A Survey of Bio Inspired Algorithms for Web Information Extraction and Optimization for Big Data Analytics

Manjunatha Swamy C, S. Meenakshi Sundaram

Abstract: Information extraction is systematic process of extracting structured information from documents which has both unstructured and semi structured data set. Data available over the web is unstructured which is processed and delivered that may be challenging due to massive data over web. Bigdata analytics approach is used in the computation field where massive data is managed and processed as information. Data from various sources like industries, institutes are processed using algorithms in efficient means employing web of things or Internet of things used to mine such a large data. Bio inspired algorithms have evolved from application of heuristic approaches to meta-heuristic and hyper-heuristic methodologies. Bio inspired techniques are categorized into human inspired algorithms, Swarm Intelligence algorithms, evolutionary algorithms and ecology based algorithms. Genetic algorithms are purely heuristic in nature and are employed for computation and extracting information and from big data. This improves the computation speed effectively for extracting web related information as evolutionary algorithm resolves information extraction problems. The Ant colony and Particle Swarm Intelligence algorithms are of meta-heuristic in nature. The Cuckoo search, Artificial Bee Colony, Firefly algorithm and Bat algorithms are of hyper heuristic in nature i.e., they employ a combination of methods. Web information extraction using bio inspired concepts and genetic operators increases efficiency, capability to search particular information in massive data in web. Some of the tools that are available for data extraction and mining are DataMelt, Apache Mahout, Weka, Orange and Rapid Miner for enhancing web data extraction efficiency. This survey on bio inspired methodologies can be extended to parameter tuning and controlling is another big strategy that can be implemented, in addition to convergence speed up.

Keywords: Web Information Extraction, Big Data Analytics, Genetic Algorithm (GA), Bat Algorithm (BA)

I. INTRODUCTION

Web information extraction in massive warehouse is not easy. Information collected from all variety of sources is in digital form, which may be structured, unstructured or semi structured. Hence algorithms are used to process data and extract information. Genetic Algorithm [5] is an evolutionary algorithm which consists of functions that represent data as a model to handle efficiency and scalability issues, with fitness function that allows selecting particular information in the data set. Data is evaluated to get information. Many bio-inspired concepts are used to handle the information in a systematic way like Ant Colony, Honey Bee [11], Bat Algorithm[10], Wolf Optimization concepts. Big Data Analysis and extraction of particular information using these approaches are the current trends. Genetic Algorithm [5] is a part of evolutionary algorithm which eases the search in huge data sets. It is suitable to implement and model an approach to data analytics with operators which enhance flexibility and functionality in processing information. Genetic Algorithm uses change in the method of searching. Raw data generated from various sources are collected by interfaces which initialize big data processing, as data generated in unstructured in all sources that need to process as information. To execute this, different key components are used as shown in Figure 1 given below. A selection operator in Genetic Algorithm does proper comparison and selection to enable information extraction. Pre-processing is by selecting a proper model to extract relevant information. Model evaluation is carried out to validate information for further analysis. Web data is very important that is collected from various sources will be pre-processed in a proper way to filter outliers or unwanted information. Later it will be gartered if necessary and data is transformed to other forms like files. Integration of various information, later extraction of information is done. Pattern evaluation also an important role played in process to verify and validate information as it will meet the requirements of users. To improve the performance of data and to extract information, many software tools are available in the market that can be used but the methodologies adopted may differ. Bigdata Analytics concept is applied to applications like health care systems for data extraction and optimization [3]. This enables to develop new algorithms to address these issues more effectively. Big data information collected from various data sources is stored in a data center. For extraction of meaningful information without any modifications, data mining is used to examine or to inspect large amount of data collected earlier. Using tools [8] which help in extraction of data, extraction of predictive information from large database is called as data warehouse

II. BIOINSPPIRED OPTIMIZATION ALGORITHMS

A. Optimization

Revised Manuscript Received on December 05, 2020.
* Correspondence Author

Manjunatha Swamy C*, CSE, GSSSIETW, Mysore, India. Email: drsmsmsc@gmail.com

Dr. S. Meenakshi Sundaram, CSE, GSSSIETW, Mysore, India. Email: 1965sms@gmail.com

Optimization problems [6][9] based on krill herd behavior in swarm with intensification and diversification, where intensification is searching in around and to select the best one whereas diversification is based on individual behavior and fitness.

