Building web crawler based on bee swarm intelligent algorithm

mohammed Ibrahim Shujaa[1] ahmed bahaa ulddin [2],

[1] department of informatics, foundation of technical education
Baghdad, State ZIP/Zone, Iraq
[2] department of informatics, foundation of technical education
Baghdad, State ZIP/Zone, Iraq

Abstract

Search engines are using web spiders to crawl the web in order to collect copies of the web sites for their databases, these spiders usually use the technique of breadth first search which is non-guided (blind) depends on visiting all links of any web site and one by one. This paper proposed a new algorithm for crawling web depending on swarm intelligence techniques, the adopted algorithm is bee swarm algorithm which takes the behavior of the bee for its work, the result in terms of speed and accuracy which means the relevancy of the collected sites.

Keywords: *web crawling, crawlers, bee swarm algorithm, swarm intelligence.*

1. Introduction

One of the most essential jobs of any search engine is gathering web pages also called crawling; web crawler is a software or program that uses the graphical structure of the web to move from page to page and add them to a local database, web crawlers used to create a copy of all visited pages for later processing by search engine], another definition for the web crawler (also known as robot) is a system for downloading of the web pages, web crawlers used for variety of purposes[1].
Most permanently considered one of the main component Of the web search engines that assemble a corpus of web pages, index them and allow users to issue queries against the index and find the web pages that match the queries, a related use is web archiving where large sets of web pages are periodically collected and archived for posterity and web data mining where web pages are analyzed for statistical purposes [2].the simplest form of web crawler starts from a seed page and then us-es the external links with it to attend other pages, the process repeated with new pages offering mo-re external links To follow until a sufficient number page-s are collected or some high level objective is reached, the working of crawler can be shown in the following algorithm:

1-start
2-initialize (frontier) query by adding seed URL to it.
3-IF frontier empty THEN go to 7
Else
Pick up a URL from frontier
4-fetch the page corresponding to the URL
5-parse the page to find new URL's
6- Add new unvisited URL's to the queue
7-end

Therefore, any search engine's web crawler should have the following objectives:-
1-it should explore and download web documents from the World Wide Web as much as possible.
2-it should bring a high quality documents so that the users gets the required relevant information within acceptable time.
3-the documents must be displayed in order to their relevance.
4-the web documents is very dynamic search engine should update it's repository.
To satisfy the first objective, search engine depend on multiple crawlers, and the ranking algorithm satisfy the second and third objective, the fourth objective can be satisfied by frequent crawling [1].

2- Problem state

in order to get an accurate functionality of the search engine, it need an efficient web crawler providing an accurate database for web sites, the problem of the web crawling is the crawling of not relevant or not useful pages and slowness of most crawlers, so this paper propose a new web crawling algorithm depends on the bee swarm intelligence algorithm in order to speed up crawling and use the collective intelligence of the bee to crawl
Most relevant (useful) web pages with good speed.

3-swarm intelligence

Swarm intelligence is a section of artificial intelligence based on study of actions in various decentralized system, this decentralized system (multi agent system) are composed of the physical individuals (robot) or virtual ones that communicate among themselves, cooperate,

exchange information and knowledge and perform some tasks in their environments [4,5].

Another definition from Dorgio and Birattori swarm intelligence is the discipline that deals with natural and artificial system composed of many individuals that coordinate using decentralized control and self – organization in particular, the discipline focuses on collective behavior that results from local interaction of the individuals with each other and with their environment[3].

A high level of view of a system suggest that N agents in the swarm are cooperating to achieve purposeful behavior and achieve some goal, this apparent "collective intelligence" seems to emerge from what are often large groups of relatively simple agents, the agents use simple local rules to govern their actions and via interaction of the entire group, a type of the (self organization) emerges from the collection actions of the group[4]. The swarm framework or meta formalism present a unified scheme or approaching the problems and investigating the ways to implement swarm intelligence, it addresses how swarm of the entities must communicate and modify their behavior in response to information from other entities and their environment for there to exist the emergent self-organized behavior known as "swarm intelligence"[5]. There are important characteristics of the swarm colony:-

- Flexible:-the colony can respond to the internal perturbations and external challenges.
- Robust:-tasks are completed even of some individuals fail.
- Decentralized :-there is no control in the colony.
- Self-organized:-paths to the solution are emergent rather than predefined.

