INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

METAHEURISTICS: A SOLUTION FROM DATABASE OPTIMIZATION PROBLEMS TO BIG DATA OPTIMIZATION PROBLEMS

AMRINDER KAUR [1], RAKESH KUMAR [2]

1. Assistant Professor, DCSA, M.D. University, Rohtak
2. Professor, DCSA, Kurukshetra University, Kurukshetra

Accepted Date: 16/01/2018; Published Date: 01/02/2018

Abstract: - Databases were conceived in the era in which systems and softwares were deployed in centralized and static environment often in a single datacenter. Today is the data age in which vast amount of data is collected daily from different sources like applications and gadgets, which exists in different fields like business, science, society and almost in every other aspect of daily life. In the computing industry the mega trend in coming decade is Bigdata; it is the driving force which is shaping the industry today. Bigdata helps analyst to unearth unknown patterns for better decisions. The genuine use of parallel and scattered structures is in colossal data examination. Traditional optimization methods are not able to cope up with those problems which scale up in size but metaheuristics procedures proved to be better solution via intensification and diversification of search space. This paper addresses different optimization problems present in database, data mining and big data, which are solved with the help of metaheuristics.

Keywords: Big Data, Database, Data mining, Metaheuristics, Optimization Problem

Corresponding Author: MRS. AMRINDER KAUR

Access Online On:

www.ijpret.com

How to Cite This Article:

Amrinder Kaur, IJPRET, 2018; Volume 6 (6): 100-117

PAPER-QR CODE

I. INTRODUCTION

Database is the foundation of analytics industry. In 1960's data were stored in flat files. Files were the only source which was used to retrieve information, which is concerned with single entity (employee, student etc) **[1].** Application programs that were used for report generation and file processing have intimate knowledge of how data was stored at physical level **[2].** In 1970's, Database management systems were developed. It was defined as relational databases which have carefully designed schemas and query language (SQL) which were used to retrieve information. This database system also follows strict consistency i.e. ACID semantics **[3]**. Database management system consists of user interfaces, query processing and query optimization by which user gained convenient and flexible data access. During 1970's, an efficient method for on-line transaction processing (OLTP) were developed. In OLTP, query is viewed as a read-only transaction and OLTP contributed significantly to the wide acceptance of relational databases. OLTP act as an efficient tool for retrieval of data, storage and management of large amount of data **[4]**.

In mid-1980s, relational database models were developed. RDBMS were conquering DBMS due to its flexibility and ease of use **[2].** In 1980s relational technology is the attention-grabbing technology for research and development activities.

In 1990s, computing resources were getting cheaper and to bridge the performance gap and to maximize the CPU utilization, new techniques like caching, memory management and data movement were developed. At the same era internet were come into existence.**[2]** All these facilities gave birth to advanced data model such as extended-relational, object-oriented and object-relational and application oriented databases which includes spatial, temporal, multimedia, scientific and engineering databases etc.**[4, 2].** In late 1990s, due to amazing progress in computing resources, data collection equipments, internet and storage media, a new kind of data repository emerged, recognized as data warehouse. If data was disparate on multiple sites than it was not possible to analyze the data completely. There was a need to store the data on one site; data warehouse came into existence. **[6]**

Data warehouse is a kind of repository which stores data from heterogeneous sources, organized under a unified schema at a solitary site. Data warehouse facilitate decision making for analysis. Before the storage of data, data should be cleaned and integrated from different sites. After this online analytical processing (OLAP), i.e. an analysis techniques with numeral functionalities like aggregation, summarization, consolidation and to view information from different perspective.

In addition, huge volume of data accumulated, which is beyond database and data warehouse. For example the data driven companies in year 2000 were Google and Amazon **[2]**. Due to this huge volume of data, data repositories become data tombs**.** A tool must be needed which were able to extract valuable information from these huge data repositories in efficient time and cost and the answer to all the problems were data mining. Data mining performs data analysis, uncover hidden patterns, knowledge extraction which contributes to organizations, business, scientific and medical research **[4]**.