Distance from the food and depth movement induced from other individuals, activity and diffusion as well. Krill density and way of reaching food solves global optimization problems effectively. Table 1 given below gives a comparison of the different bio-inspired algorithms that can be suitably employed for big data analytics

TABLE- 1: COMPARISON OF DIFFERENT BIO INSPIRED ALGORITHMS

Sl.no	Algorithm	Concept	Advantages	Limitations
1	Genetic Algorithm Approach, Heuristic scheduling algorithm	Generation set, Population size, crossover rate, mutation rate	reduces processing time , setup time	Identifying particular set of population size is poor
2	Deterministic algorithm, Stochastic algorithm Swarm algorithm Animal behaviour	Particular input with output	works better with same sequence of sets	Pseudo random number used is not sufficient
3	Cukoo search, Chicken swarm,	Behavior	to find best way to schedule task	Easily fallen to local optimal solution with low rate of convergence
4	Bat algorithm	Frequency, Pulse rate generated	Intensificant and diversification to get better relation, high mobility, diversity	Converge very quickly and later its rate is slow Accuracy is limited
5	Bacteria behavior algorithm	prototype developed for evaluation	Heuristic approach as best alternative for getting optimal solution …Solve symmetric problems	focus on Intensification, Diversification area need to be improved
6	Wolf optimization (For leadership), Grey wolf algorithm	Hunting Mechanism as Global optimization concept	Avoid local optima, High performance in unknown, Challenging search space	Local searching ability is less, Slow convergence rate, Low accuracy
7	Elephant search optimization algorithm (hybrid evolutionary algorithm)	2 factor used Fitness values as output, Fitness values as output, Stability	Stable performance, Fast Solution	Fitness value can be improved Searching in local space is low accuracy
8	Kmean, Evolutionary algorithms	Accuracy raet, precision, Recall and F-measure	Feasibility, Efficiency, Intelligence, Parallel PSO	Distance factor
9	Pittsburgh learnt fuzzy rule base for feature subset selection, , PSO algorithm	temporal complexity optimization, attribute type support expansion and evaluation of the possible influences of indifference threshold values.	pre-process the data, by selecting the relevant subset of features,	change in position function, change in velocity function is minimal

10	HITS, TF, Genetic algorithm, Thematic search algorithm,	Finds best search localization using term concept, frequency of repetition	HITS address many issues with searching frequent topics	HITS having low accuracy can be addressed

III. RELATED WORK

Data available over the web is unstructured which is processed and delivered that may be challenging due to massive data over web. Bigdata analytics approach is used in the computation field where massive data is managed and processed as information. Data from various sources like industries, institutes are processed using algorithms in efficient means employing web of things or Internet of things used to mine such a large data. Any improper extraction, partial information cases genetic algorithm strategy can be deployed to improve the efficiency. Figure 3.1 given below gives the suitability of Genetic Algorithm to extract web information which uses big data.

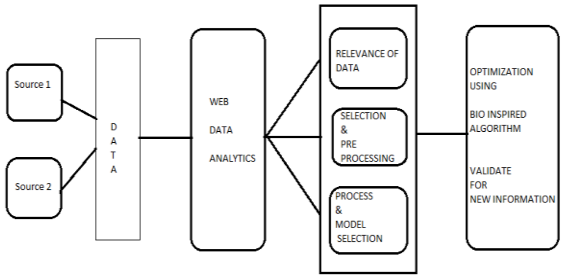

Fig 3.1: Big data Analytics using Genetic Algorithm [5] to extract web information

A. Bat Algorithm

This algorithm is based on behavior of bats in the eco system to search its food, which emits pulse rate and loudness based on availability of its prey. Rate of pulse in frequency is increased to obtain accuracy. Bats[15] have flexibility in adjusting the wavelength and frequency to find its prey effectively.

B. Wolf Optimization Algorithm

This algorithm mainly focuses on leadership and hunting techniques of wolf like searching for prey based on a social structure and attack towards it prey. Global search algorithm [9] ensures best optimization techniques altogether to protect as well to search best to find its prey and to attack.

