Artificial Bee Colony (ABC) Approach for Ranking Web Pages

G.Anuradha
Research Scholar
Dept. of CS&SE
Andhra University
Visakhapatnam
Andhra Pradesh, India

G. Lavanya Devi, Ph.D
Assistant Professor
Dept. of CS&SE
Andhra University
Visakhapatnam
Andhra Pradesh, India

ABSTRACT

The World Wide Web (WWW) is rapidly growing on all aspects and is a massive, explosive, data resource in the world. In information retrieval approach web Search engines are predominant tools for finding and getting access to the contents of web. The primary goal of Search engine is to provide relevant information to the users according to their needs, Usually Search engines gives large result set for a user's query. To limit the result list, it is necessary to assign ranking the web pages in an efficient and effective manner. Artificial Bee Colony (ABC) is one of the new approaches used to solve optimization problems. This paper proposes Artificial Bee Colony (ABC) approach as a new method for web mining particularly in ranking web pages. It considers users interest, total web site linkage and growth analysis rate are used to assign rank the web pages. Proposed ABC approach for ranking web pages is implemented and tested on real datasets. The experimental results shows efficiency of the proposed method compared with tradition page Rank Algorithm.

Keywords

WWW, Search engine, Artificial Bee Colony (ABC), User interest, Total web site linkage, growth analysis rate

1. INTRODUCTION

The exponential growth of the data on www is a challenge to Search Engines. User needs to use information retrieval tools for desired information. Search engine is a tool used to find required information from the World Wide Web. The architecture of search engine is shown in fig1, it consists of three major components: Crawler, Indexer and Ranking mechanism[1].Crawler traverses the web and collects web pages from the web. Collected web pages are sent to index module. Indexer creates and maintains the index. When user poses a query in the interface of the search engine, query processor component match is the query keywords with the index and returns the URLs of the pages to the user. Ranking mechanism is applied before showing results to the user. Page ranking was first introduced to rank the importance of web pages on the web. It is a fundamental requirement of search engines to make the search results up-to-date and very fast[2].

In any information retrieval system ranking plays a main Role. Most of the Search engines return million of pages for a given query, It is highly impossible for a user to preview all the returned results, ranking is helpful in web searching.Based on content and connectivity, ranking is divided into two categories.Content based ranking is depends on content of web page,Connectivity ranking based on link analysis technique.

There are two famous link analysis methods [3]:- i)Page Rank Algorithm and ii) HITS Algorithm.

