Design and Implementation of Focused Web Crawler Using Genetic Algorithm: An Approach to Web Mining

Prashant Dahiwale, M. M.Raghuwanshi, Latesh Malik

Abstract— The speed at which World -Wide -Web (WWW) is growing round the clock spreds its arms from smaler collections of web pages to a massive hub of web information which gradually increases the complexity of crawling process.search engines handles enourmous quaries from different part of the univers to retrieve most of the relevant results in response to answer the user queries, and it is solely depends on knowledge that it gathers by means of crawling. To tackle this issue the Focused web crawlers are emerging. The crawler is kept focused to the user interests toward the topic, thus crawling processes should be optimum.to make optimum crawling one should use available optimization techniques. This paper proposes a web carawler using genetic algorithm. For selecting more truthfull and proper web pages by web crawler the genetic algorithm as optimization technique has been used. It uses similarity measures which is use to determine the relevancy of the web pages.The results showed that our approach displays with higher quality expected result than traditional focused crawling techniques.

Index Terms—Webcrawler, Focused Crawling, Genetic Algorithm, Mutation, crossover.

1 Introduction

In these days of the spirited world, where each and the every second is measured the important backed up by the information. The World Wide Web is a big set of data. The data keeps rising continuesly round the clock. It is very vital to categorize data as important or non important in accordance with client query. Researchers are operational on techniques which would help to download related web pages. Researchers say that the vast size of data outcomes in reduced coverage of totoal data while search is performed and it is anticipated that only 33%of the data gets listed in to indexer [1]. The web is so gigantic that even the number of important or appropriate web pages that get download is too huge to be discovered by the client. This scenario creates the requirement of downloading the most relevant and excellent pages first. Web search is prsently creating near about 13% of the traffic to Web sites[2]. Web crawler requires to search for information between web pages identified by URLs. If it can believe each web page as a node, then the WWW can be visualise as a data structure that look like a Graph'. To navigate a graph crawler will require traversal mechanisms much nerely equal to those needed for traversing a graph like BFS or DFS, proposed Crawler follows a BFS approach. The Crawler is the most vital part in the search engine. It can traverse 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 user's queries efficiently [3].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 fetches 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.

This paper proposes a web crawling technique which makes use of genetic algorithm for optimum crawling. It uses Jaccard similarity measure to determine relatedness amongst searched result. The section 2 discusses the related work Section 3 gives details about Basics of Web Crawler. Section 4 gives a detailed description about proposed work. Section 5 deliberates with result, section 6 deals with conclusion and future work. And last section is references.

2 Related Work

For search engines so many focused crawling programs are available[4]. Each of which are having advantages and disadvantages[5], some of it uses the different probabilities for the particular input and it gives good result by using the different mutation rates. The techniques presented in [5] are applicable to any domain for which it is possible to generate term-based characterizations of a topic. It gives the total description about

- *Prashant Dahiwale , PhD scholar in computer science and engineering in RTM NagpurUniversity, India,and Assistant Professor at RGCER,Nagpur India,E-mail: prashantdd.india@gmail.com*
- *Dr M.M. Raghuwanshi ,Yeshwantrao Chauhan College of Engineering,Nagpur, India, E-mail: m_raghuwanshi@rediffmail.com*
- *Dr. Latesh Malik,Professor & HOD, Dept of CSE,GHRCE,Nagpur,India, E-mail:lateshmalik@gmail.com*

mutation rate.Sushil Kumar et.al[6] proposed the context model for the focused web search, it describes a Focused Crawler which look for gain, make the index, and keep the collection of the pages on a particular 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. [7] 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. X.Chen et.l [8] 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 [9] 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[10] Hati, D. was proposed block partitioning technique in which the blocks are partitioned by using VIPS algorithm In paper [11] 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[12] 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 [13] , 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 [14] proposed the new hybrid approach to the focused crawling based on the meta search algorithm The paper[15] it extend the performance of the focused web crawling by using the Gcrawler technique. The paper [16], obtains the real mixture of subject-based and link-based Web analysis, concurrently with the capacity to make the universal searching. In paper [17] it proposed the OFC which is based on the reinforcement learning and the fuzzy clustering theory for a focused crawler. In paper [18] Safran et.al proposed the new learning based approach to make better relevance forecast in focused web crawler. Initially, instruction set is built to guide 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.