C. Elephant Search algorithm

Meta heuristic concepts are based on evolutionary concepts. In this searching, the elephants indicates iteratively refining solutions itself, local search is taken up by the female elephant[12] to find best, other male elephants can act as rangers. Some elephants live in a group and others in remote spaces where the visual ability has Euclidean distance concept. The male elephants have better vision compared to female elephants. Fitness in the group is maintained by other elephants with stability as well.

IV. WEB INFORMATION EXTRACTION AND OPTIMIZATION USING BAT ALGORITHM AND GENETIC ALGORITHM

Bat algorithm [10] is a Meta heuristic algorithm used in Optimization Algorithm categorized under Bio inspired behavior concept in bats. Many species of bats use sonar based location tracing in finding its prey. Compared to mini bats and mega bats, micro bats extensively use this strategy to find prey, to find or avoid obstacles in its path, to move in the dark environment. By generating various sound sequences in the range of 25 KHz to 150 KHz 8 to 10 mSec.

A. Bat Algorithm for web Information Extraction

The flow chart for implementation of Bat Algorithm is given in figure 3.2 below.

step 1: Assign position, velocity and parameters with frequency

step 2: Increment velocity and location each time based on equation

step 3: Increment position and find fitness value in locating prey based on distance

step 4: Based on loudness factor and pulse rate identify or avoid obstacles to get best position

step 5: algorithm is done and go back to step 2

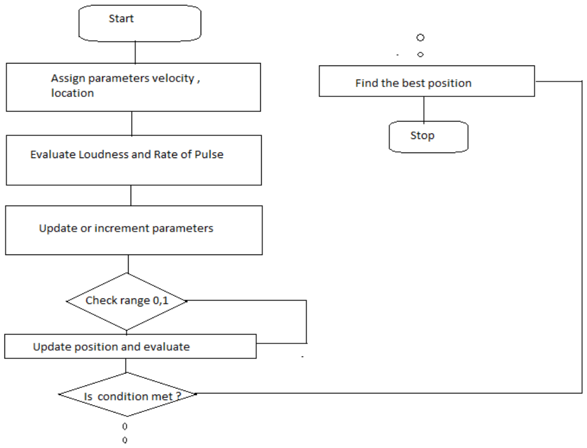
Start
Assign parameters velocity , location
Evaluate Loudness and Rate of Pulse
Update or increment parameters
Check range 0,1
Update position and evaluate
Is condition met ?
Find the best position
Stop

Fig 3.1: Flowchart for Bat algorithm

Loudness and pulse rate [15] emission factors used with frequency range to formulate equations

$a^{t+1} = B(a^t, k(t))$

B is non linear mapping

p(t) = (p1, p2, p3…..pn)

Time sequences for the algorithm

Initial state is denoted with Q

$Q(\chi) \to Q(\psi)$

To get optimal solution modified equation is as below

$$Q(\chi) \xrightarrow{B(t, D, C)} Q(\psi)(a^{**})$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

where B is an algorithm factor, D criteria factor,
C is behavior controller
Q(ψ) final converted state of solution

Bat Movement:

Based on velocity (v) at different locations y_i^t with particular time (t), frequency ranges from [0,1] with distance. Loudness factor and rate of pulse emission bat can do its search of prey. When loudness is zero then its prey is found and pulse will be set to zero

