International Journal of Embedded Systems and Emerging Technologies
Vol. 1: Issue 2
www.journalspub.com

# Comparative Study of Searching Algorithms for Databases and Present New Idea For Better Searching

*Javad Saghi*[*], *Morteza Zahedi*
Department of Artificial Intelligence Engineering, Kerman Science and Research Branch, Islamic Azad University, Kerman, Iran

***Abstract***

*Meta-heuristics encompass a very wide range of algorithms, all having very different underlying principles. But all these algorithms are stochastic in nature. Big data is an emerging field. Big data has some of its own problem, some of them being the optimization problems. This paper gives a brief glimpse of Meta-heuristics and its algorithm, big data and its corresponding fields. In the last section we also explain how these algorithms can be incorporated with big data*.

**Keywords:** algorithms, big data, heuristics

***Corresponding Author***
*E-mail: seyedmousavi@gmx.com*

## METAHEURISTICS

### Introduction

In today's world, most of the problems are complex, non-linear and have a very large solution space. This requires the use of a technique that is capable of solving such problems in real time. Several of these are optimization problems. A technique used to solve an optimization problem is called an optimization technique. An optimization problem in computer science is a problem in which we aim to find a solution that attains a maximum/minimum score in some aspect. When a feasible solution of the problem reaches its maximum/minimum with respect to some quality measure, then it is called an optimal solution. Optimization techniques can be classified as stochastic and deterministic. Stochastic optimization methods are optimization methods that use random variables to decide for control flow. On the other hand, deterministic optimization methods are the ones which involve no randomness. A deterministic model will thus always have the same output from a given starting condition.[1-5]

Another classification for optimization algorithms may be - exact methods and approximation techniques. Exact methods guarantee the optimum solution though the time taken may be enormous for large instances of NP-hard problems, the examples are - branch and bound algorithm, dynamic programming, etc. On the other hand, approximation methods give results in a comparatively less amount of time but the solution may not be the best possible. The approximation techniques can be further classified as approximation algorithms and heuristic methods. Approximate algorithms are those algorithms which find a rough, coarse-grained solution to the optimization problem. These algorithms guarantee a certain quality in the solution but do not guarantee to reach the best, however, using comparatively small time frame. On the contrary, a heuristic search is an approach to problems solving that incorporates the

custom information about the problem; it may not achieve perfect result, but reaches somewhere close to the optimal result in reasonably fast. Within heuristic algorithms, there is a class of algorithms called Meta-heuristics. Meta-heuristics are higher level procedures that may be used for a category of problems.[6,7] The basic structure of the algorithms remains same for entire category but the exact design of the algorithm may be problem specific, again. These algorithms are stochastic in nature and are problem independent in general. This group of algorithm is better than heuristics and iterative methods as they provide better chances to find the global optima. The Meta-heuristics algorithms can be classified along various different dimensions as follows-

a) Single solution v/s population based- A single solution based algorithm focus on modification and improvisation of a single solution, e.g. - tabu search, simulated annealing, etc. Population based algorithms are based on the improvisation of the multiple candidate solutions, e.g. - genetic algorithms, particle swarm optimization, ant colony optimization, etc.
b) Biologically inspired v/s Non-biologically inspired Meta-heuristics- Biologically inspired Meta-heuristics are the ones that are inspired by some biological system, e.g. -Genetic algorithm, ant colony optimization, etc. Non-Biologically inspired Meta-heuristics are the ones which are based on some pure mathematical methods or inspired by some natural phenomena other than biological phenomenon, e.g. - tabu search, simulated annealing, etc.
c) Iterative algorithm v/s greedy algorithm- The first set of algorithm start from a solution or from a set of solution that keeps getting manipulated as the search moves forward, e.g.- particle swarm optimization, simulated annealing. However, the second set of algorithms start from a clear and vacant solution space and with every thread a variable is assigned a value, which is added to the solution space, e.g. -Ant colony optimization.[8-13]
d) Parallel Meta-heuristics- Parallel Meta-heuristics are the ones which uses parallel programming techniques to run multiple Meta-heuristics searches in parallel. In this, various algorithms, of different nature, run simultaneously on the same search space. These algorithms, all work together through some interference mechanism from time to time to improve the solution space.