4. Behavior of bees in nature

Before illustrating bee swarm algorithm, we need to understand bee in nature, the best example is the collection and processing of the nature, the practice of which is highly organized each bee decides the nectar source by following a nest mate has already discovered a patch of flowers, each hive has a so called dance area in which the bee that have discovered nectar source dance in dance in that way trying to follow them if the bee decides to leave the hive to the nectar, this bee follows one of the bee dancers to one of the nectar areas, bees communicate through this waggle dance which contain the direction, the distance and the quality (fitness) of the flowers patch , these information helps the colony to send it's bees[6,7], Tereshko explains the main components of his model as below :-

1- food source: in order to select a food source, a forager bee evaluates several properties related with the food source such as its closeness to the hive, richness of energy, taste of its nectar, and the ease or difficulty of extracting this energy. For the simplicity, the quality of a food source can be represented by only one quantity although it depends on various parameters mentioned above.

2- Employed foragers: an employed forager is employed at specific food source which she is currently exploiting. She carries information about this specific source and shares it with other bees waiting in the hive. The information includes the distance, the direction and the profitability of the food source.

3- Unemployed foragers: a forager 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 number of scout s is about 5-10%. The exchange of information among bees is the most important occurrence in the formation of collective knowledge. While examining the entire hive it is possible to distinguish some parts that commonly exits in all hives, the most important part of the hive with respect to exchanging information is the dancing area. Communication among bees related to the quality of food sources occurs in the dancing area. The related dance is called waggle dance, since information about all the current rich sources is available to an onlooker bees are those bees that are waiting on the dance area in the hive for the information to be shared by employed bees about their food sources, and then make decision to choose food source[5]. Probably she could watch numerous dances and chooses to employ her self at most profitable source. There is greater probability of onlookers choosing more profitable source since more information is circulating about more profitable sources. Employed foragers share their information with probability which is proportional to the profitability of the food sources, and the sharing of this information through waggle dancing is longer in duration. Hence, the recruitment is proportional to profitability of a food source [6,7]. In order to better understand the basic behavior characteristics of foragers, let us examine figure 1

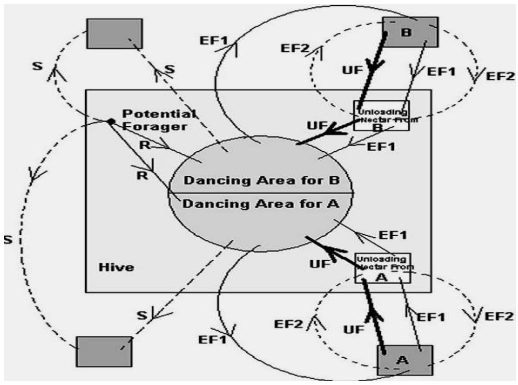

Fig. 1 bee in nature

Assume that there are two discovered food sources; A and B. at the very beginning, a potential forager will start as unemployed forager, that forager bee will have no knowledge about food sources around the nest. There are two possible options for such a bee:-

1-it can be a scout and starts searching around the nest spontaneously for food due to some internal motivation or possible external clue (S on fig 1).
2-it can be recruiting after watching the waggle dances and starts searching for food sources (R on fig 1). After finding the food source, the bee utilizes its own capability to memorize the location and then immediately starts exploiting it; hence, the bee will become an employed forager. The foraging bee takes a load of nectar from the source to the hive, unloading the nectar to food store. After Unloading the food, the bee has the following options:-
1-it might become an uncommitted follower after abandoning the food source (UF in fig1.)
2-it might dance and then recruit nest mates before returning to the same food source (EF1 in fig.1).
3-it might continue to forage at the food source without recruiting bees (EF2 in fig.1).
It is important to note that not all bees start foraging simultaneously. The experiments confirmed that new bees begin foraging at rate proportional to the difference between the eventual total number of bees and the number of the bees presently foraging [6, 7].

5-artificial bee algorithm

Artificial bee colony (ABC) algorithm is a new swarm intelligence algorithm , which was first introduced by Karabog in Erciyes university of turkey in 2005[4], and the performance of ABC was analyzed in 2007. The ABC algorithm imitates the behaviors of the real bees on searching food source and sharing the information of food sources with other bees. Since the ABC algorithm is simple in concept, easy to implement, and has fewer control parameters, it has been widely used in many fields, such as constrained optimization problems, neural networks and clustering [6, 7]. In the ABC algorithm, the colony of artificial bees is classified into there categories: employed bees, onlooker, and scouts. Employed bees are associated with particular food source which they are currently exploiting or employed at. They carry with them information about this particular source and share information to onlookers. Onlooker bees are those bees that are waiting on the dance area in the hive for the information to be shared by the employed bees about their food sources, and then make decision to choose food source. A bee carry out random search is called a scout. In the ABC algorithm, first half of the colony consists of the employed artificial bees and the second half includes the onlookers. For every food source, there is only one employed bee. In other words, the number of employed bees is equal to the number of food sources around the hive. The employed bee whose food source has been exhausted by bees becomes a scout. The position of a food source represent a possible solution to the optimization problem and the nectar amount if a food corresponds to the quality (fitness) of the associated solution represented by that food source. Onlookers are placed on the food sources by using probability based selection process. As the nectar amount of food source increases, the probability value with which the food source is preferred by onlookers increases too[7]. This is the main steps of the algorithm are given below:-