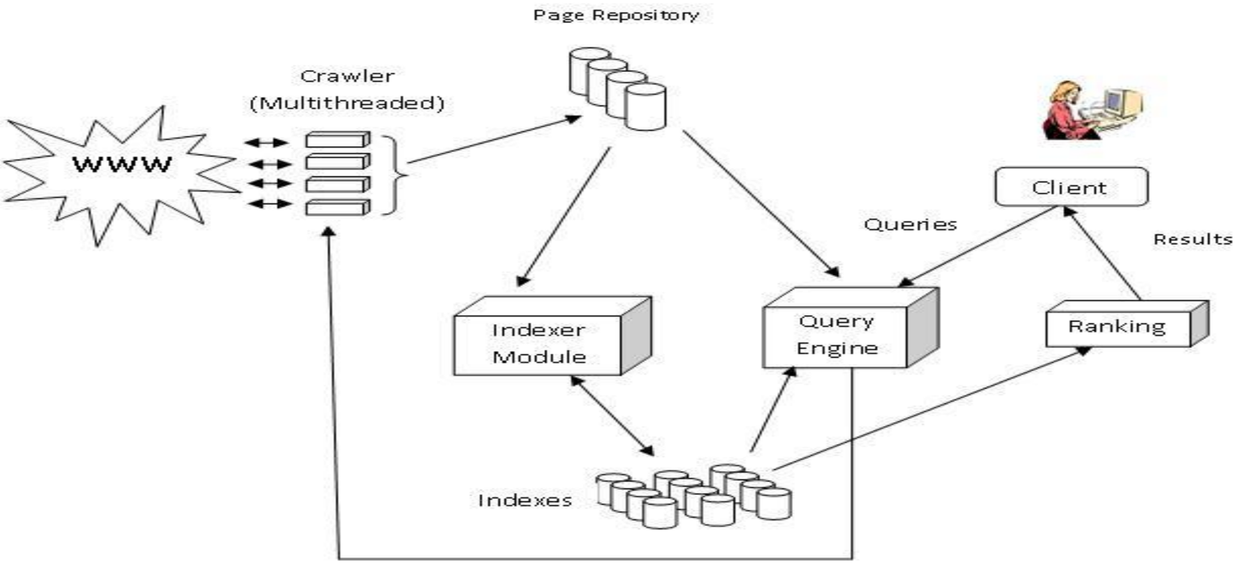
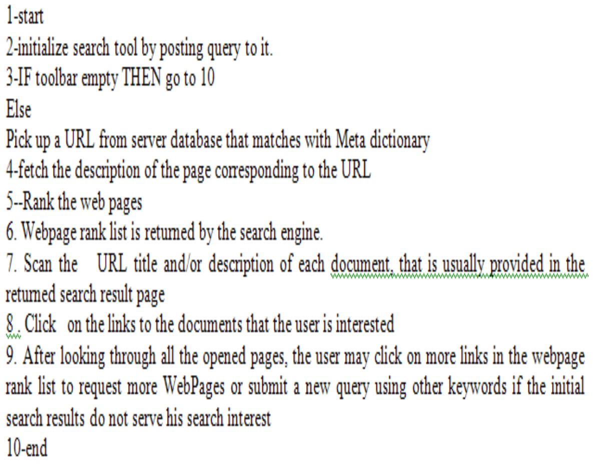


Fig1: Basic search engine Architecture

Ranking process is shown in the following algorithm:



1-start
2-initialize search tool by posting query to it.
3-IF toolbar empty THEN go to 10
Else
Pick up a URL from server database that matches with Meta dictionary
4-fetch the description of the page corresponding to the URL
5--Rank the web pages
6. Webpage rank list is returned by the search engine.
7. Scan the URL title and/or description of each document, that is usually provided in the returned search result page
8 . Click on the links to the documents that the user is interested
9. After looking through all the opened pages, the user may click on more links in the webpage rank list to request more WebPages or submit a new query using other keywords if the initial search results do not serve his search interest
10-end

Fig2: Shows Ranking Process

Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score .Web mining techniques are employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users. If the search results are not displayed according to the user interest then the search engine will lose its popularity[1]. So the ranking algorithms become very important. Most important link analysis algorithm is "PAGERANK" developed by Google. If the content of the web page is frequently updated by the owner with most relevant data, obviously the user heuristically gets attracted towards that web page and this makes the web page to get more interests than his competitors .On the other head this is not possible with Page Rank algorithm as the referential concept only gives URL irrespective of the content. As a reason through the content is improved, it is not depicted properly, Page Rank algorithm considers only URL's but not updates of contents. This article tries to address the above mentioned drawbacks by proposed ABC Approach. This proposed approach calculates User interest, Growth Analysis rate, and Total site linking. This algorithm ultimately gives more relevant information for the query posed by the user when compared to the exists Page Rank algorithm. Experimental analysis is performed by considering 5000url's.The results are proved to be encouraging. The proposed algorithm can be adopted by any Search Engine for that this algorithm can be extended for different datasets.

This paper as organized as follows: Section 2, presents Page rank algorithm overview
, Artificial Bee Colony, Section 3, introduce the Artificial Bee Colony Algorithm for web page ranking Section 4 explains implementation of the proposed method.

2. BACKGROUND

In this section explains Page Rank algorithm and ant colony algorithm

2.1 Page Rank Algorithm [2]

The Page Rank algorithm presented by Brin, Page et'al is one of the factors used by Google to calculate the relative importance of the web pages. The Page Rank value of a Web page depends on the Page Rank values of pages pointing to it and on the number of links going out of these pages. In this algorithm those web pages with more citations are more important. The of the advantage Page Rank is that it does not only depend on the count of referrals, but also considers the importance of the cited web page.