## OBJECTIVES OF METAHEURISTICS

a) To explore the search space efficiently and effectively
b) To exploit the local area in quest of better solution without repeating the sample points.
c) Avoiding redundancy in solution representation in order to reduce the search space size.
d) Using direct or indirect information learnt from points explored so far in order to generate quality new sample point to continue search with. Combining the properties of good solution to generate an even better solution.
e) To escape local optima, i.e., not to get stuck at one point.[14-16]

## GENETIC ALGORITHMS

Genetic algorithm is a nature inspired evolutionary algorithm. It is a population based Meta-heuristics. Genetic algorithm is a global search algorithm. In the initial stage, it takes a big search space and then navigates through it looking for the optimal solution. It does not give a good solution; rather it gives a robust solution which is measured against a fitness function. Since

the solution is measured against fitness function so it avoids local optima and searches for global fitness of the solution. Genetic algorithm is based on the natural selection process. It starts with an initial population which can be generated or gathered using any technique. There are 3 main rules of genetic algorithm-

a) Selection- selection rules select the individuals called parents from the initial population, which contribute towards the future population
b) Crossover- crossover combines the 2 parents to form child for the next generation
c) Mutation- the mutation change applies random changes to individuals (generated from the crossover stage) to form children for the next generation.

Genetic algorithm works on the survival of the fittest. All the individuals compete for the resources. Those individuals that are fitter get the resources and produce children for the next generation. Thus, good genes from parents propagate throughout the population, resulting in children with even better genes. So that all the children of the next stage don't have the same genetic structure and the population does not converge to a local point, mutation is done. Each of these children is called chromosomes and a single variable in these chromosomes is called a gene. Although the selection process is randomized; genetic algorithm is not random, it exploits the genetic information to move the search into a region of better results within the search space. Many factors affect the effectiveness of genetic algorithm, like

a) Population size
b) Selection rule
c) crossover rule
d) mutation rule

The general steps of a genetic algorithm are-

a) Initialize population
b) Assign a fitness value to each individual using the fitness function
c) Individuals with highest fitness value are selected as parents.
d) Children are created from the parents using the selection, crossover and mutation operator.
e) With the help of fitness function, a fitness value is given to each child.
f) Children with the highest fitness value are selected as for new population.
g) Steps c, d, e, f are repeated until a termination condition is met.

## PARTICLE SWARM OPTIMIZATION

Particle swarm optimization algorithm is an algorithm which is based on the concept of a group of particles (population) and is stochastic optimization technique which was derived by the community based behavior of the particles (birds/fishes). The initial population for the problem is a set of random solution from which we search our best solution. Unlike genetic algorithm, this algorithm has no operators. All the potential solutions are called particles. Since there are no parameters to adjust, is algorithm is comparatively easy to implement. Initially, it was used to imitate the patterns of the swarms, later; however, it was found that it can be used as an optimization algorithm. In particle swarm optimization, every single solution is a particle in the solution space. All these particles have their fitness function which is evaluated by the fitness function that we want to optimize. An analogue of directed mutation moves the particle in the space based on the globally best position attained by any particle in the swarm( the global best) or the position secured by the individual particle (the local best). The position is updated by

$x(t+1)= x(t) + v(t+1)$,
where, $x(t+1)$ is the position if the particle at $t+1$ time( the next position),
$x(t)$ is the current position of the particle, and
$v(t+1)$ is the velocity of the particle.

The performance of the algorithm is independent of the size of the swarm. Each particle in the solution space keeps a track of the previous best positions it has achieved till now, which is called p best. One best value that is kept track by the algorithm is the best value that has been secured so far by the particle in its neighborhood, called l best. When a particle takes all the population as its topological neighbor, the best value is the global value called g best. The p best and l best are the local best. This algorithm is easy to apply and is fast and cheap over the other methods. This algorithm never guarantees that a globally optimal solution will be ever found. The advantage of this algorithm is that it doesn't require any assumption about the problem that is to be solved and even explore very large space. The algorithm works in the following way-

a) Initialize a population.
b) Calculate fitness merit for all particles.
c) If the fitness value of the particle is improvised than the previous fitness value of the particle(p best), then the old value is changed by the new fitness value(p best).
d) Choose the particle that has the incomparable fitness value of all of the particles. This is called g best.
e) For each particle, calculate its speed and update its position.
f) Repeat step b, c, d, e till the maximum number of iterations are achieved or the required stop condition is met.