1-cycle=1
2-initiate the food source position Xi,I=1,…Sn
3-evaluate the nectar amount (fitness FIT) of food sources
4-repeat
5-employed bees phase
  For each employed bee
  Produce new food position Vi
  Calculate the value fit (i)
  Apply greedy selection mechanism
  End for
6-calculate the probability values Pi for the solution
7-onlooker bees phase
  For each onlooker bee
  Choose food source depending on Pi
  Produce new food source position Vi
  Calculate the value fit (i)
  Apply greedy selection mechanism
  End for
8-scout bee phase
  If there is an employd bee becomes scout
Then replace it with new random source position
9-memorize the best solution achieved so far
10-cycle=cycle+1
11-until cycle=maximum cycle number

In the initialization phase, the ABC algorithm generates a randomly distributed initial food source positions of SN

solutions, where SN denotes the size of employed bees or onlooker bees. In the employed bees phase, each employed bee find a new food source Vi in the neighborhood of its current source Xi. The new food source is calculated using the following equation (1):

$$vij = xij + \phi ij(xij - xkj) \quad \text{----------(1)}$$

Where $k \in (1, 2,..., SN)$ and $j \in (1, 2,..., D)$ are randomly chosen indexes, and k has to b different form i . ij is random number between [-1,1]. And then employed bee compares the new one against the current solution and memorizes the better one by means of a greedy solution mechanism. In the onlooker bees phase, each onlooker chooses a food source with a probability which related to the nectar amount (fitness) of a fork) od source shared by employed bees. The probability is calculated using the following equation (2):

$$pi = fiti \sum_{n-1}^{Sn} fiti \quad \text{----------(2)}$$

In the scout bee phase, if a food source can not be improved through a predetermined cycles called “limit”, it is removed from the population and the employed bee of that food source becomes a scout. The scout bee finds a new random food source position using the equation(3) below :

$$x_i^j = x_{\min}^j + rand[1,0](x_{\max}^j - x_{\min}^j) \quad \text{----------(3)}$$

Where min j and max Xj are lower and upper bounds of parameter j respectivelyv[6, 7, 8, 9].

5-crawling techniques

This section will explain some important crawling algorithms, the beginning will be with “breadth first algorithm” which is the simplest strategy, blind traversing approach, this algorithm was explored in 1994, it uses two frontiers(queues) FIFO , it is crawling links in order in which they are encountered, the problem with this algorithm is that when the frontier is full, the crawler add only link from a crawled pages , the other problem is that it traverse URL’s in sequential order as these were inserted into frontier, it does not calculate the relevancy of the link , so many useless page can be crawled . to overcome this problems of the blind traversing approach[1] , a heuristic approach called out best first crawling have been studied by Cho in 1998, the following naïve best algorithm uses relevancy function to compute similarity between desired keywords and each pages, all of the above algorithms was static approach. The followed (fish and shark) algorithm are dynamic that means fetching data in time the query is issued. In fish algorithm the intuition that relevant often have relevant neighbors, thus it searches deeper under documents have been found to be relevant to the search query and stops searching in the dry areas. In shark search algorithm is development for fish, one improvement is that instead of binary (relevant/irrelevant) evaluation of document relevancy, it returns a fuzzy score (i.e a score between 0 and 1) [1, 2].

Navarat in 2006 proposed an approach to web search based on bee hive metaphor comprising of dance floor, an auditorium and dispatch rooms two simple model that describes the process of web search, in 2007 Navarat used the bee hive metaphor for online search of users predefined group of pages, authors claim that the hive determines the best routes of search and reject bad ones by experiments reported in the paper [6, 9].

6- The proposed system

The proposed system (bee crawling system) is built using VC++, and a relational database management system to store web pages, we can use any relational database like Microsoft Access , the user interface of the crawling system contains four buttons (crawl a single page, crawl a single page using bee swarm algorithm ,crawl the web from Google directory, crawl using breadth first search), the most important button is the second one (crawl a single site using bee swarm algorithm, and crawl web from Google directory) because these two buttons work depending on bee swarm algorithm , the crawl using breadth first search was putted for comparison between the traditional approach and the new proposed one see figure 2 .