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR\ (Tn)\ /C\ (Tn))$$

Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one.

or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

2.2 Artificial Bee Colony(ABC) Behavior

Swarm intelligence is the field of computer science that designs algorithms and studies efficient computational methods for solving problems inspired by the collective behavior of social insect [5]. Karaboga in 2005 introduced as swarm based algorithm ABC ,it inspects the behaviors of real bees on finding nectar amounts and sharing the information of food sources to the other bees in the hive. The process of searching for nectar in flowers by honeybees can be observed as an optimization [6,7]process.

Tereshko developed a model of foraging behavior of a honeybee colony based model .This model shows collective intelligence of honeybee swarms consists of three essential components: food sources, employed foragers, and unemployed foragers,

Tereshko explains the main components of his model as below[8-10]:

(i)***Food Sources***: To select a food source, a forager bee evaluates some properties related with the food source i.e closeness to the hive, richness of the energy, taste of its nectar, and the ease or difficulty of extracting this energy..All these parameters represents quality and of a food source

(ii)***Employed foragers***: An employed forager is employed at a specific food source which she is currently exploiting. She carries formation about this specific source and shares it with other bees waiting in the hive. The information includes the waggle distance, it consists Direction in which it will be found and Distance from the hive. Its quality rating (or fitness)

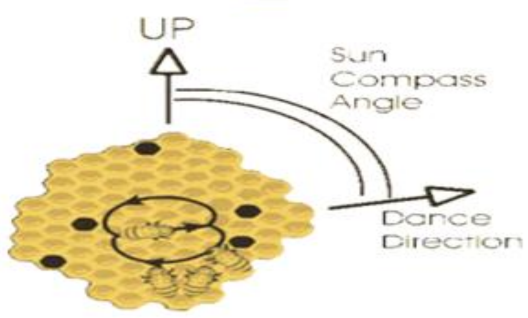


Fig3: Shows honey Bee waggle dance

(iii)***Unemployed foragers***: A forager bee that looks for a food source to exploit is called unemployed. It can be either a scout who searches the environment randomly or an onlooker who tries to find a food source by means of the information given by the employed bee. The mean number of scouts is about 5–10%.

The Basic ABC algorithm:

```
Bee Initialization Phase
Set the Loop
Employed Bee Phase
Onlooker Bee Phase
 Scout Bee Phase
Memorize the best solution found so far
 Until the loop is terminated
```

3. PROPOSED ALGORITHM

Key words: Scout Bee, Direction (for more honey) , Quality of the nectar, Distance of the nectar from hive

Inputs

Web pages
Meta Keywords
New Interest value of the user on the web page(I) //calculated as Ii-Ij where Ii : Initial interest and Ij : Current calculated Interest
Time Stamp (T) //Time of last update of the interest value on the web page
Growth Analysis of the web page(Di)
Distance of the target web page from root(Ds)

Outputs

Optimized relevance on the search results for the query posed.

Variables

// Variables considered

t[]	-	For storing tokens from the input string
num_s	-	To store passed query (type : String)
num_count	-	Search Dimension (no. of results depicted to be relevant)
num_i	-	To update/increment the user interest (type : static)
num_t	-	To store the elapsed time interval between last and previous update
num_ Di	-	To store the growth analysis value of the web page
num_Ds	-	To store the distance of web page

Algorithm

```
//Initialization
1.num_count = 0
2.t[]=tokens(num_s)
// Query Processing
3.while(each(t[]) matches with meta)
4.Display as hyperlink to the location of the web page order by      ( num_i + num_Di + num_Ds )
5. Increment the value of num_count  /* To show dimension value (To show how many results found relevant) */
6.//End while
7. num_i is incremented upon user click on the respective result link
 //server side upgradation
8. If ( (I=i/X) >= Threshold)

9. Perform I=I/X for every X days to keep sites in race condition
10.End if
Repeat step (3) each time the query is posed and step (8) for every X days.
```

4. IMPLEMENTATION OF PROPOSED SYSTEM

The performance of waggle dance by an employee bee informs his fellow bees about the Direction of availability nectar or pollen, Distance of food zone from the hive and Quality of the nectar available.