## ANT COLONY OPTIMIZATION

The ant colony optimization algorithm is a well-known probabilistic technique, used for answering the computational problems which can be reduced for finding the best path with the help of the graph. The algorithm searches for a best path in the graph, based on the locomotive behavior of the ants searching the shortest path between the food source and colony. As this algorithm works well with dynamic changing systems, this algorithm can be used for the graphs with changing topology, e.g. -Computer network.
This algorithm is a population based Meta-heuristics. In the natural world, ants move randomly, starting from their colony, in search for food. On finding food, they return to their colony, leaving a chemical signature along the way, also called a pheromone. So if the other ants back in the colony want to go to their food source, they instead of moving randomly follow the chemical signature trail. But after a time gap that chemical signature starts evaporating. As the ants in search of food follow the shortest path, the chemical density on the path increases. The evaporation of the chemical signature trail is beneficial as the ants don't just follow one path (i.e. no local optimality). The evaporation of the chemical signature prevents the solution from premature convergence. In ant colony optimization, ants (particles) search for a good solution to the problem. To apply the algorithm, the given problem is modified to finding the optimal pathon a weighted graph, where the edges are the paths taken are the ants and the vertices are the food sources and/or ant colony. The particles (here ants) incrementally build solution by moving along the edges. In simple terms, ant colony optimization is a stochastic model which follows an iterative procedure for optimization.

## SIMULATED ANNEALING

The simulated annealing algorithm is a probabilistic algorithm technique which

approximates the global optimization of a large search space. Simulated annealing is preferable when finding the precise global optima is less important than finding acceptable local optima in a fixed amount of time. This algorithm employs a random search technique. It interprets slow cooling as a slow decrease in the probability of accepting worse solution because it examines the solution space. Accepting worse solution is a bottom line property of Meta-heuristics because it motivates for more extensive and exhaustive search of the sample space for the optimal solution. This algorithm is based on the annealing process of mechanics- heating a substance and then gradually cooling it to obtain a high strength crystalline structure. In the mechanics, the aim of annealing is to reach the lowest energy state while minimizing the total entropy production. This algorithm is good as it doesn't run into local optima and thus, avoids premature convergence of the solution. It gives a good amount of randomness in the early stages. Simulated annealing is more likely considered as a memory less algorithm that does neighborhood search. It maintains a current assignment of values to variables. At every iteration, it randomly selects a variable. If the assignment of the value to the variable improves the function, then the assignment is accepted and there is a new current assignment. Else, it accepts the assignment with some probability, depending on the value and then computes hoe worse it is to the current assignment. If the change is good, the current assignment is unchanged. The most important parameter of simulate annealing is temperature. It requires an annealing schedule, which specifies how the temperature is reduced. The initial temperature should be high enough as too low initial temperature can get stuck in local optima, but even a very high temperature can cause difficulty in reaching the solution. Iterations at each step should be enough stabilize the system. The annealing schedule used affects the solution quality.

## ARTIFICIAL BEE COLONY

It is a population based optimization algorithm which is based on the clever foraging behavior of the bees. It is a simple flexible and a robust algorithm. It has 3 main parameters against which the working of the algorithm is judged- the size of the population, maximum number of cycles and constraints, all of which are pre-determined by the user. This model has 3 types of bees- the employed bees, the onlooker bees and the scout. At a given time, a single employed bee searches for the food source and each food source has only 1 employed bee associated to it. Thus, the number of employed bees and the number of food sources closer to the hive are equal. The employed bees search for food source, go back to their hive and dance on their food source. The onlooker bees watch the dance of the employed bees and select their food source accordingly. For the food sources that have been abandoned or are now empty, the employed bees of that food source become scout bees and search for new food source.

The food sources represent the possible solution. Initially, very random food source positions are generated. After generation, these food source positions are subjected to repeated cycles by the bees. The employed bees and the onlooker bees share the information about food by dancing on their food source. The onlooker bees evaluate the dance of all of the employed bees and then pick one food source which has the highest nectar. If the next food source has higher nectar concentration as compared to the

present one, it is then only that the employed bees make modification to their food source, by leaving the present food source and moving on to the next food source which has higher food concentration. The deserted food sources are replaced by then new food sources, which are randomly generated by the scouts. This model combines the advantages of the local search method, done by the employed bees and the onlooker bees, with the advantage of the global search method, done by the scouts. The algorithm works as follow-

a) Initialize the population
b) position the employed bees on the food sources
c) position the onlooker bees on the food source depending on the amount of nectar
d) Send the scout bees to search for new food sources
e) Store the location of the new found food source and update them in the memory if they are better than the present food sources
f) Repeat step b, c, d, e until the stop condition is met

## TABU SEARCH

The tabu search is a local search method which searches the entire neighborhood. In tabu search, the information gathered during the iterations is gathered, and stored in memory, which makes the future search process more efficient. The upside of tabu search is that it accepts the non-improvising solution which helps the algorithm to move out of local optima. Since tabu search maintains a memory, it prevents the the revisiting of the previously visited nodes. This accelerates the achievement of the optimal solution. This algorithm starts with a randomly selected solution. If the new solution is of superior quality than the present or the previous solution, in context to the contents of the tabu list, the solution is accepted. If the solutions achieved are not better than the present solution, then the neighboring best solution replaces the current best solution. Tabu search is better than other search methods as it starts with a very simple problem which can further be upgraded to more advanced and complex problems. The tabu search can be used in collaboration with other algorithms so as to prevent them from getting trapped in local optima.