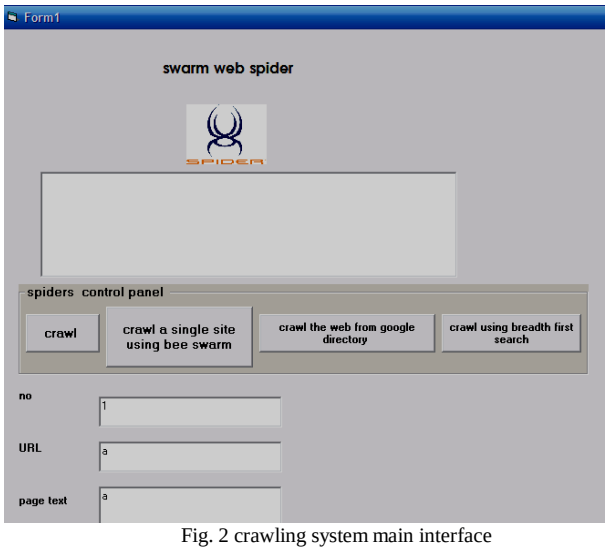

Fig. 2 crawling system main interface

The crawling algorithm depends on bee swarm intelligence algorithm as mentioned above, so the following algorithm illustrates the bee swarm crawling work:

I-start

II-initialize a crawling function by entering a seed URL to it

III-fetch the URL and save it's URL keywords and page text in the database table called(web data)
IV-find the URL's(links)to another pages
V-calculate the link relevancy using the following equation (page rank equation)
PR(A)=(I-D)+D(PR(T1)/C(T1)+…+PR(Tn)/C(Tn))
VI-for each link in the queue
VII-if the link rank>0.7 then initialize a new function (bee) to crawl this link goto I
VIII-end for
XI-END
Where PR(A) is the page rank of page A, PR(T1) is the page rank of page T1, C(T1) is the number of outgoing links from th e page T1 , d is a damping factor in the range 0<d<1 , usually 0.85,the page rank of the web page is calculated as a sum of the page ranks of all pages linking to it (it's incoming links), divided by the number of links on each those page (outgoing links).

7-the experimental results

in order to prove the efficiency of the proposed crawler algorithm "bee crawling algorithm", we have test the system to crawl a single site using the proposed "beed algorithm" and "breadth first search", for this test we have choose for example five sites to be crawled see figure 3

www.toyota.com
www.sony.com
www.freecomputerbooks.com
www.amazon.com
www.developnew.com/books

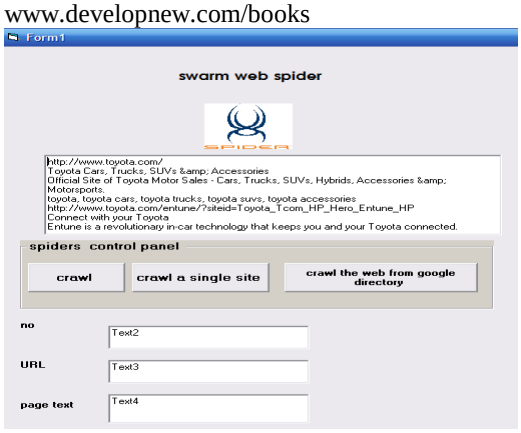

Fig. 3 system interface after crawling Toyota web site

These sites have been crawled using the proposed bee algorithm and breadth first search and the result for crawling speed for each sit was as follows

Web site	Time of breadth search crawling	Time of bee algorithm crawling
www.toyota.com	20 sec	15sec
www.sony.com	30 sec	20 sec
www.freecomputeoks.com	40 sec	35 sec
www.amazon.com	120 sec	100 sec

Table 1: speed of breadth search crawling and proposed bee algorithm

We can understand the results in this chart

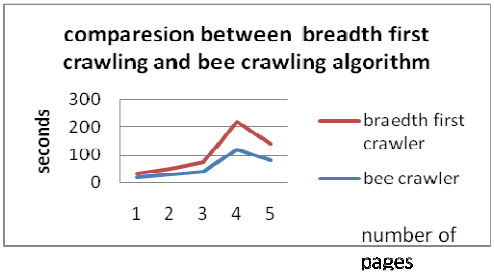

Fig. 4 chart of the results

In the obvious chart the reader can recognize that bee algorithm crawling give a good result in time than the breadth first crawling. Now, the second test is the test of crawled pages relevancy by crawling the web for specific subjects, so the user can input to the system terms like (free, computer, books, and download) with spaces between each term figure 5

