International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 1, January 2016

A Brief Review On The Application Of Swarm Intelligence To Web Information Retrieval

[1] Ramya C, [2] Dr. Shreedhara K S
[1] Research Scholar, [2] Professor
Department of Computer Science and Engineering,
U.B.D.T. C. E, Davangere, Karnataka, INDIA
[1]cramyac@gmail.com [2]ks_shreedhara@yahoo.com

Abstract: **Web Information Retrieval process has become one of the most focused research paradigms because of large quantity of growing web data as internet is ubiquitous. To this distributed, uncertain and volatile data, accurate and speed access is required. So there is a need to optimize the search process using some efficient approaches. For such novel approach a literature survey is presented on evolutionary bio-inspired Swarm Intelligence techniques to optimize search process in Web Information Retrieval Systems.**

Keywords: **web information retrieval, swarm, PSO, ACO, BSO**

I. INTRODUCTION

When the user issues a query for some information it is the responsibility of the Web Information Retrieval (WIR) system to provide relevant and most recent information to the user. Since the Internet users are more and is used by everybody frequently, there is a huge accumulation of web data. It is not easy to search information from such a vast collection of web documents available on WWW. As the web page around the world is increasing day by day, the need of search engines has also emerged. WIR system is facing so many challenges in handling this vast collection of web data. Few are mentioned in section 2. So there exist a need for optimization of search process of WIR system.

The main purpose of this paper is to deal with swarm technologies which are applied to optimize the search process. A deep survey on such works is done and is given in section. 4. During survey it is observed that swarm approaches to search process dominate the already existing traditional approaches like exact and exhaustive algorithms. Swarm methods are more suitable for large scale web data. They provide statistical results about the searching process. They are able to obtain the most frequently used words on time distinguishable web pages and to retrieve top ranked documents. More over they are simpler and easy to implement. Table. 1 provides the comparison of some of the considerable works done recently. Section.3 provides a brief look at the popular swarm approaches.

II. WEB INFORMATION RETRIEVAL

As the WWW is having more than three billion of web documents , it is called to be a huge tremendous repository. These documents are searched on the web using search engines. Search engines provide the documents to the user based on the specified set of keywords in the query. They return a list of documents where any or all of the specified keywords in the query were found. So the process of retrieving information from the web is called to be WIR.

WIR System is facing a huge challenge due to vast amount of information stored on the web. This data may be inconsistent, distributed, incorrect and imprecise. The users are not familiar with this system and their information seeking varies often. The web data stored is distributed [9]. Because of this, search engines sometimes fail to retrieve more relevant documents for the user queries rather they are fetching irrelevant web documents. So there is a need for the optimization of this search process in WIR system. Many research works are going on to optimize the search process using the technologies like Artificial Intelligence, Symbolic Learning [14], neural networks, Genetic Algorithms [13] and Swarm Intelligence. Recent works in the literature showed superior and considerable performance of swarm technology in context of optimization of WIR process.

III. SWARM INTELLIGENCE

Swarm Intelligence (SI) is a technology taken inspiration from the behaviors observed in flocks of birds, schools of fish, or swarms of bees. SI is best suited for optimization problems like TSP, quadratic assignment, graph coloring, optimization, network routing, cluster finding, job scheduling, search engines, load balancing, etc.. Fig. 1 shows the natural forms of SI Techniques.

Particle Swarm Optimization (PSO) is inspired from the swarming behaviors observed in birds. PSO is a population-based optimization tool which could be implemented and applied easily to solve various function optimization problems. The main strength of PSO as an algorithm is its fast convergence. For applying PSO successfully, one of the key issues is finding how to map the problem solution into the PSO particle, which directly affects its feasibility and performance [15]. Ant Colony Optimization (ACO)

incorporates the foraging behavior of real ants which are used to solve distinct optimization problems. The main idea is the indirect communication between the ants by means of chemical substance pheromone, which enables them to find short paths between their nest and food. The Bee Colony Optimization (BCO) is inspired by bee's behavior in the nature. The basic idea behind BCO is to create the colony of artificial bees capable of successfully solving difficult optimization problems. The survey on how swarm technology has been applied to solve the optimization problem of WIR process is briefly presented.