3 Basics Of Webcrawler

Researchers have developed many techniques to schedule downloading of web pages. It has been a fact that whenever a algorithms have been published they are deficient in terms of much information so that the algorithms cannot be reproduced. The basic rudiments of crawler are that it is a mechanism, a method or a piece of code whose prime focus is to travel across the web, intending for relevant data[19]. Crawlers run in a infinite loop for months together to gather relevant information. These crawlers are also known as spiders, web robots, bots etc. Habitually the crawling starts with a set of seed Uniform Resource Locator (URLs). These URLs are often relevant URLs. It is mandatory that these URLs must be as fine as possible, common practice is to search for the particular keywords on Google, Yahoo, etc. and treat the first five to six URLs as the seed URL.

The crawling process initiates with seed URL which are fetched and downloaded. The next step is to apply some relevancy technique to trace the page is relevant or irrelevant. After the decision of relevancy has been taken and decision is “Yes” then it implies that the page is relevant and the links on it can also be relevant. Hence the links on relevant page are extracted and added to the URL frontier. URL frontier is a queue in which the URLS to crawled while complete crawling process are placed. If the decision is “No” then it implies that the page is not relevant.

```
Initialize the URL Frontier with Seed URL;
While (the URL Frontier is empty)
{
    Fetch/Download the URL;
      If(page is relevant)
      {
          Extract the links on the page;
         Add the extracted links in URL frontier;
       }
        Else
           {
            Store the irrelevant linkin Irrelevant log;
           }
}   End;
```

The process starts by initializing the URL Frontier with seed URL and continues until the complete URL frontier goes empty. Each URL in the frontier gets downloaded one by one. Subsequently the page are getting whether they are relevant or not, if yes then links on that page are extracted and added to the URL frontier else the link is stored in the log of irrelevant links. Once the URL frontier gets empty the crawler loop terminates resulting in relevant and irrelevant links as output of the complete process.

4 Proposed Work

The main aim of proposed work is to choose the most promising links in order and try to maximize the relevancy of a new, unvisited URL by applying the concepts of genetic algorithm for focused crawling.This algorithm will try to produce more optimal result, and it will also helps to improve the accuracy. The proposed method yields promising results.

Step1- The URL frontier is initialized with seed URLs and also placed in Relevant links log.
.Step2- Communicate the User Query. Perform Stop words removal technique and stemming technique on User Query to get refined User Query.
Step3- Process URL in URL frontier one by one to apply GA for optimized crawling result.
Step4- download page for URL.
Step5-compuet score of page/URL by performing link and content analysis on page.
Step6-take out links from processed page and process all taken links for finding score of each link.
Sep7- Perform mutation.
Step8: Select only few top score links from processed pages and send it to frontier.
Step9-if selected link is not earlier visited then do crossover
Step10-Send the page to resultant repository.
Step11- repeate process for all selected pages recursively.
Step12 -Get the output.

The brief decription about stop word removel,steming ,special symbol removel,selection,mutation,crossover is consider as follows.

4.1 Remove special symbol, stop words, stem words

special symbol removel algorithm will removes all the special symbol that are (. @ % & + - / # $! * , etc). Stop word algorithm will remove all the stop words. E.g.-the, are is about all, by etc.stem words algorithm will do steming of all the 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 is consist of keywords. By using these algorithms we can focused on only original refined keywords.