Big Data means a data which is too big, too fast and too hard for current tools to process. In technical term big data is defined as three V's- volume, velocity and variety by Gartner. Relational databases do pretty well with respect to data size but for velocity and variety, database don't act upon well. Reason being relational engine doesn't integrate well with streaming technologies and databases don't parallelize effectively to massive qualities of data **[3].** Huge information is not simply information; rather it has turned into a complete subject, which includes different algorithms, tools and systems. Google, Amazon, Microsoft and Yahoo are technology based companies, which have collected and maintained data that is measured in petabytes or even in more proportion. Social media sites like Facebook, Twitter, YouTube and many more have billions of users that continuously generating large a very large scale of data **[5].** At CERN, about one petabytes of data is processed by the data centers, which is equivalent of around 210,000 DVDs. In databases of CERN some 6000 changes are performed every second **[7].** The factual problem associated with huge data are capturing data, storage of data, its curation, searching, sharing, presentation and analysis of data **[8].** The remainder of the paper is structured as follow: Section" Metaheuristics" presents an overview of different types of metaheuristics; Section" Database Optimization Problems" , include different optimization problems solved with the help of metaheuristics; Section" Data Mining Optimization Problems ", discusses optimization problems present in data mining and their solution via metaheuristics; Section" Big Data Optimization Problems" presents optimization problems present in big data and these problems specifically solved with the help of metaheuristics. Section "Conclusion" reiterates the focus of the paper and summarizes the work presented.

II. METAHEURISTICS

In formative years, to solve complex optimization problems specialized heuristics were developed. But picture was changed drastically when general solutions were emerged i.e. Metaheuristics. But the challenge is how to adapt metaheuristics to respective problem or a problem class **[9].** Metaheuristics was coined by Fred Glover in early of 1980s. Metaheuristics is a combination of two words meta- beyond in the sense of high level and heuristic- to search

[10]. It is a class of approximate methods, which are designed to solve complex optimization problems where classical heuristics and other optimization methods failed to be effective and efficient. Metaheuristics formally defined as an iterative generation process which guides the assistant heuristics by different concepts of exploiting and exploring the search space to find reasonable optimal solution **[11]**. In metaheuristics same pattern is repeated during iteration and it never stops until a stopping criterion is achieved which is specified in the beginning**. [12]**

Metaheuristics are used to solve combinatorial problems as well as continuous problems. Combinatorial optimization is defined as finding the optimal arrangement of discrete objects, grouping of discrete objects and their ordering **[11].** It is also defined as efficient allocation of constrained resources to meet up desired objectives. Decision variables can take values from bounded and discrete objects set. In Combinatorial optimization problems, solution space is limited from which best solution will be determined **[13]**. Different combinatorial optimization problems are travelling salesman problem, routing problems in networks, timetable and scheduling problems **[10].** Continuous optimization problems are different from combinatorial optimization problem. It is concerned with optimal setting of parameters. Decision variables can take continuous values. In continuous optimization problems, there is no limit on the solution space but optimal values for continuous variables have to be determined **[13]**.

Different metaheuristics techniques includes ant colony optimization (ACO) **[12] [9]**, particle swarm optimization (PSO) **[14]**, Genetic Algorithm **[12],** tabu search **[9],** Cuckoo search **[17],** Artificial Bee colony **[19].** These algorithms depict the behavior of different species and derive algorithms, which are useful in optimizing complex problems in an effective way.