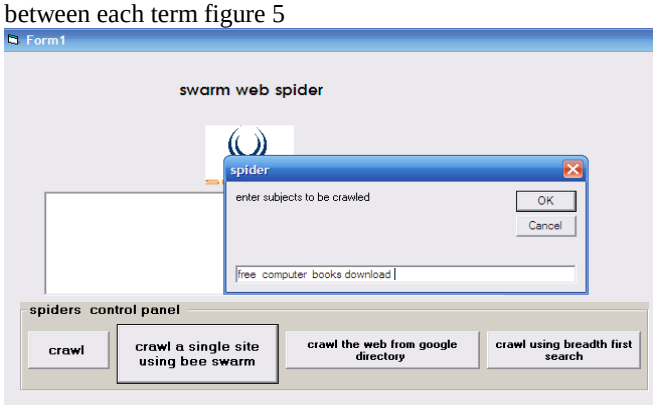

fig. 5 enter specified keywords into the system to be searched and crawled in Web

then press on "crawl the web from Google directory", this system will crawl a sample of 10 sites , now repeat the operation by using the last button "crawl using breadth first search" and also 10 sites will be crawled, see the table

Table2: the time between old and proposed method

Sites crawled using Bee algo.	Sites crawled using BF. Algo.

1-freecomputerbooks.com 2-freebookcenter.com 3-freetechbooks.com 4-appsapps.com 5-techbooksforfee.com	1-overstock.com 2-kobobooks.com 3-amazon.com/stael-this-computer-book 4-eindiabooks.com 5-buy-ebooks.com 6-nextag.com/ebook-reader/stores.html

These sites has been crawled using breadth first search in 30 minutes while in bee algorithm in 17 minutes as the system use a function for each site, the second issue to prove the effiecny of bee swam algorithm is the relevancy of the crawled pages to the user request by measuring the relation of the frequencies of the word in these crawled sites ,this measuring has been done using a statistical tool for word mining called stat miner , so we can see the frequency of the terms and it's related to the subject of these sites which is free e-books see the table below of the frequencies of words in the crawled sites

a.txt - Notepad

File Edit Format View Help

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES
BOOKS	308	7.3%	5.0%	3.8%	1	100.0%
FREE	254	6.0%	4.1%	3.1%	1	100.0%
FREETECHBOOKS	245	5.8%	4.0%	3.0%	1	100.0%
HTTP	203	4.8%	3.3%	2.5%	1	100.0%
WWW	183	4.4%	3.0%	2.3%	1	100.0%
COMPUTER	182	4.3%	2.9%	2.2%	1	100.0%
DEVELOPNEW	145	3.4%	2.3%	1.8%	1	100.0%
ONLINE	116	2.8%	1.9%	1.4%	1	100.0%
SCIENCE	105	2.5%	1.7%	1.3%	1	100.0%
PROGRAMMING	90	2.1%	1.5%	1.1%	1	100.0%
EBOOKS	74	1.8%	1.2%	0.9%	1	100.0%
REGISTER	72	1.7%	1.2%	0.9%	1	100.0%
DOWNLOAD	69	1.6%	1.1%	0.8%	1	100.0%
HTML	69	1.6%	1.1%	0.8%	1	100.0%
LECTURE	61	1.5%	1.0%	0.8%	1	100.0%
NOTES	61	1.5%	1.0%	0.8%	1	100.0%
IGNOU	59	1.4%	1.0%	0.7%	1	100.0%
TEXTBOOKS	58	1.4%	0.9%	0.7%	1	100.0%
DESIGN	54	1.3%	0.9%	0.7%	1	100.0%
WEB	51	1.2%	0.8%	0.6%	1	100.0%
CONTACT	48	1.1%	0.8%	0.6%	1	100.0%
SEARCH	48	1.1%	0.8%	0.6%	1	100.0%
FAQ	45	1.1%	0.7%	0.6%	1	100.0%
NET	45	1.1%	0.7%	0.6%	1	100.0%
RSS	44	1.0%	0.7%	0.5%	1	100.0%
FORM	43	1.0%	0.7%	0.5%	1	100.0%
LOG	42	1.0%	0.7%	0.5%	1	100.0%
FEED	41	1.0%	0.7%	0.5%	1	100.0%
HOME	40	1.0%	0.6%	0.5%	1	100.0%
HOMEPAGE	40	1.0%	0.6%	0.5%	1	100.0%
MEMBERLIST	40	1.0%	0.6%	0.5%	1	100.0%
RESOURCES	40	1.0%	0.6%	0.5%	1	100.0%
READ	35	0.8%	0.6%	0.4%	1	100.0%
JAVA	34	0.8%	0.5%	0.4%	1	100.0%
USER	34	0.8%	0.5%	0.4%	1	100.0%