The tabu search maintains a short term memory, which prevents the algorithm from the previous bad solutions. To make the search method more better, medium term and long term memory stores can be used. The medium term memory store the best solution achieved during the search process while the long term memory helps in exploring the unexplored areas of the search space. In tabu search, the main process is to find the local optima and then to continue the search process by allowing non-improvising moves/illegal moves, also called tabu moves, to the best solution in the neighborhood of local optima. Despite all these advantages, the downside of tabu search is that it takes a lot of iteration to reach the optimal solution.

## BIG DATA

Data is any piece of information that can be measured, collected, reported and analyzed. But when this data becomes very large in size and complex in nature that the traditional processing techniques don't work on it, then this data is called big data. Big data includes both the structured and unstructured data. Those entire data sets lie outside the capacity of the traditional software tools to capture, manage and process within a considerable amount are included in big data. This huge quantity of data can be characterized by the following properties-

a) Volume

b) Velocity
c) Value
d) Veracity
e) Variety

Along with the above defined characteristics, big data can also be explained and distinguished on the following parameters-

a) Variability- this refers to inconsistency shown by data, which hinder the process of understanding, handling and managing the data effectively.
b) Complexity- Managing data can be very difficult, especially when huge amount of data comes from multiple sources. Data must be correlated so users can get the information the data is supposed to convey.

Thus, the term big data refers to the capturing, storing, managing and analysis of the very large amount which cannot be handled by the traditional software practices or tools. Types of big data-

a) Structured Data- the data which has a definitive length and format is called structured data. The structured data resides within the fixed fields of a record. Structured data includes all of the data stored in databases and spreadsheets. Since structured data has a format and resides within a database, thus, it is easy to enter, store, query and analyze a structured data. They are managed by SQL. Structured data depend on the data model that has been employed. The sources of structured data can be computer generated and human generated. Structured data can be very powerful and can be utilized for various different purposes. They generally reside within a relational database, due to which they are sometimes called as relational data.
b) Semi-Structured Data- a semi structured data is a data that is neither raw nor typed in a database. It is more-or-less like a structured data which has not been organized into any database. This type of data does have some organizational properties attached to it which makes it easier to analyze the data. They don't have a strict data model, unlike structured data. In semi-structured data arrangement, some types of markers are used which identify certain elements within the data source. Files that are semi-structured may or may not contain relational data, but that data won't be having any organizable structure as some fields may be missing, or they might be arranged in different order. Although semi-structured data have a data model and they cannot be organized into any category, but they do have their meta-data.
c) Unstructured data- those data items that do not have a specified data format are called unstructured data. This type of data is not stored in any database or any other data structure for that matter. The unstructured data can be textual or non-textual. The term big data generally means unstructured data. Big data refers to very large datasets that are difficult to examine with traditional tools and since unstructured data don't have any format attached to them, so no traditional processing techniques can be applied to them. That's why the big data is generally considered unstructured data. Unlike structured and semi-structured data, unstructured data don't have any meta-data tagging.

## BUSINESS INTELLIGENCE

Business intelligence is a set of tools that take raw data and transform it into forms

that business can use for improvisation of the business. This involves data preparation, data analytics and data visualization. It a key factor in decision making in business, which is entirely data driven. Thus, we can say that business intelligence in entirely business dependent. Business intelligence answers the question- what is happening to business. It includes creation, aggregation, examination and perception of data to inform and ease business management. It is not strictly technology dependent. It supports those non-technological tools that involve the processes and the procedures that support data collection, data sharing and data reporting. Business intelligence uses descriptive statistics measure things, detect trends, etc. There are 4 types of big data that really aid business (according to analytics)-

a) Prescriptive analysis- this type of analysis directs the course of actions which should be taken. This is the most valuable kind of analysis and usually gives rules and recommendations for later stages. This type of analysis reflects the next best option.
b) Predictive analysis- it is an analysis of what might happen in the nearby future or even late. The deliverables of the analysis is a predictive forecast. This type of analysis reflects what will happen.
c) Diagnostic analysis- in this analysis, we look in at the previous performance and determine what happened and why. The output is an analytic dashboard. This type of analysis answers the why of the business performance.
d) Descriptive analysis- this type of analysis shoes what is happening based on the incoming data. This type of analysis answers what happened and what is happening, not showing anything about the future of the business performance.

