irrelevant pages in order to the results would be as accurate as possible. Thus, a focused crawler is a program which extracted as much as possible relevant pages. The focused crawling can be done by using Genetic Algorithm. The genetic algorithm is a kind of the searching algorithm. It penetrates the result place for
the best possible solution to the tricky Problem. Genetic Algorithm uses the genetic operators which are selection, crossover and mutation. It produces the result for the succeeding generations. The process of genetic algorithm ended when the best possible solution is found. Applying the genetic algorithm for the web crawlers is possibly to produce a excellent outcome.

II. RELATED WORK

In paper [1] gives the details of crawling algorithms it compares all the algorithm and give the advantages and disadvantages of every algorithm In paper [2], it uses the different probabilities for the particular input and it gives good result by using the different mutation rates. The techniques presented in this paper are applicable to any domain for which it is possible to generate term-based characterizations of a topic. It gives the total description about mutation rate. In paper [3] the Sushil Kumar et.al proposed the context model for the focused web search. The paper [4], depict a Focused Crawler which look for gain, make the index, and keep the collection of the pages on a particular of area that represent a somewhat thin portion of the web. Thus, web substance can be handled by a scattered group of the focused web crawlers, each concentrating in one or a small number of area. The focused crawler is directed by a category which discovers to be familiar with the relevance from the examples surrounded in a particular topic, Classification, and a distiller which discover relevant vantage points on the World Wide Web. In paper [5] says that it uses to grouping of the link structure analysis and the subject matter similarity while building their focused crawling. The idea behind that it is based on the ordinary hyperlinks in pages are illustration to the authors sight about additional pages. Also the matter of the pages are the another source to relate them to the domain. In paper[6] combines the search strategy based on the content and the link construction. The study of the Link is based on the anchor score, close relative score etc. In paper [7] S. N. Sivanandam, S. N. Deepa describe the whole genetic algorithm and gives the detail of genetic operators and working of the genetic algorithm. In the paper[8] Hati, D. was proposed block partitioning technique in which the blocks are partitioned by using VIPS algorithm In paper [9] is named as the Genetic Algorithm: A tutorial review it represent approximately all conservative technique search from a particular point. Genetic Algorithms all the times manage on a entire population. These add a lot of the toughness of the genetic algorithms. It decreases the risk of proper attentive in a local fixed place. Jon M. Kleinberg[10] proposes the notion of authority on the basis of the algorithmic formulation, which is based on the relationship between the collection of relevant reliable pages and a collection of center pages that bond them collectively in the link structure. In paper [11] , named as the Context Focused Crawler(CFC), it manages the partial ability of the search engines like Google. It is used to permit the users to question for pages connecting to a particular document. This data can be used to create a statements of the pages that happen within the accurate link space of the goal documents. In paper [12] proposed the new hybrid approach to the focused crawling based on the meta search algorithm The paper[13] it extend the performance of the focused web crawling by using the Gcrawler technique. The paper [14], obtains the real mixture of subject-based and link-based Web analysis, concurrently with the capacity to make the universal searching. In paper [15] it proposed the OFC which is based on the reinforcement learning and the fuzzy clustering theory for a focused crawler. In paper [16] Safran et.al proposed the new learning based approach
to make better relevance forecast in focused web crawler. Initially, instruction set is construct to direct the system. Instruction set contain the amount of four relevance attributes: URL word relevancy, anchor text relevancy, parent page relevancy, and surrounding text relevancy. By using Naïve Bayesians, which is used to guess the relevancy of unvisited link. In paper [17] says that it proposed the precedence based focused crawling. The web page consequent to the URL are downloaded from the web and it measures the relevant value of the download page with the center word.

III. PROPOSED WORK

The main aim of the project is, to choose the most promising links in order and try to maximize the relevancy of a new, unvisited URL. The main aim of the project is to apply genetic algorithm for focused web crawling, by using this algorithm we will try to produce more optimal result, and it also helps to improve the accuracy. The main goal of the proposed system is to apply genetic algorithm for focused web crawling. The proposed method yields promising results.

Step1- Give the Input( seed url and query)
Step2-Remove special symbol, stop words, and stem words
   algorithm
Step3- Apply basic web crawling
Step4- Apply genetic algorithm in selection, crossover and mutation phase.
Step5 -Get the output.

A. REMOVE SPECIAL SYMBOL, STOP WORDS AND STEM WORDS

In the special symbol algorithm it will removes the special symbol that are (., @ % & + - / # $ ! * , etc). In the stop word algorithm it will remove the stop words. E.g.-the, are is about all, by etc. In the stem words algorithm it remove the stem words e.g.- summing, interested, like ing, est, ed, these words are removed. By using these three algorithm we get the refined query, that refine query we call that keywords. By using these algorithms we can focused on only original keywords.

B. WEB CRAWLING

The process of web crawler first it start with the seed URL, then crawler start downloading a group of seed pages. Parse through the download page and extract all the links The links to pages that have positioned in a queue. After extraction all the links the procedure is frequent. Crawlers are designed for different purposes In the High performance crawlers, their goal is to improve the working of the crawler by downloading as a number of documents as feasible as in a definite time. By using this process the easiest algorithms is the Breadth First Search (BFS) Algorithm.