fig. 6 the words and frequencies in the crawled sites

These results can be explained by the following chart

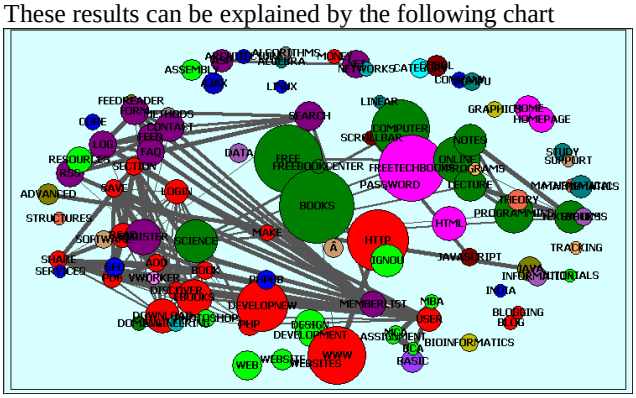

Fig. 7 the relations between keywords in sites crawled using bee algorithm

The reader can recognize from the obvious chart the close relation between terms like "free", "books", "computer", "science", "programming", "notes", and "lecture" , which appears in green color as these words coming together too much, and the relation between "http", "www", and "pdf", "websites", "make", "share", "save", so the brief result will be as the table below

Table3: the term and its frequencies

The term	Frequency
Books	308
Free	254
Free techbooks	245
Computer	182
science	105
ebooks	74
download	72

so it is a good result for "free e books" sites crawled using bee swarm algorithm, then when analyze the web sites crawled by the breadth first search algorithm from web , and analyze the relation between terms in these sites . the reader can recognijze that these sites don't represent strongly "free computer e books"web sites , see the figure

sample3.txt - Notepad

File Edit Format View Help

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF * IDF
OVERSTOCK	449	7.2%	5.5%	4.4%	1	100.0%	0.0
HARDWARE	307	4.9%	3.8%	3.0%	1	100.0%	0.0
COMPUTERS	269	4.3%	3.3%	2.6%	1	100.0%	0.0
SOFTWARE	261	4.2%	3.2%	2.5%	1	100.0%	0.0
ELECTRONICS	213	3.4%	2.6%	2.1%	1	100.0%	0.0
HTTP	189	3.0%	2.3%	1.8%	1	100.0%	0.0
WWW	187	3.0%	2.3%	1.8%	1	100.0%	0.0
HTML	173	2.8%	2.1%	1.7%	1	100.0%	0.0
COMPUTER	159	2.6%	2.0%	1.5%	1	100.0%	0.0
SHIPPING	152	2.4%	1.9%	1.5%	1	100.0%	0.0
DEPT	130	2.1%	1.6%	1.3%	1	100.0%	0.0
ONLINE	111	1.8%	1.4%	1.1%	1	100.0%	0.0
CARS	95	1.5%	1.2%	0.9%	1	100.0%	0.0
SHOPPING	95	1.5%	1.2%	0.9%	1	100.0%	0.0
BUY	93	1.5%	1.1%	0.9%	1	100.0%	0.0
TRAVEL	92	1.5%	1.1%	0.9%	1	100.0%	0.0
INSURANCE	91	1.5%	1.1%	0.9%	1	100.0%	0.0
INFO	90	1.4%	1.1%	0.9%	1	100.0%	0.0
ACCESSORIES	86	1.4%	1.1%	0.8%	1	100.0%	0.0
PRICES	86	1.4%	1.1%	0.8%	1	100.0%	0.0
FACEBOOK	82	1.3%	1.0%	0.8%	1	100.0%	0.0
TWITTER	82	1.3%	1.0%	0.8%	1	100.0%	0.0
COMMUNITY	81	1.3%	1.0%	0.8%	1	100.0%	0.0
GIFTS	81	1.3%	1.0%	0.8%	1	100.0%	0.0
PINTEREST	80	1.3%	1.0%	0.8%	1	100.0%	0.0
LIST	79	1.3%	1.0%	0.8%	1	100.0%	0.0
OMAIL	79	1.3%	1.0%	0.8%	1	100.0%	0.0
AK	77	1.2%	1.0%	0.7%	1	100.0%	0.0
ENTIRE	77	1.2%	1.0%	0.7%	1	100.0%	0.0
EXCLUDES	77	1.2%	1.0%	0.7%	1	100.0%	0.0
INT	77	1.2%	1.0%	0.7%	1	100.0%	0.0
ORDER	77	1.2%	1.0%	0.7%	1	100.0%	0.0
REVIEWS	74	1.2%	0.9%	0.7%	1	100.0%	0.0
DISCOUNT	73	1.2%	0.9%	0.7%	1	100.0%	0.0
EVERYDAY	73	1.2%	0.9%	0.7%	1	100.0%	0.0

