Volume 6, Issue 7, July 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Improving URL Analysis Model for Focused Crawler

K. Varun Kumar Reddy
M.Tech Student, Department of Computer Science and Engineering, GATES Institute of Technology, Gooty, Andhra Pradesh, India

O. Bhaskar
Assistant Professor, Department of Computer Science and Engineering, GATES Institute of Technology, Gooty, Andhra Pradesh, India

Abstract*— proposing paper analyses the URL analysis models of the existing focused crawler, and also their pros and cons, then we propose a URL analysis model based on the improved genetic algorithm, in which the selection operator, crossover operator and mutation operator are optimized. The user query is introduced to construct the virtual documents to participate the genetic process.

***Keywords*— AVG, HITS,**

I. INTRODUCTION

With the explosive growth of the information on the Internet, there are many challenges for the general purpose search engine, such as the size of the index, the speed of the update and personalized needs. Facing these challenges, the vertical search engine which is suitable for the special theme and personalized search is proposed to meet the new needs. The vertical search engine based on the focused crawler is the hotspot and difficult part for the research of the search engine. The aim of the general search engine is to collect the information pages as many as possible, in which the collection order of the pages and the theme of the collected pages are not the main concern. It consumes a lot of system resources and network bandwidth. At the same time the consumption of the resources does not bring a higher utilization for the collected pages. The focused crawler improves the utilization rate of the collected pages by traversing as quickly as possible and collecting web pages relevant to the predetermined theme as many as possible. The most important problem of the focused crawler is how to propose an effective URL analysis model to measure whether the web page pointed by the new URL is relevant to the predetermined theme. The result of the measuring is very important for the resource utilization and accuracy of the vertical search engine The most important problem of the focused crawler is how to propose an effective URL analysis model to measure whether the web page pointed by the new URL is relevant to the predetermined theme. The result of the measuring is very important for the resource utilization and accuracy of the vertical search engine.

In this proposed work, we propose a method to learn the theme automatically based on the search keyword and use the theme description to crawl efficiently. We implemented our scheme on a crawler and measured the accuracy in terms of time to crawl and the accuracy of matching pages found against the time of crawling.

II. RELATED WORK

De Bra et al.[1] propose the FishSearch Model. This model is evolved from the Bionics and biological swarm intelligence[2]. In this algorithm, the Web crawler is modeled as fish in the sea. When fish find the information relevant to the food, they start to breed, and enlarge their population. And they could hunt more food in this way. When the food becomes less(there is no relevant information), or the environment deteriorates there is not enough bandwidth, the population is forced to reduce, and the individual gradually disappears. The core of the algorithm is how to adaptively update and maintain the URLQueue waiting for crawling, according to the theme seed sites which the user is interested in and the change of the theme key words[3].

The method has the advantage of simplified model and dynamic search. But there are also some disadvantages, such as the values for the relevant degree are discrete and the number of the values is small (just only three values, namely 0, 0.5, 1), and the relevant degree is matched only by string; the relevant degree to the theme is difficult to compute accurately by the allocated weights. In the URLQueue, the difference of the priority between different kinds of URL is small.

SharkSearch Model[4] is the improved version of the FishSearch model, the improvement mainly lies in the correlation calculation. The relevant degree of SharkSearch model is not discrete, but is continuous between 0 and 1. Compared with FishSearch model, SharkSearch model is more accurate. The basic idea of PageRank Model[5] is that if a

page is referenced by many other pages, then this page is probably the important page. If a page is referenced by an important page, then the page may also be an important page even though the page is not referenced many times. The importance of a page is divided equally and delivered to the pages referenced by it. The PageRank value is the quantified grade of the page importance. This order is iteratively calculated according to the link information between the web pages. The link information is relatively static, not considering the dynamic information used by the web pages.

HITS(Authorities and Hubs)model divides the important pages into two kinds: the Authority page and the Hub page. Authority page involves the pages that are well-known Authority pages, Hub page which provides the set of links pointing to the Authority pages involves one or more general pages. Generally, good Hub often points to many good Authority pages, good Authority pages are generally pointed by many good Hubs. HITS algorithm utilizes the interaction between Hub and Authority. The procedure of HITS algorithm is as follows[5]: Submiting the query q to the general similarity-based search engine, then search engine returns many pages, take the first n pages as the root set RootSet, denoted by S. Then expand S into a larger set T by adding the pages referenced by S and the pages S refers to into S, as the base set BaseSet. Firstly, assign a non-negative Authority weight ap and Hub weight hp to every page in T, and initialize all ap and hp with the same constant.