In [18][19] the aims of this algorithm is the standardized search from one side to other side of the neighbor nodes. It starts by the source node and find all the neighbor nodes at the parity. If the goal is found, then it is reported as the success and the search is ended. If the goal is not found, then it will go down to the next level far reaching the search from corner to corner the next neighbor nodes at that stage and so on till the objective is not found. When all nodes are searched, but objective is not found then it is reported as the failure. Breadth first is well performed when the objective is found on the upper level in a deeper tree. This strategy gives us more relevant result.
The main goal of the paper [20] is to determine the algorithmic feature of the focused crawler or topical crawlers. A genetic algorithm is a kind of the searching algorithm. It penetrates the resolution place for the optimal result to the difficult Problem. Genetic algorithm is an iterative procedure which represents its applicant result as sequence of the genetic material called as the Chromosomes. When the folks come together then population form. Population is customized in the every iteration. Genetic Algorithm’s process are continuously repeated are called the generation. Genetic Algorithm used the genetic operative such as selection, crossover and mutation. It produces the solutions for the consecutive generations. The genetic algorithm is used to optimize the Web crawling and it select a more proper Web pages to be extracted by the web crawler. Numerous estimation experiments are conducted to determine the advance to the success. Genetic algorithm is used to calculate the relevancy of page by using fitness function. Link is extracted on the basis of fitness function. For fitness function is calculated by using jaccard function. In Genetic algorithm it contains a three process:- selection , crossover and mutation.

Selection—In selection process find out the similarity on web page on the basis of links and keywords. In selection process the formula for finding the score of links by using the jaccard function. Jaccard function similarity

\[ J_{\text{similarity}}(P,R) = \frac{X \cap Y}{X \cup Y} \ldots (1) \]

P is P is a one web page and R is another web page. In web page P it contains a set of links called X and in web page R it contains a set of links called Y. Page P is constant but page R is vary. By using this formula it find out the score of the link by finding the similarity between the two pages on the basis of links.

In selection process formula to find the weight of keywords and data.

\[ D_{\text{pr}} = \sqrt{\frac{M_p + N_r + C_{W\text{pr}}}{U_{W\text{pr}}}} \ldots (2) \]

Dpr is the weighted term, p is a one web page and r is a another web page. Mp is how much time keyword appeared in web page p. Nr is how much time keyword appeared in web page r. CWpr is common words in both web pages(p,r). UWpr is uncommon words in both web pages (p,r).By using these formula we have to find out similarity between the two web pages on the basis of keywords.

Finally for all the web page that has been visited by a web crawler, we have to make the addition for link and keyword score..

\[ J(P,R) = J_{\text{similarity}}(J_{\text{link}}) + D_{\text{pr}}(J_{\text{keyword}}) \]

Crossover—Crossover generally combines two higher fitness value chromosomes, to gain a new offspring. These offspring are passed to the next iteration for further
evaluations. After selecting the fittest individuals, we perform the crossover operator to produce the children of the next generation. In crossover phase it select the main link’s sub-links after the selection process.

**Mutation**—In the mutation phase is goal at provide the crawler capability to search widely various network area properly. Keywords are taken out after applying algorithm that is remove special symbol, stop words, stem word. The chosen keywords run like a query in the famous search engines, that are Google. Their links are http://www.google.com. Get the links from the well-known search engine and pass to the selection phase.

**IV. EXPERIMENTAL RESULT**

Here we have evaluate our crawler with the Simple Focused Crawler on some average Parameters. By using genetic algorithm the parameter are compared i.e. accuracy. The results openly prove that our method in mostly all parameter are comparatively better than the other method. Our main goal of optimization the crawler is accomplish when we manually measure the retrieved links with the other as relevancy is greatly achieved. This is proved in our next table displaying the results obtained dynamically executing our crawler.

Overall performance analysis
Comparing proposed crawler with other technique based crawler we get the following results:

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Type</th>
<th>Query related to</th>
<th>Total relevant link out of 100</th>
<th>Precision in %ile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Focused Web Crawler Using Genetic Algorithm</td>
<td>What is java classes</td>
<td>66</td>
<td>66%</td>
</tr>
<tr>
<td>2</td>
<td>Simple Focused Web Crawler</td>
<td>What is java classes</td>
<td>35</td>
<td>35%</td>
</tr>
</tbody>
</table>

The following table shows the rate of precision for simple focused web crawler and focused web crawler using genetic algorithm

<table>
<thead>
<tr>
<th>Query</th>
<th>Limit for URL to crawls</th>
<th>Simple focused web crawler (Precision) in %</th>
<th>Focused web crawler using genetic algorithm (Precision) in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is java classes</td>
<td>10</td>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>44</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>35</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>27</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>27</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>48</td>
<td>83</td>
<td></td>
</tr>
</tbody>
</table>

Fig.2 shows the analysis of rate of precision from the above table, where precision is percentile ratio of number of relevant links and limit for url to crawl i.e. total links.
V. CONCLUSION

By using genetic algorithm we will get more relevant links in less time. Genetic algorithm is used to find results with more precision. It also helps to improve the crawling performance. The genetic algorithm process terminates when an optimum solution is found. The advantages of this approach is that it may possibly construct certain-area collections with superior quality than fixed focused web crawling methods and will give us more relevant result in less time. The work will comprehensive by making use of some additional similarity coefficients for evaluating the result.

REFERENCES

[16] Mejdi S. Safran and Abdullah Althagafi and DunrenChe "Improving Relevance Prediction
for the Focused Web Crawlers” ,11th International Conferenc on Computer and Information Science, 2012 IEEE/ACIS