Along with helping the business, big data does have lot of other advantages attached to it. They are-

a) Real time observation and prediction of events that impact business and other endeavors.
b) Ability to search, collect, deduces, shape, analyze, co-join and picture data with the tools of choice.
c) Reducing the risk by optimizing the complex decisions of unplanned events more rapidly
d) Identifying important data which can improvise the decision making process and the quality of the decision
e) The capacity of Hadoop to manage vast amount of data with validation and verification
f) Integration of both structured and unstructured data
g) Reducing cost and time in collection and analysis of data
h) Timely insight to vast amount of data

On the other hand, big data does some challenges to it. They are-

a) Scalability- big data is supposed to have very rapid and elastically growing database to store data
b) Performance- big data moves at a very high speed irrespective of the scale of the project and the workload
c) Quality of data- the data should be accurate and timely, else the process of decision making is jeopardized
d) Understanding of data- it takes a lot of effort to get the data to right shape and have a proper domain knowledge
e) Data security
f) As big data is very huge and complex, it takes lot of time and expensive technology to manage it

## KNOWLEDGE DISCOVERY

Knowledge discovery and data mining (KDD) is an interdisciplinary region that focuses upon the methodologies to derive

useful information from the data. The immense growth of data has lead to the creation of knowledge discovery and data mining methodologies. The need to extract information form data helps in research. Knowledge discovery from data can allow the organization to have a deeper perception and understanding of data. The major threats to knowledge discovery are timeliness and security. The existing principles for deriving new knowledge form the existing data are-

a. Assist different types of analysis methods- different types of users perform different types of operation on big data. Restricting them to asset of tool is not fair as some user are not comfortable or unaware of the tool.
b. One size does not fit all- as the big data being captured is extremely large and is varied, storing all the data in the same relational database is not useful. Moreover, scaling the traditional relational database to accommodate new data is difficult.
c. Making data available- data should easily available. Not only the result of analysis process but also other information like nature of input, assumptions being made, etc. should be available.

But since the rate at which the data is generated is very high, it is necessary that we perform timely analysis of data and ensure data security. The process of knowledge discovery can be summarized the following steps-

a. Define the application domain and the purpose of process
b. Generate a data subset from the big data set for knowledge discovery
c. Remove all the unnecessary details and the missing data fields and finding the time information and known changes
d. Depending on the purpose of doing the task, find the important information
e. Finding the most required and most for data mining method
f. Select the data mining method and algorithm for searching data patterns
g. Revising the selected data patterns
h. Iterating from step 1 to 7
i. Combining the information with another system and reporting

## HADOOP AND MAP REDUCE

Hadoop is a set of tools whose main objective is to support the running of applications on big data. In other words, Hadoop is a framework of tools that help in the smooth running of all the applications that use big data. It is a open source software which is distributed under the Apache license. Big data creates challenges- velocity, volume, variety- that Hadoop handles. Hadoop divides the big data into fragments. There are 2 main components- MapReduce and Hadoop distributed file system. Hadoop works on distributed model and supports the modularization. The advantage of Hadoop is that it is built on keeping hardware failure in mind. It provides fault tolerance. Hadoop is highly scalable and provides enormous processing power. Hadoop is written in Java and it is and since it is written in Java, it is platform independent. Importance of Hadoop-

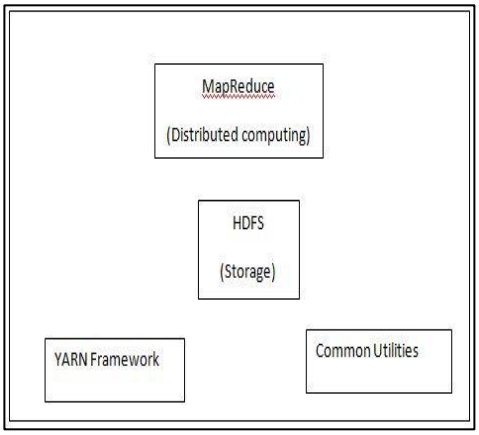
a) It has the ability to store and process large amount of data quickly.
b) It is very high computing power. It works on a distributed computing model which processes data very fast. The more fast computing nodes are used, the more computing power increases
c) Fault tolerance- data and application working on big data need protection against hardware failure. If a computer goes down, the job is automatically

directed to other computer to ensure that the distributed computing does not fail. Multiple copies of data are kept on different computers.

d) Scalability- easily helps to grow the system to handle more quantity of data simply by addition new nodes.
e) Low cost- the open source framework is free of cost.
f) Flexibility- unlike the conventional databases, there is no need to pre-process the data before storing it. We can store as much data as we want and can use it later as per our need.