The problem which exists in the traditional focused crawler URL analysis model described previously is that the local optimal solution is often easily gotten in the process of searching the relevant pages according to the predetermined theme, namely only crawling around the related web pages, which results in some related web pages which are linked together through hyperlinks with lower degree of relevance are not crawled, then the effective coverage of the focused crawler reduces. The genetic algorithm is a global random search algorithm that based on the evolutionism and molecular genetics, whose prominent feature is the implicit parallelism and the capacity to effective use of the global information, and it can effectively find the global optimal solution jumping local optimum, which is the focused crawler URL analysis model needs.

But the genetic algorithm also has some disadvantages, for example, it can not use the feedback in the system and lots of unnecessary redundancy iterations come out when the solutions reach a certain extent; and the capacity of local search is weak, also may not get the optimal solution. For the crawling strategy of the current common focused crawler, the content of the web pages is generally provided by the editors, which results in some information irrelevant to the predetermined theme involved in the web pages. The whole web page documents will be often used in the genetic process, when the genetic algorithm is used in the focused crawler in the past, which results in that the theme drift easily comes out in the process.

III. OVERVIEW OF PROPOSED SOLUTION

The steps of the URL analysis model of focused crawler based on improved genetic algorithm in this paper are as follows:

(1) Set the theme description vector using a automatic learning method.
(2) Preprocessing of web page documents.
(3) Define the fitness degree function Fitness.
(4) Define crossover probability Pc, mutation probability Pm and expansion probability Pe and so on.
(5) Initialize the generation group P.
(6) Calculate the fitness degree value—Fitness of every individual in the group, get the average value AVG of group fitness degree.
(7) According to the genetic strategy, use selection, expansion, crossover and mutation operation to act on the group to form the next generation group.
(8) Determine if the average value new AVG of the new generation group fitness degree is less than AVG, or has completed the scheduled number of iterations. If it does not fit, return to step (7), or change the inheritance strategy and return to step (7); otherwise it ends..

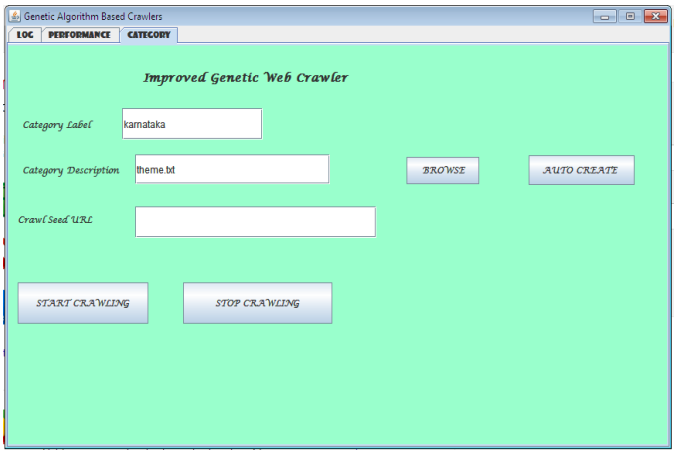
IV. DETAILS OF PROPOSED SECURITY MECHANISM

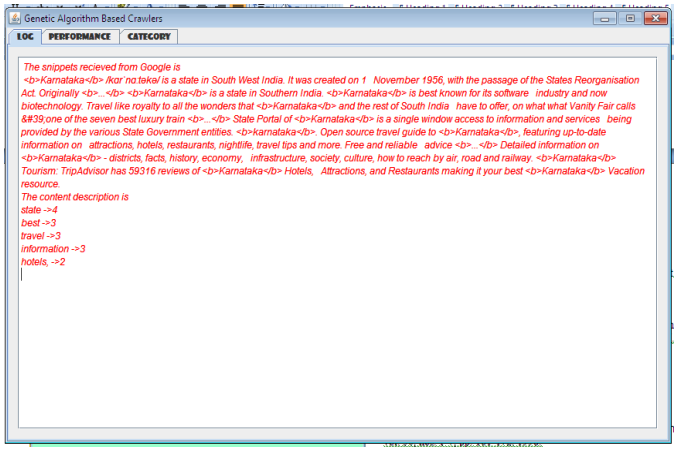
A. Learning Theme for a search keyword automatically