$$f_i = f_{min} + (f_{max} - f_{min})\Upsilon$$

$$V_i^t = V_i^{t-1} + (x_i^{t-1} - x^*) f_i$$

Si is the loudness, Ri is rate of pulse emission

$$S_i^{t+1} = \acute{y} S_i^t$$

$$R_i^{t+1} = R_i^o[1- \exp(\,^{-}\Upsilon t)]$$

B. Genetic algorithm for web Information Extraction

Big data Analytics Algorithm [7] with genetic operators [5] can build the new architecture where it resolves many Web Data extraction issues. Architecture is software and collection of concepts which is a solution later to the entire data mining system. Traditional approach of extraction is no more effective with huge data set with multiple sources. Genetic Algorithm is designed with fitness function, selection operator, mutation operator, cross over and initialize concept which have ability to provide effective flexible and reliable mode to meet different criteria of extraction. Figure 3.3 given below shows the procedure for Web extraction using Genetic Algorithm. Web information [1][2] extraction of information, traffic and routing management, optimization is application where bioinspired and genetic algorithms will help to build a model. Fitness function of GA which gives maximum accuracy of relevance. Accuracy flexibility in addressing data of N – sources with multi behavior to breakdown data with proper optimization. Model is new concept with algorithm notation having multiple sources feature selection, subset evaluation the selected data sets also lead to evolution. Transform the model into algorithm which suits the requirement. Mutation [4] is carried out to create multiple redundant copies to increase availability of information all time. genetic algorithms to design fuzzy systems to provide how to learn and adaptation capabilities through a terminology genetic fuzzy system. How this system influences the computing Intelligence [2] by incorporating fuzzy rule-based system, evolutionary algorithms, machine learning and computation intelligence to mine data in data set and enable system to provide suitable information while retrieving

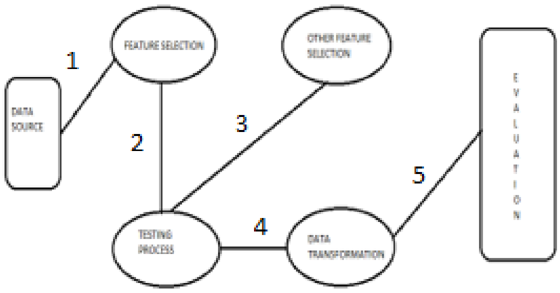

Fig 3.3: Web Extraction using Genetic Algorithm

The algorithm below shows the complete analysis of data processing to get suitable Information starts with raw data which is unstructured and improper. Evaluate the system repeatedly to obtain 'Information' as shown below:

Sep1: Select the data from data center for Initialization.
Step2: Use Fitness () to evaluate suitability or relevancy
Step3: Continue this Fitness evaluation until fitness is obtained
Step4: Allocate rank based on relevancy individually
Step5: Do mutation or recombination to obtain Information
Step6: Recombine the information to find new data model
Step7: Repeat the process by Evaluation
Step8: Stop

V. CONCLUSION & FUTURE WORK

Information extraction is systematic process of extracting structured information from documents which has both unstructured and semi structured data set. Data available over the web is unstructured which is processed and delivered that may be challenging due to massive data over web. Massive data is managed and processed as information using big data analytics. Data from variety of sources are processed using algorithms efficiently employing web of things or Internet of things that are used to mine such a large data. Bio inspired techniques are categorized into human inspired algorithms, Swarm Intelligence algorithms, evolutionary algorithms and ecology based algorithms. Genetic algorithms are purely heuristic in nature and are employed for computation and extracting information and from big data. Bio inspired algorithms have evolved from application of heuristic approaches to meta-heuristic and hyper-heuristic methodologies. The Ant colony and Particle Swarm Intelligence algorithms are of meta-heuristic in nature. The Cuckoo search, Artificial Bee Colony, Firefly algorithm and Bat algorithms are of hyper heuristic in nature i.e., they employ a combination of methods. Any improper extraction, partial information cases genetic algorithm strategy can be deployed to improve the efficiency. For improving the computation speed effectively and for extracting web related information from the survey it is found that the bat algorithm and genetic algorithms are more suitable. Web information extraction using bio inspired concepts and genetic operators increases efficiency, capability to search particular information in massive data in web.

Big data with natural behavior of animal concepts have more research opportunities in the Big data Analytics research area. Bio inspiration drives to learn new models, concepts and algorithms to deduce new methodologies to store data and process it. Many tools are available for data extraction and mining like Data Melt, Apache Mahout, Weka, Orange, Rapid Miner which enhances web data extraction efficiency. Further this survey on bio inspired methodologies can be extended to parameter tuning. Controlling is another big strategy that can be implemented, in addition to convergence speed up.