Table 1. Comparison of various Swarm Optimization based WIR methods

SL No.	Paper Referred	Data sets Used	Algo./Tech	Purpose	Similarity Function Used	Pros
1	Dr. Hasanen S. Abdullah & Mustafa J. H. [4]	CACM and NPL collections.	Artificial Bee Colony algorithm	To cope with the complexity induced by the huge volume of information on web	cosine formula $f(d,q)=\frac{\sum i\,(a_i * b_i)}{\sqrt{\sum i\,(a_i)^2 * \sum i\,(b_i)^2}}$	1. More efficient than exhaustive search And exhibits superiority over classic approaches 2. System is more suitable to large scale IR
2.	Pavol Navrat and Anna Bou Ezzeddine [8]	Web pages from: www.pravda.sk, www.sme.sk and www.ta3.com.	Modified Bee Hive Model	To search parts of the web online.	To Find Incidence number quality: $q_{count}=\frac{-1}{2\left(n+\frac{1}{2Q_{COUNT}}\right)}+Q_{COUNT}$	1. It provides statistical results about the searching process 2. Able to obtain the most frequently used words on time distinguishable web pages
3.	Priya I. Borkar and Leena H. Patil [6]	Database located at the local server	Hybrid Genetic Algorithm -PSO	To reformulate a user query to improve the results of the corresponding search.	Jaccard coefficient: $\mathrm{Sin}\,(x,y)=\lvert x\cap y\rvert \div \lvert x\cup y\rvert$	1. Top ranked documents are retrieved
4	Habiba Drias [1]	CACM and RCV1 collections	Parallel PSO	To validate direct search methods are more suited & simpler to implement than the other	$f(d,\hat{q})=\sum_i (\hat{a_i} * b_i)$	1. exhibits the robustness & superiority of parallel PSO on all the others in terms of scalability while yielding comparable quality. 2. Simpler to implement
5	Peiyu Liu, Zhenfang Zhu and Lina Zhao [3]	Experimental data are downloaded from the Internet	Clustering principle based on Ant-foraging	To make a more precise and rapid clustering to increase the speed and efficiency of information retrieval	Density of similarity: $f(a_i)=\max\left(0,\frac{1}{s^2}\sum_{a_i\in N_{s\times s}(x,y)}\left[1-\frac{d(a_i,a_j)}{\alpha(1+\frac{(v-1)}{v_{max}})}\right]\right)$	1. the results of clustering algorithms have more accuracy and accelerate the cluster speed
6	Anna Bou Ezzeddine [2]	Online search to track current events like earthquakes, floods etc.	hierarchical Bee Hive Model	To retrieve information and also to trace story that is developing on the Web on-line.	To find quality: $q_{count}=\frac{-1}{2\left(n+\frac{1}{2Q_{COUNT}}\right)}+Q_{COUNT}$	1. Able to obtain the most frequently used words on time distinguishable Web pages 2. provides statistical results about the searching process

IV. RELATED WORKS

Habiba Drias [1] showed the designing of two novel PSO algorithms as search techniques for information retrieval on the web. Here the PSO is parallelized by executing the code independently and in parallel to each particle said in the algorithm. Due to parallelism, the PSO threads will return a high similarity of the best found document with a less runtime as the best solution. When the

collection is very high, the parallelism achieves good scalability. It is shown through experimental results that parallel PSO outperforms all the other heuristic search methods [10][11][12] like exact algorithm available in the literature.

Anna Bou Ezzeddine [2] proposed a upgraded bee hive model inspired by bee swarm behaviour for retrieving information from the web. An adapted model in the form of bees is used to trace the story that is developing on the Web on-line. A hierarchical interconnection among several bee hive models is proposed to obtain more number of high quality documents. The model performs story tracking with the aim of finding a relevant set of of pages which would form together a story. It was supposed to be used on any sites containing frequently updated or added information but not as a general search engine. By this method, a high quality searching of information with a less improved speed can be achieved.

Peiyu Liu, Zhenfang Zhu and Lina Zhao [3] introduced ant-based clustering and sorting to increase the speed and efficiency of information retrieval. This was compared with the incremental learning (IL) information retrieval. Experimental results showed that this cluster analysis based on the ant heap principle provides more precise and rapid clustering than IL based one.

Dr. Hasanen S. Abdullah and Mustafa J. Hadi [4] used Artificial Bee Colony (ABC) algorithm with the aim of addressing information retrieval with the huge volume of information in terms of response time and good solution quality. The comparison with classic approach in terms of response time and document quality showed explicitly less efficiency of the classic approach. However this work is inspired by [5] where the adaptation of heuristic search techniques to large scale IR and their comparison with classical approaches can be seen.