fig. 8 keyword frequency for web sites crawled by using breadth first algorithm

The reader can observe the frequency of words from the crawled sites as below

Table4: terms and its frequencies

The term	Frequency
HARDWARE	307
COMPUTERS	269
SOFTWARE	261
SHIPPING	152
ONLINE	111
SHOPPING	95
BUY	93
EBOOK	11

so the web sites retrieved by breadth first search algorithm have not the same relation to the “free computer site” like the retrieved by bee swarm algorithm , see the chart with the full analyzing for the retrieved sample below

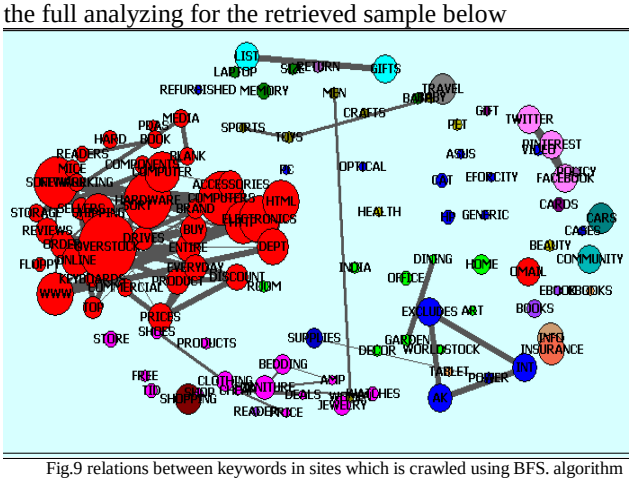

Fig.9 relations between keywords in sites which is crawled using BFS. algorithm

Lets us see the clustering analyses below

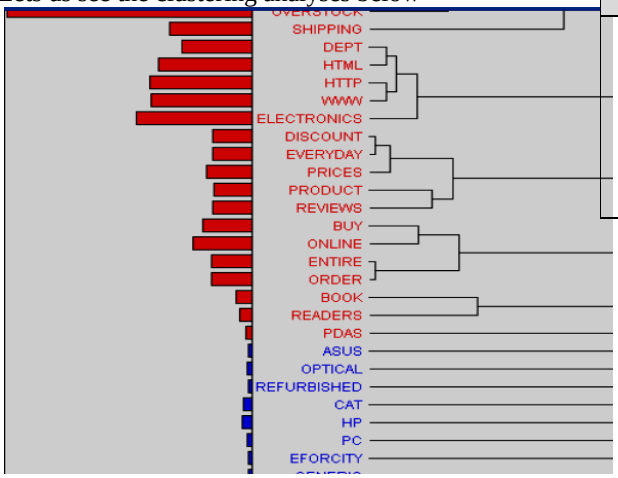

Fig. 10 clustering analysis for keywords in site crawled using breadth first algorithm

Now the second test for another terms “computer networks and books”, we try to use the system to crawl the web sites for computer networks books , because this term cause ambiguity to any searching system as it’s includes the computer network hardware and computer networking books , first we try the traditional breadth first search algorithm and second using the proposed swarm bee algorithm , so enter the term to the system as below

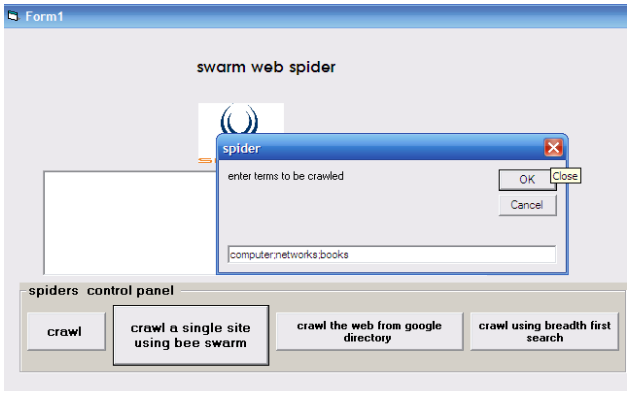

Fig. 11 enter “computer”,”network”,”books” into the system

Then press on “crawl the web from directory”, this system will crawl a sample of 10 sites, repeat the operation but using the last button “crawl using breadth first search “ and also 10 sites will be crawled