4.2 Web Crawler

The process of web crawling 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 [20][21] 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

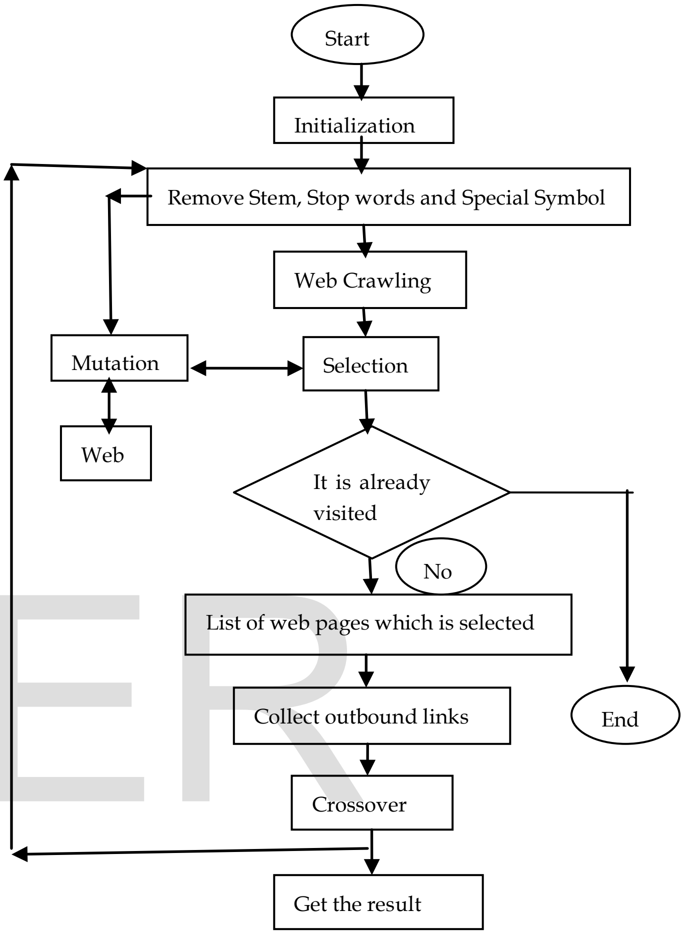

Fig 1Craling flow

4.3 Genetic Algorithm

The main goal of the paper [22] 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 is 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. It also used to optimize the Web crawling and it select a more proper Web pages to be extracted by the web crawler.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 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

$$Jsimilarity\ (P,R)=(X\cap Y)/(X \cup Y) \qquad (1)$$

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.

$$Dpr=\sqrt{(Mp+Nr+ CWpr)} / UWp \qquad (2)$$

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..

$$J(P,R)=Jsimilarity(Jlink)+Dpr(Jkeyword) \qquad (3)$$

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

5 Experimental Result

The proposed Web Crawler' was tested and run on the standard environment and evaluate its performance with videly accepted link and content based web crawler i.e.intelligent web crawler , after operation both the web crawlers in similar hardware and software surroundings with equal input parameters, experimental following rate of precisions for mutually the crawlers as shown in Table. calculate the performance of proposed structure by equating it with Intelligent web crawler by operating both the crawlers to the equal problem set.

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

Table 1: Performance Comparison table

	intelligent crawler		Crawler with GA	
	Precision	Crawl Time in Seconds	Precision	Crawl Time in seconds
What is java?	5.43	252	19.43	195
Book a movie ticket	14.29	642	43.14	502
Books Of ChetanBhagat	12.86	591	32.00	459
Samsung galaxy E7 mobile	14.57	612	42.00	543
Recruitment for PO	4.00	363	6.57	259
Java classes	9.43	297	26.29	222
Anna hazare in anti cor-	7.71	338	19.71	249
Earthquake in Nepal	11.71	438	24.86	307
Cricket Match	16.29	479	36.86	388
Jewelry	17.43	389	38.57	341
Artificial Intelligence	8.86	392	19.14	291
Health and medical in-	6.86	493	18.00	368
Pizza	14.00	413	30.00	317
Student special tours	6.29	513	17.43	429
Special offers	15.14	542	31.14	472

Following figure shows precision comparision between two crawlers as per experimental results noted.