Priya I. Borkar and Leena H. Patil [6] presented a model of hybrid Genetic Algorithm -Particle Swarm Optimization (HGAPSO) to produce the new keywords that are related to the user search. The jaccard similarity function is used to find the fitness value of each document. Document with high fitness value will be picked in the selection operation. Thus the evolutionary algorithm was used to reformulate a user query to improve the results of the corresponding search.

Sridevi U.K and Nagaveni N [7] proposed a ontology based annotation method to improve the retrieval. The quality of the solution obtained can be improved by using annotated weights and optimized clustering algorithm. The particle swarm optimization clustering algorithm is to discover the proper centroids of clusters for minimizing the intra-cluster distance as well as maximizing the distance between clusters. PSO algorithm can generate more compact clustering results. The Fuzzy Particle Swarm Optimization algorithm is a hybrid method developed in order to combine the properties of fuzzy clustering and Particle Swarm Optimization.

Pavol Navrat and Anna Bou Ezzeddine [8] developed a modified bee hive model for tracking story, to collect relevant web pages, to reconstruct the story backwards and to monitor the story that is developed during the search, there by searching pages of the web online. It is showed that it's possible to obtain the most frequently used words on time distinguishable web pages by the system.

V. CONCLUSION

This paper has presented a survey on the research work done on WIR based on swarm optimization techniques. This survey contains a brief introduction about web information retrieval and swarm techniques and explored various research papers related WIR using swarm approaches. More research works have to be carried out based on semantic to improve the quality of search process.

REFERENCES

[1] Habiba Drias, "Parallel Swarm Optimization for Web Information Retrieval", in proceedings of Third World Congress on Nature and Biologically Inspired computing, pp. 249-254, 2011.

[2] Anna Bou Ezzeddine, " Web information retrieval inspired by social insect behaviour", Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 3, No. 1, pp. 93-100,2011.

[3] Peiyu Liu, Zhenfang Zhu and Lina Zhao, "Research on Information Retrieval System Based on Ant Clustering Algorithm", Journal of Software, Vol. 4, No. 9, pp. 1032-1036, 2009

[4] Dr. Hasanen S. Abdullah and Mustafa J. Hadi, "Artificial Bee Colony based Approach for Web Information Retrieval", Eng. & Tech. Journal, vol.32, Part (B), No. 5, pp. 899-909, 2014.

[5] Habiba Drias and Hadia Mosteghanemi, "Bees Swarm Optimization based Approach for Web Information Retrieval", In proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 6-13, 2010.

[6] Priya I. Borkar and Leena H. Patil, "Web Information Retrieval Using Genetic Algorithm-Particle Swarm Optimization", International Journal of Future Computer and Communication, Vol. 2, No. 6, 2013.

[7] Sridevi U.K and Nagaveni N, "Ontology based Optimization Techniques for Information Retrieval", A The is from http://shodhganga.inflibnet.ac.in/handle/10603/15038, 2012.

[8] Pavol Navrat and Anna Bou Ezzeddine, "Bee Hive at Work: Following A Developing Story on The Web", Max Bramer. Artificial Intelligence in Theory and Practice III, 331, Springer, pp.187-196, 2010.

[9] Shruti Kohli and Ankit Gupta, "A Survey on Web Information Retrieval Inside Fuzzy Framework", Proceedings of the Third International Conference on Soft Computing for Problem Solving, pp. 433-445, 2014.

[10] R. Baeza-Yates and B. Ribiero-Neto, "Modern Information Retrieval", Addison Wesley Longman Publishing Co. Inc., 1999.

[11] C.D. Manning, P. Raghavan and H. Schutze, "Introduction to Information Retrieval", Cambridge University Press, 2008.

[12] J. Kennedy and R.C. Eberhart, " Particle Swarm Optimization", In Proceedings of the IEEE Int. Conf. On Neural Networks, Piscataway, NJ, pp. 1942-1948, 1995.

[13] P. Pathak, M. Gordon and W. Fan, "Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaptation," 33rd IEEE HICSS, 2000.

[14] C. Hsinchun, "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning and Genetic Algorithm", Journal of the American Society for Information Science, pp. 194-216, 1995.

[15] A. Abraham, "Swarm Intelligence: Foundations, Perspectives and Applications", Studies in Computational Intelligence (SCI) 26, pp. 3–25, 2006.