Challenges of big data- Map-Reduce programming is not a good match for all problems. It is good for easy information requirement and problem which can be divided into independent smaller problems, but it is not efficient for iterative and interactive analysis tasks. Map-Reduce is file intensive. Because the nodes don't communicate through shuffling, iterative algorithm requires multiple map shuffles phases to complete, thus, creating multiple files between Map-Reduce phases which is not good for advanced computing.

**HADOOP ARCHITECTURE**




a) Hadoop common- these are Java classes, libraries and utilities needed by other hadoop modules for their execution. These libraries make available the much needed file system and the operating system abstraction and contain all the necessary Java files and scripts which are needed to start Hadoop
b) Hadoop YRAN- this is a framework is responsible for scheduling the jobs to all the nodes and managing the resources
c) Hadoop distributed file system (HDFS)- it is the in-built distributed file system

**WORKING OF HADOOP**

Step 1- the user submits a job to the Hadoop for required process by mentioning the following terms-

a) Position of the input and output files in the distributed file system.
b) The Java classes and libraries in the form of jar file containing the implementation of map and reduce function.
c) Job configuration by setting different parameters specific to the job.

Step 2- the Hadoop job client then submits the job and configuration of the job to the job tracker which has the duty of circulating the job and configuration to the slave nodes, scheduling the tasks and tracking them, giving status and diagnostics information to the client

Step 3- The task tracker on all the individual nodes executes the task according to the Map-Reduce implementation and the output of the reduce function is stored into the output files on the file system.

of hadoop which provides high throughput connection to application data. It is a scalable, distributed and portable file system, encoded in Java. It uses the TCP/IP socket for communication. All the clients nodes use the remote procedure call (RPC) for communicating with one other. It is highly fault tolerant and holds a large amount of data, which can be streamed to the user applications at very high bandwidth. The distributed file system is based on the Google file system. It uses a master slave architecture, in which the master consists of a single node that controls and manages the file system and one or more slave node that store the data needed for the required and do all the processing.

d) Hadoop MapReduce- this is a YARN based system responsible for the parallel processing of huge data sets.

MapReduce and HDFS are the most important tools of Hadoop.

## MAPREDUCE

The Hadoop Map-Reduce is a software framework used for writing applications which process large amount of data, concurrently. The input and the output are stored on the file system. The framework tales care of the task scheduling, monitoring these tasks and re-executing the failed tasks. The framework consists of a single master job tracker and one slave task tracker per cluster node. The master has the duty of managing the resources (consumption, availability, and tracking) and scheduling the job component to the slaves nodes, monitoring them and re-executing the failed jobs, if any. The slave task tracker executes the tasks are directed by the master and provides status of the task to the master periodically.

The MapReduce algorithm contains 2 important tasks- map and reduces. Map takes a set of data and changes/modifies it into another set of data, where all the individuals are broken into tuples (smaller entities). The reduce task takes the output from the map stages as an input and combines those data tuples into smaller set of tuples, giving the data a more better meaning. The advantage of MapReduce algorithm is that it can very easily scale up to multiple data nodes for processing. In the MapReduce algorithm, the data processing elements are called mappers and reducers. Breaking the data processing elements to mappers and reducers is some time not important. Once we have written an application in the MapReduce form, scaling the application to run over multiple machines in a cluster is just mere change in the configuration. This is why the MapReduce algorithm is used so extensively.

Algorithm for MapReduce-

1. The MapReduce algorithm is based on sending the computer where the data is stored. It is executed in 3 stage- map stage, shuffle stage and reduce stage
   a) Map stage- the map or mapper's job is to alter the input. The inputs are given in the form of a file or directory which is stored in the HDFS. The input file is passed to the mapper function, one line at a time. The mappers process the data and create several smaller chunks of data, by decomposing the data.
   b) Reduce stage- this stage combines the shuffle stage and reduce stage of the MapReduce algorithm. The operates on the output coming from the mapper stage. After processing, it produces output which is further stored in the HDFS.