Table 5: sites crawled using the traditional method and proposed one

Sites crawled using bee algorithm	Sites crawled by breadth first algorithm
• http://compnetworking.about.com.h • http://www.amazon.com/ • http://www.target.com/ • http://freecomputerbooks.com • http://more-networking.blogspot.co • http://forums.about.com • http://linux.about.com	• http://www.cengage.com • http://www.gaebler.com • http://wps.prenhall.com • http://www.computernetworksinc.com/ • http://www.computercompany.net • http://www.cnsyn.com/ • http://www.cisco.com

To understand the relations between the keywords for the sites crawled using bee swarm algorithm see the chart below

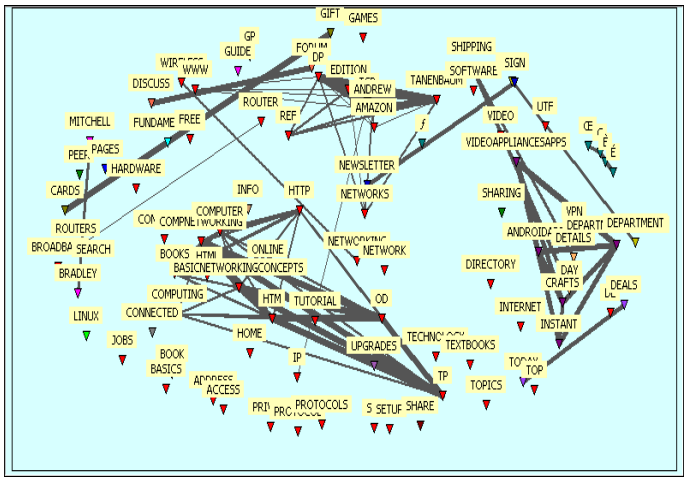

Fig. 12 the relations between words in sites crawled by bee algorithm

The reader can recognize the following frequencies for essential keywords

Table 6: keywords and it's frequencies

Keyword	frequency
Computer	137
Networking	274
Books	179
Hardware	11

So the higher rank for computer networking books as the reader can see from the table above so the accuracy in crawling using bee algorithm is more the accuracy using breadth first search algorithm you can see also good relation between "computer networking" and books and tutorial as you see in the figure above

8-Conclusion

this research reach to a good result when using bee swarm algorithm for crawling a single web site in speed, and also we have good results in crawling the web using keywords for a group of sites in sped and relevancy of the crawled sites to the desired topics, so the a swarm intelligence can be viewed as a good improvement in web crawling area.

Refrences

[1]Sandeep Sharma,"web-crawling approachesin search ",thapar university, Patiala, India, 2008.
[2]Christopher oilstone and Mark najork, "web crawling", yahoo research, 2010.
[3]Mrco dorigo and Mauro birattari, "swarm intelligence", http://www.scholarpedia.org/article/swarm_intelligence,2007
[4]Gianni di carro, " introduction to swarm intelligence", http://www.idsia.ch/~gianni/lectures/swarm-part-1.pdf
[5]Mark Fleischer, "foundation of swarm intelligence", Institute of system research, university of Maryland ,2005.
[6]Dervis karaboga and Bahriya akay, "a comparative study of artificial bee colony algorithm", ELSEVIER, www.elsevier.com/locate/amc, 2009.
[7]Wenping Zou, Younlong Zhu, Hanning Chen, Zhu Zhu , "cooperative approaches to artificial bee colony algorithm" , international conference on computer application and system modeleing, 2010.
[8]Panta lucic and dusan teodorovic, 'computing with bees" Virginia polytechnic institute and state university ,2003.
[9]dervis karaboga and bahriye akay, "a survey: algorithms simulating bee swarm intelligence" , springer science and business media,2009.

First Author Muhamed I. Shujaa is a lecturer in the technical college of management in Iraq, has Bsc in Electrical and Computer in engineering in 1996 American university politehnical Bucharest, PhD in computer engineering 2003 at the same university and participated in many conferences and workshops about informatics inside and outside Iraq. Lecturer in informatics dept. technical college Baghdad 2004-2006. Lecturer Syrian university of technology 2006-2009.lecturer informatics dept,2010 till now. Teaching Networks,Iss,C++.

Secound Auther Ahmed bahaa aldeen is an assistance lecturer in the technical college of management in Iraq, has Bsc in computer science in 2001 from technological university and Msc. In computer science from Iraqi commission for computers and informatics in 2005, participate in many conferences and workshops about informatics inside Iraq, from 2006 till now teaching algorithms and databases in the technical college.