Currently available solutions for genetic crawl are guided by theme vector set by user manually. In this paper we propose a theme learning method. For a given search key word, we extract the first 10 pages snippet from google search. The snippet is broken into terms. Stop works like is.was,are etc are removed from snippet. The words are then converted to adjective. For converting to adjectives we are using Porter Stemmer algorithm. This will bring terms like computing , computed to compute. The remaining terms is then done FI(Frequest Item) analysis to find the most occurring Frequent Item . The frequent item found are used as theme description for the further steps in the genetic algorithm. When visiting the page we found the count of Frequent Item in theme description. If the count is less than the threshold, we don't consider the page or its links to be crawled and drop it. Due to this step, of automatic theme description construction and using it to filter pages, the unnecessary pages are dropped and this improves the accuracy of page retrieval.

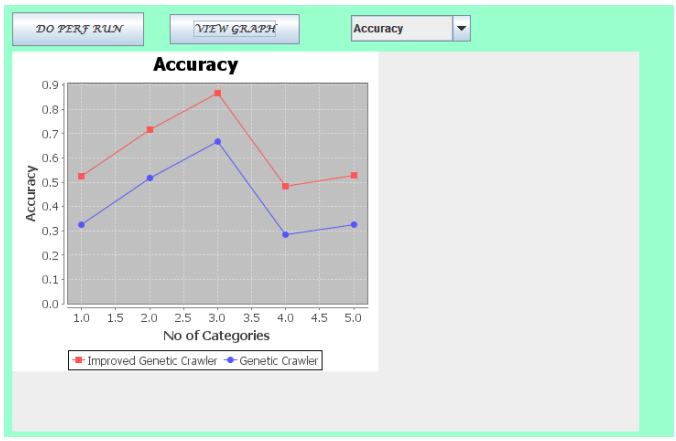
V. RESULTS

Proposing solution in JAVA. We used crawl4J for crawling the genetic library is used for implementing the genetic algorithm proposed in this work. The automatic theme construction for a search term Karnataka is shown below with the proposed solution.





The accuracy of the genetic crawler with and with proposed extension is shown below



The F1 score for the genetic crawler with and without the proposed extension is shown below

VI. CONCLUSION AND ENHANCEMENTS

Proposed system analyzes the dilemma of low relevance and intensive resource when the traditional search engine faces the rapid growth Internet data. It points out that vertical search engine can effectively solve these problems. Focused crawler is the key technology of vertical search engine, and the relevance analysis of URL topic is the problem faced by focused crawler which must be solved firstly. After analysis the existing URL topic relevance analysis algorithm and their own deficiency, this paper proposes a new focused crawler analysis model based on improved genetic algorithm, introducing user query constructed page virtual document into genetic procession to correct theme description, optimizing the crossover operator, selection operator and mutation operator, using vector space model to calculate the similarity degree of anchor text and topic description. The experiment shows that the focused crawler URL analysis model based on improved genetic algorithm proposed in this paper can improve accuracy rate, recall rate and other quotas effectively, and avoid getting into the local optimal solution. Vertical search engine has become the hot and difficult area in the field of search engine, and focused crawler plays a vital role in the area of vertical search engine. Nowadays the integrated use of multiple methods has become a common method in academic search. How to combine genetic algorithm and other method to obtain better results needs to be solved in the following days.

REFERENCES

[1] Paul De Bra and Licia Calvi."Creating Adaptive Hyper documents for and on the Web"in Proceedings of the AACE Web Net Conference, Toronto, 1997: 149-155.

[2] Shang Gao, Jingyu Yang. The Swarm intelligence algorithm and its application[C], China Water Power Press, 2006,

[3] J.Carri`ere, R.Kazman, "WebQuery: Searching and visualizing the Web through Con-nectivity,"Proc.6th International World Wide Web Conference, 1997. 29

[4] Michael Hersovici. The Shark-Search algorithm an application: tailored web site mapping.http://www-2.cs.cmu.edu/~dpelleg/bin/360.html, 1998.

[5] Jiankang Song, Liping Zhang. Web Structure Mining Algorithm. Journal of East China University of Technology. 2003(10): 537- 540

[6] Ling Zhang, Fanyuan Ma. Accelerated Ranking: A New Method to Improve Web Structure Mining Quality[J]. Journal of Computer Research and Development, 2004, 41(1): 98-103

[7] P.M.Aoki.Generalizing search in generalized search trees.Ehrig. Maenchea.Proc of ACM Symposium on Applied Computing□Hawai□ 2007□223-234ng., pp. 63-72, 2004.