2. During the MapReduce job, the map and the reduce tasks are send to appropriate nodes by Hadoop.
3. The framework manages and controls all the details of data passing.
4. Majority of the computation tasks takes place on the nodes; all those having their own copy of data needed by task, thus reducing the work load the servers.
5. After completing the given task, the slave nodes collect the data to form the result and send it back to the server.

**TOOLS USED IN HADOOP**

MapReduce and HDFS are the most important tools for Hadoop. But along with MapReduce and HDFS, Hadoop uses a lot of other tools to function correctly. Some of them are as follow-

a) HBase- it is a database management system which is column oriented that runs on the top of HDFS. Although it is a DBMS, but has no support SQL. It is not a relational database system. It is written in Java and does not support the applications written in Auro, REST, and Thrift, which are used for random read and write access.

b) Hive- it is data warehousing application. Unlike HBase, it provides SQL like access and supports relational model. It facilitates querying and managing large datasets that reside in distributed storage. Hive has its own version of SQL language which is called HiveQL. HiveQL still supports the traditional MapReduce algorithm.

c) Sqoop- it is tool used for sending data between relational Hadoop file system and relational database. Sqoop can be used to importing data from relational databases such as MySQL or Oracle into HDFS, transferring the data into Hadoop MapReduce and then sending the data back into a relational database. Sqoop is used for connecting the application to the database server and for controlling parallelism. It imports data to Hive and HBase.

d) Pig- it is a high level data programming language used for the analysis of data for Hadoop computation. Pig is used for evaluating huge data sets which consists of a high level language which is used for data analysis program, combined with infrastructure for the evaluation these programs. The important feature of Pig program is that their structure is accountable to significant parallelization which helps in handling large data sets.

e) Mahout- Apache Mahout is a library of flexible machines learning algorithms which are implemented on top of Apache Hadoop and use the MapReduce prototype. The important features of Mahout are that is supports clustering and classification

f) Oozie- it is workflow management tool used for the Hadoop jobs that are dependent on each other. It is Java web application which is used for scheduling Hadoop jobs. It combines many jobs into 1 logical unit of work. It is mixed in the Hadoop stack and provides support for Hadoop jobs for Apache MapReduce, Pig, Hive and Sqoop. this tool can be used for scheduling. It is a flexible and reliable.

**NEW USES OF HADOOP**

a) Low cost storage and data achieve- the low cost of hardware makes Hadoop useful for storage and combination of data such as social media, sensors, etc. The low storage cost lets you to keep information that is not needed currently but may be needed later

b) Sandbox for discovery and analysis- As Hadoop was designed to deal with large volumes of data of all form; it can run analytic algorithm easily. Big data

analytics on Hadoop helps the organization to operate more dynamically; uncover more opportunities and derive the next level competitive advantage. This approach is best as it gives innovation with minimum investment

c) Data lake- Data Lake supports storing data in its native form. The goal is to give the original view of data to the data scientist and data analyst for understanding and analysis. It helps them to ask new questions without constraints, as the data is available in its original format. Data lakes have not replaced data warehouses; they just compliment the data warehouses.

d) Internet of Things (IoT) and Hadoop- things in the IoT need to know what to communicate, how and when. The bottom-line of IoT is streaming. Large quantity of data and processing capacity also allow you to use Hadoop as a sandbox of exploration and exploitation of patterns to be monitored for prescriptive instructions. It is the easy to improvise these instructions as Hadoop is constantly updated with new data which doesn't match earlier defined patterns.

## METAHEURISTICS IN BIG DATA

The big data is defined as a set of data which are beyond the processing capability of database and computers. 4 main components that are emphasized in the definition of big data are- capturing, storing, managing and analysis. The definition of big data focuses on the quantity of data; however, the complexity and variety are also an important factor that influences big data. Metaheuristics are algorithms that help in organizing the data along some line and give a good solution out of it. The algorithms work on the data sets and give a best solution to the current problem. Metaheuristics is used in big data for the following main purposes-

a) Clustering- it is an unsupervised learning technique that is used to group similar instances on the basis of features. It does not have any kind of training set and does not use any labels of any kind. Clustering is basically a way of grouping the objects together in such a way that objects with similar features come together. It is a common technique used in statistical data analysis. It is not a specific algorithm but a general method to solve the task. The aim of clustering is grouping a set of objects in order to find whether they have any relationship between them.

b) Classification- it is learning technique that is done under supervision and used to assign defined tags to instance on the basis of feature. It is a process of categorization where the objects are recognized, differentiated and understood on the basis of the training set of data. The training set is used for finding similarities. The aim of classification is to find out that the new belongs to which class, from the set of predefined classes.

c) Dimension reduction- This process of reducing the number of variables in the data set. The size of the data affects the performance of the algorithm. Some methods don't work well when the dimension space increases, rather, their performance deteriorates. The dimension reduction process can be divided into 2 processes- feature selection and feature extraction.