These three are the agents that contribute for the successful foraging of food. Hence it is called multi agent system. In the proposed method Direction of more honey is considered as growth analysis of the web page, Distance of food zone from the hive is considered as total sites referring this web page and Quality is considered as the User interest values .

order by *(num_i + num_Di + num_Ds)*

Implementation is divided into 3 phases:

Phase-1: Calculation of Users interest (*I*)

User interest(I) calculated using 5000 Urls that are collected and hosted on goongo server named www.goongo.in. Goongo was popularized by using social media and oral advertising. The time stamp of 15 days was considered. The user's interest for every 15 days of usage of Goongo was observed[13].

Phase-2: Growth analysis Rate *(num_di)*

The growth analysis of the web page(num_di) is obtained from the web information site "alexa.com[12]. The value of each web page present on the dataset is collected from the web site. These values are assigned to the web page based on various facts like network traffic analysis, number of distinct users of the site, average time spent by the user on the web page etc… This value can be positive or negative. It is purely dependent on its value versus the previous three months. Based on the above factors the growth analysis can be determined. This value can be compared to direction of more honey because the direction of more honey and direction of more relevant data spells one and the same.

Phase-3: Total site linking *(num_ds)*

The distance of hive from the food zone is considered as total sites linking in to the web page(num_ds). The ultimate food zone is the most relevant web page in the real world. The total sites referring the web page considered as the distance of web page from the root. This makes sense because no web page either positive or negative can be reached without a referential back links. However this referential concept is hammered as back drop in the ancestor algorithms. But here it is considered as one of the agents but not the main and only agent.

The values of Interest (I), Direction (dir) and distance will be continuously varying that makes no web site to be at the top always. These values are needed to be frequently updated into the database. Because of this continuous process of server up gradation we find more relevant web pages on the top unlike the native page ranking algorithms.

Calculate ABC Ranking : Using Page rank algorithm (Total sites linking to the web page, user interest ,Growth Analysis rate

Implementation of Artificial Bee Colony Algorithm for searching various topics on the given dataset revealed considerable results. First comes the manual comparison of obtained search results:

Table1: Shows relevant pages for the query "tourism" produced by page rank and ABC

Size Of the Page Set	Number of relevant Pages	
	Page Rank	ABC
10	4	8
20	10	15
30	16	23
40	23	31
50	30	38

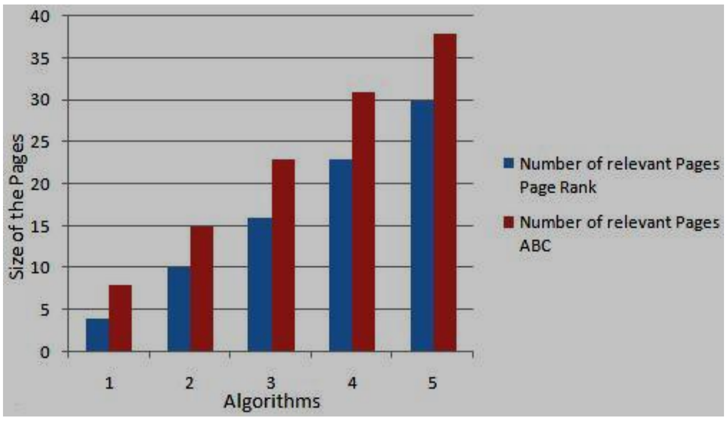


Fig4: Shows relevant pages of given query "Tourism" for page rank and ABC

Table2: Shows relevant pages for the query "Health" produced by page rank and ABC

Size O f the Page Set	Number of relevant Pages	
	Page Rank	ABC
10	2	3
20	6	7
30	11	11
40	13	15
50	19	18