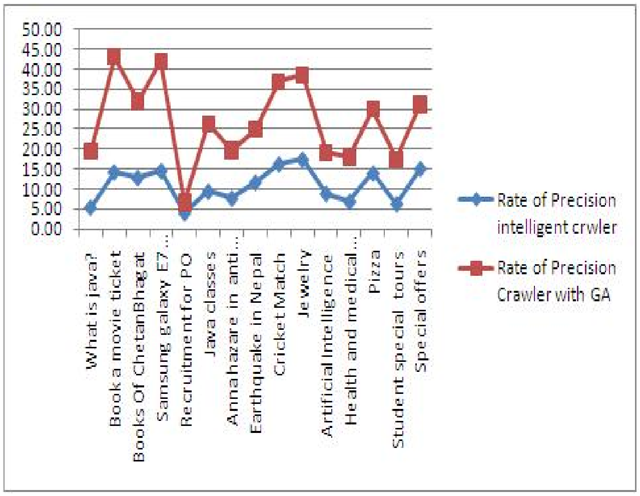

Fig 2 precision Comparision

For each query fired , precision for proposed crawler (crawler with GA) shows far better than available intelligent crawler that is link and content based crawler.

Also next figure shows comparision of time of crawl for both the crawlers.

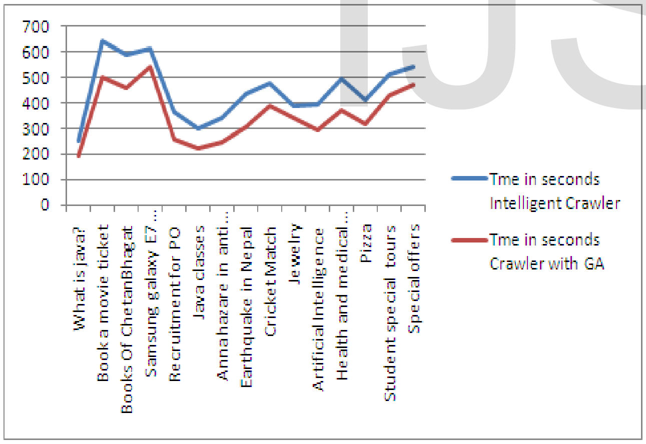

Fig 3 time of Crawl comparision

Above graph shows that proposed crawler taking much lesser time than intelligent crawler..

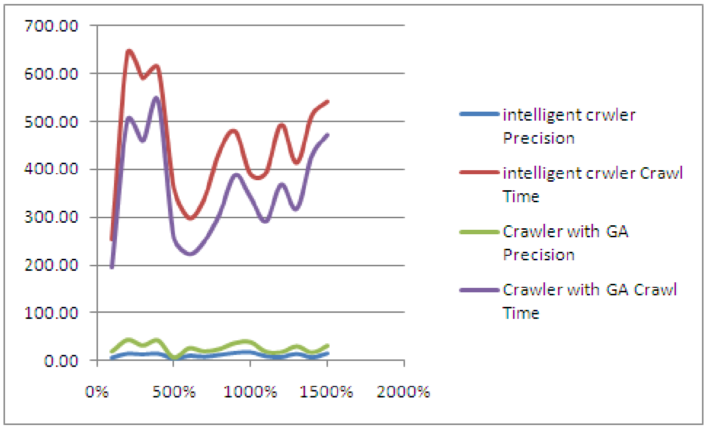

Fig 4 Overall performance of both crawlers

From fig4 it is clearly identified that rate of presion of proposed crawler is very high in comparision with intelligent crawler on the other hand time of crawle is very less which meance proposed craler gives double benefits .while in intelligent crawler ,some queries gain very little precision and more time.but for same queris proposed crawler gain very high precision and less time.

6 Conclusion

Form above experimental observations it is concluded that by using genetic algorithm crawler is geting more relevant links in less time meance more precision in less time. It also helps to improve the crawling performance. And proposed crawling process terminates when an optimum solution is found. The advantages of this approach is 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.

Acknowledgment

I would like to express thanks to dr M M raghuwanshi, Dr L G Malik who helped constant guidance and motivation througt experimentation .Thanks for your advice during the work. I would like to thanks all the authors of the papers I referred for this paper.