Feature Selection- This process chooses an optimal subset of features on the basis of an

objective function. This approach tries to find the subset of the original value.
Feature extraction- This process modifies the data to a smaller dimension from a large dimensional space. The modification can be non-linear or linear.

**Swarm Intelligence Algorithms In Big Data( PSO, ABC, ACO)**

In the swarm intelligence, important information is obtained from the competition and cooperation on the particles. There are 2 types of approaches that can be apply swarm intelligence as data mining techniques. The 1st category consists of techniques where the individual of a group (swarm) move through a solution space to obtain a solution. This kind of approach is called a search approach. The swarm intelligence algorithm are applied to improvise the data mining approach, e.g. - parameter tuning. In the 2nd approach swarms of data instances are placed in a low dimensional space so as to apply suitable clustering or classification to the space. This approach is called the data organizing approach.
Swarm intelligence, especially PSO or ACO is used in data mining to resolve the single objective and/or multiple objective problems. Based on the 2 characteristics of swarm intelligence- self cognitive and social learning- the particle swarms are applied in data clustering technology. In swarm intelligence algorithm, there are multiple answers at the same time. Due to the solutions getting clustered together too fast, there may be pre-mature convergence. But, however, the solution convergence is not always a problem.
The big data analytics is needed to manage large amount of data easily. The increasing dimension of data is also increasing the hardness of the problem.

**REFERENCES**


1. Rodriguez F.J., Martínez C.G., LozanoM. Hybrid metaheuristic based on evolutionary algorithms and simulated annealing: taxonomy, comparison and synergy Test. *IEEE Trans On Evolutionary Computation.* 2012 December; 16(6): 787–800p.
2. Almeida F., Giménez D., López-Espín J.J. *et al.* Parameterized schemes of metaheuristics: basic ideas and applications with genetic algorithms, scatter search and GRASP. IEEE Trans on Systems, Man, And Cybernetics: Systems. May 2013; 43(3): 570–86p.
3. Kumar S., Rao C.S.P. Application of ant colony, genetic algorithm and data mining techniques for scheduling. *Robotics and Computer Aided Manufacturing*. 2009; 901–8p.
4. Plastino A., Fonseca E.R., Fuchshuber R. *et al.* A hybrid data mining for metaheuristics.
5. Padhy N., Mishra P., Panigrahi R. The survey of data mining applications and feature scope. *Int J Comp Sci Engg Information Tech.* June 2012; 2(3): 43–58p.
6. Singh D.K., Swaroop V. Review and analysis of data security in data mining. *Int J Comp Sci Information Tech Security.* 2012 August; 2(4): 831–5p.
7. Vijayarani S., Sakila A. Multimedia mining research – an overview. *International Journal of Computer Graphics & Animation (IJCGA).* 2015 January; 5(1): 69–77p.
8. Jaseena K.U., David J.M. *Issues, Challenges and Solutions: Big Data Mining.* 2014; 131–40p.
9. Khan S., Sharma A., Zamani A.S., *et al.* Data mining for security purposes and its solitude suggestion. *Int. J Scientific & Technology Research*. August 2012; 1(7): 1–4p.
10. Bora S.P. Data mining and ware housing. *IEEE*. 2011; 1–5p.

11. Kesavaraj G., Sukumaran S. A study on classification techniques of data mining. *IEEE*. 2013.
12. Gomaa, Wael H., Aly A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications.* 2013; 68.13x: 13–8p.
13. Belkin N. J., W.B.C. Information filtering and information retrieval: two sides of the same coin? Commun. *ACM*. 1992; 35: 29–38p.
14. Hu M., Wang S., Wang A., Wang Lei. Feature extraction based on the independent component analysis for text classification. In Fuzzy Systems and Knowledge Discovery. *FSKD'08. Fifth International Conference*. 2008; 2: 296–300p.
15. Dasgupta A., P.D., Harb B. *et al.* Feature selection methods for text classification. Proceedings of the 13th ACM SIGKDD International conference on Knowledge discovery and data mining. San Jose, California, USA, 2007**.**
16. Khan A., B.B., Hong Lee L., Khan K. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*. 2010 Feb; 1(1).