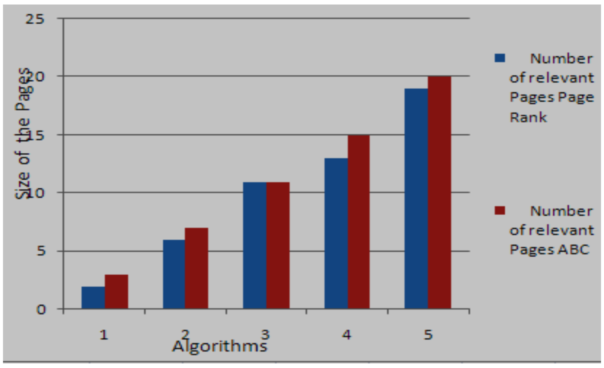


Fig4: Shows relevant pages for the query "Health" produced by page rank and ABC

5. CONCLUSION

This paper has proposed an algorithm for ranking web pages inspired by artificial bee colony. The goal of this algorithm is to assign rank for web pages based on Users Interest, Total sites linking to the web page (Page Ranking), Growth Analysis rate. The process of updating the interest on the web pages is continued .This makes no web page to be always at the top. If the content of the web page is frequently updated by the owner with most relevant data, obviously the user heuristically gets attracted towards that web page and this makes the web page to get more interests than his competitors. On the other head this is not possible with Page Rank algorithm as the referential concept only gives URL irrespective of the content. As a reason through the content is improved, it is not depicted properly, Page Rank algorithm considers only URL's but not updates of contents. The proposed algorithm has overcomes this issue. Experimental analysis is performed by considering 5000url's.The results are proved to be encouraging. The proposed algorithm can be adopted by any Search Engine for that this algorithm can be extended for different datasets.

6. REFERENCE

[1] Laxmi Choudhary, Bhawani Shankar Burdak. 2008," Role of Ranking Algorithms for Information Retrieval " Discrete Algorithms Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms; Pages: 1010-1018.

[2] Manju Patel1 and Shweta Modi. 2011," A Survey on Distributed Page Ranking", International Journal of Chemistry and Applications. ISSN 0974-3111 Volume 3, pp. 201-208

[3] Hema Dubey, Prof. B. N. Roy. 2011,"An Improved Page Rank Algorithm based on Optimized Normalization Technique", International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2183-2188

[4] http://www.cs.princeton.edu/chazelle/courses/BIB/pagerank. htm

[5] E. Bonabeau, M. Dorigo, G. Theraulaz. 1999, "Swarm Intelligence: From Natural to Artificial Systems", New York, NY: Oxford University Press

[6] D. Karaboga, b. Akay. 2005, "Artificial bee colony (abc), harmony search and bees algorithms on Numerical optimization", Erciyes University, the dept. Of computer engineering, 38039, melikgazi, kayseri, turkiye

[7] Ashita S. Bhagade, Parag. V. Puranik. 2012, " Artificial Bee Colony (ABC) Algorithm for Vehicle Routing Optimization Problem", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2.

[8] Milos Subotic, "Artificial bee colony algorithm with multiple onlookers, for constrained optimization problems", Proceedings of the European Computing Conference

[9] V. Tereshko. 2000, "Reaction–diffusion model of a honeybee colony's foraging behavior", in: M. Schoenauer (Ed.), Parallel Problem Solving from Nature VI,Lecture Notes in Computer Science, vol. 1917, Springer–Verlag, Berlin, , pp. 807–816.

[10] V. Tereshko, T. Lee. 2002, "How information mapping patterns determine foraging behaviour of a honeybee colony", Open Systems and Information Dynamics 181–193.

[11] V. Tereshko, A. Loengarov. 2005, Collective decision-making in honeybee foraging dynamics, Computing and Information Systems Journal 9 (3).

[12] www.alexa.com

[13] G. Anuradha, G. Lavanya Devi and M.S Prasad Babu. 2014, “Antrank: an ant colony algorithm for ranking web pages”, ijettcs