References

[1] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280(5360):98.100,1998.

[2] StatMarket. Search engine referrals nearly double worldwide.http://websidestory.com/pressroom/-pressreleases.html?id=181, 2003.

[3] Soumen Chakrabarti and Martin van den Berg and Byron Dom, "Focused Crawling: A New Approach to the Topic-Specific Web Resourc Discovery" Proceedings of the 8th International WWW Conference.

[4] Pavalam S M, jawahar M, S V Kashmir Raja and Felix K Akorli "A Survey of Web Crawler Algorithms" International Conference on Machine Learning(ICMLC 2011)

[5] Rocio L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman and N'elida Beatriz Bringnole"Genetic Algorithms for Topical Web Search: A Survey of Different Mutation Rates" In 2005.

[6] Sushil Kumar ,Naresh Chauhan , "A Context Model For Focused Web Search", International Journal of Computers & Technology Volume 2 No. 3, June, 2012.

[7] Brin, S. and Page, L. (1998),"The Anatomy of a Large- Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, 30(1–7)

[8] X.Chen and X. Zhang , "HAWK: A Focused Crawler with Content and Link Analysis", Proc. IEEE International Conf. on e-Business Engineering ,2008

[9] S. N. Sivanandam, S. N. Deepa "Introduction to Genetic Algorithms" Springer, 2008.

[10] Hati, D.; Kumar, A. "Improved focused crawling approach for retrieving relevant pages based on block partitioning", 2nd International Conference on volume3 Communication, Networking & Broadcasting ;Computing & Processing (Hardware/Software) ; Robotics & Control Systems,2010,PP-269.

[11] Deep MalyaMukhopadhyay, Maricel O. Balitanas, AlisheroyFarkhod A Seung-Hwan Joen and Debnath Bhattacharyya "Genetic Algorithm: A Tutorial Review International Journal of Grid and Distributed Computing Vol.2, No.3,September, 2009.

[12] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Journal of the ACM, 1999, 46(5), 604-632

[13] M. Diligenti and F. M. Coetzee and S. Lawrence and C. L. Giles and M.Gori, "Focused Crawling Using Context Graphs", In Proceedings of the 26th V.L.D.B Conference, Cairo,Egypt, 2000.

[14] Y. Sun, P. Jin, and L. Yue. A framework of a hybrid focused web crawler. Future Generation Communication and Networking Symposia, 2008. FGCNS '08. Second International Conference on, 2, 2008.[5]

[15] Shokouhi M, Chubak P, Raeesy Z. "Enhancing Focused Crawling with genetic Algorithms" Proceedings of the International Conference on International Technology: Coding and Computing (ITCC). 2005; 2: 503-508.

[16] Anshika Pal, Deepak Singh Tomar, S.C.Shrivastava, "Effective Focused Crawling Based on Content and Link Structure Analysis", I.J.C.S.I.S. Volume 2, In June 2009.

[17] Qiang Zhu "An Algorithm OFC for the Focused Web Crawler.Machine Learning and Cybernetics, 2007 International Conference Vol. 7 Page(s): 4059 –4063.

[18] Mejdl 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

[19] Kim, S. J. and Lee, S. H. "An improved computation of the PageRank algorithm"in Proc. of the European Conference on Information Retrieval (ECIR', 2002, pp. 73—85).

[20] Jaytrilok Choudhary and Devshri Roy ," A Priority Based Focused Web Crawler" , International Journal of Computer Engineering and Technology , Volume 4 ,Issue 4, july-august 2013.

[21] Steven S. Skiena "The Algorithm design Manual" 2nd Edition, Springer Verlag London Limited, 2008, Pg 160.

[22] Ben Coppin "Artificial Intelligence illuminated" Jones andBarlett Publishers, In 2004,

[23] BanuWirawan Yohanes1, Handoko2, HartantoKusuma Wardana3"Focused Crawler Optimization Using Genetic Algorithm" December 2011.