REFERENCES

1. J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase : A repository of web pages. In Proc. of the 9th Int. World Wide Web Conference, May (2000).
2. Yang Sun, Isaac G. Councill, C. Lee Giles," The Ethicality of Web Crawlers" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology2010.
3. Yang Sun, Isaac G. Councill, C. Lee Giles," The Ethicality of Web Crawlers" 2010
4. S. M. Alaoui, O. Frieder and T. El-Ghazawi, A parallel genetic algorithm for task mapping on parallel machines. Computer Science, 1586, 201-209,2015
5. C. Low and Y. Yeh, Genetic algorithm-based heuristics for an open shop scheduling problem with setup, processing, and removal times separated. Robotics and Computer-Integrated Manufacturing, 25(2), 314-322,2014.
6. G., Zhang, et al., Solving satisfiability problems with membrane algorithms. Proceedings of the Fourth International Conference on Bio-Inspired Computing: Theories and Applieations, 2012, pp.29-36.
7. Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. and Buyya, R. "Big Data computing and clouds: Trends and future directions," Journal of Parallel and Distributed Computing (79), 2015, pp.3—15.
8. Analytics, Revolution. "Packages in RHadoop Toolkit."(2017).
9. M. V. C. d. Silva, N. Nedjah, and L. d. M. Mourelle, "Power-Aware
10. Multi-objective Evolutionary Optimization for Application Mapping on NoC Platforms," in Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, ed, 2014, pp. 143-152.
11. Damasceno, N.C.; Filho, O.G. PI controller optimization for a heat exchanger through metaheuristic Bat Algorithm, Particle Swarm Optimization, Flower Pollination Algorithm and Cuckoo Search Algorithm. IEEE Lat. Am. Trans. 2017, 15, 1801–1807
12. Sun, H.; Wang, K.; Zhao, J.; Yu, X. Artificial bee colony algorithm with improved special centre. Int. J. Comput. Sci. Math. 2017, 7, 548–553.
13. T. N. C. Vidya and R Sukumar, "Social and reproductive behaviour in elephants ", Current Science, Vol. 89, No. 7, 10 October 2005, pp.1200- 1207
14. X.S. Yang, A.H. Gandomi, Bat algorithm: a novel approach for global engineering optimization, Engineering Computation 29 (5) (2012) 464– 483
15. V. Priyadharshini, P. Divya, D. Preethi, N. Pazhaniraja, P. Victer Paul, A novel Web service publishing model based on social spider optimization technique, IEEE InternationalConference on Computation of Power, Energy Information and Communication (ICCPEIC),April 2015, India, pp.0373-0387.
16. Tharakeshwar, T.K.; Seetharamu, K.N.; Prasad, B.D. Multi-objective optimization using bat algorithm for shell and tube heat exchangers. Appl. Therm. Eng. 2017, 110, 1029–1038.

AUTHORS PROFILE

Mr. Manjunatha Swamy C is currently working as Senoir Assistant Professor in the Department of Computer Science and Engineering at New Horizon College of Engineering, Bengaluru. He obtained B.E in CSE form PES College of Engineering, Mandya in 2006 and M. Tech Software Engineering from SJCE Mysuru in 2011. He is pursuing Ph.D at VTU Belagavi. He has eleven years to teaching expereinece and has good industry exposure. His research areas of interest include Bigdata Analytics, Web analytics, Information retrieval system and Cloud Computing. He has published 4 papers in refereed International Journals, presented 1 papers in International Conferences He is a Life Member of Indian Society for Technical Education (ISTE), Life member of Computer Society of India (CSI) and also a Life member of International Association of Engineers (IAENG).

Dr. S. Meenakshi Sundaram is currently working as Professor and Head in the Department of Computer Science and Engineering at GSSS Institute of Engineering and Technology for Women, Mysuru. He obtained his Bachelor Degree in Computer Science & Engineering from Bharathidasan University, Tiruchirappalli in 1989, M.Tech from National Institute of Technology, Tiruchirappalli in 2006 and Ph.D. in Science & Engineering from Anna University Chennai in 2014. He has 30 years of expereince in teaching and 12 years in research. He has published 53 papers in refereed International Journals, presented 3 papers in International Conferences and has delivered more than 40 seminars. He is a reviewer of Springer – Soft Computing Journal, International Journal of Ah Hoc Network Systems, Journal of Engineering Science and Technology, Taylor's University, Malaysia and International Journal of Computational Science & Engineering, Inderscience Publishers, UK. He has organized more than 40 seminars / Workshops / FDPs. He has attended more than 45 Workshops / Seminars. His area of interest includes Computer Networks, Wireless Communication, Software Engineering, Optical Networks and Data Mining. He is a Life Member of Indian Society for Technical Education (ISTE) and a member of Computer Society of India (CSI). He has 30 Years of teaching experience and 10 years of research experience. Currently 7 research scholars are pursuing Ph.D. in VTU Belagavi, India under his guidance.