III. Database Optimization Problems

Database system or Database Management System (DBMS), is the collection of database (set of interrelated tables) and software programs, which are used to access and manage the data. Database could be a bunch of spreadsheets, customer data or anything which can be represented in tabular form. For example, e-commerce website may have tables for customer, orders, items, products or categories. These tables can be used for preparing invoices, orders, customer relations and content management, etc. Some query language is used to access data from databases. But to extract information from the database a massive query (contains several sub queries) will be fired, which will be time consuming if it is not an optimal query. There are several methods which can optimize the speed of database up to a limit; like indexing, cache, by reducing the number of queries by using joins or by using backups and security **[15].** As the number of databases in the queries increased, further many issues will be raised, which can

only be resolved with the help of metaheuristics. Metaheuristics are used to deal with those problems which are large in size. Different optimization issues in the database are discussed below:

A. Query Optimization

Query is request for desired data or information from the database. This requested data may be generated with the help of structured query language. Query optimization is defined as the task of improving the strategy for database query processing. And Query optimization is NP-hard problem When a query is fired it is analyzed for syntax error with the help of parser **[16].** Query graph (standard format) is prepared if no error is found. Query graph is input to optimizer, which selects the best execution plan out of different query execution plans prepared by the optimizer. Query optimization decides the space and time complexity of query.

In **[17**] a variant of cuckoo search metaheuristic i.e. adaptive cuckoo search is used for generating optimal query plan for resource development framework (RDF) for World Wide Web Consortium. For querying RDF, SPARQL language is used.

In **[18]** genetic algorithm is used for query optimization. With the help of genetic algorithm large join queries (LJQ) are optimized. Large join query is defined as that query which includes more than 10 relations. After creating different query execution plans (QEP) join query optimizer determines the lowest query execution plan. Join query optimization is combinatorial problem which is solved in reasonable time with the help of metaheuristics. [18**]** use simulated annealing and iterative improvement strategy which perform well but stuck in local optima. But genetic algorithm outperforms due to its global searching.

B. Multi Join Query Optimization (MJQO)

In relational database management system, the most expensive operation is join operation. Join operator relates two different tables through common attribute. According to join attribute, the matching of all the rows is done in Join operation. Query optimizer can lower the cost of query execution. The optimizer first task is to decide the order of join, which is called multi join query optimization (MJQO). Join processing tree represents the order of execution of join operation. In this tree internal nodes are join operation and leaf nodes are base relations. Different join operation trees are left deep tree, right deep tree and bushy tree **[20]**. Due to extensive database, tables and blocks in query gave birth to multi join query optimization. Multi join query optimization (MJQO) includes different problems like large size of query (in terms of joins), long query execution time and high processing cost. Search engine, data mining, banking

system, decision support system, information retrieval etc. are various applications of multi join query optimization (MJQO) **[16, 19].**

The process of query evaluation is shown in fig 1.

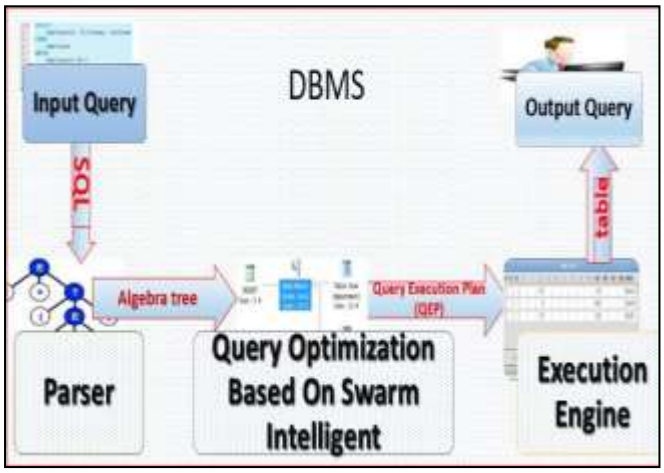

Fig 1: Query Evaluation

In **[16]** swarm intelligence approach i. e. Bee algorithm is used to solve multi join query optimization problem (MJQO). It simulates the behavior of honey bee to compute time. Bee algorithm used for MJQO reduces the time of query processing as compared to particle swarm optimization technique.

In **[19]** a variant of bee algorithm is used, which is known as two phase artificial bee colony algorithm is used to solve multi join query optimization. In first phase, the target is to find interesting plans which includes common sub expressions. These common sub expressions can be reused by different queries. List of all interesting plans are maintained for second phase. In second phase, global optimal plan is obtained from the interesting plans maintained in first phase. This algorithm is very promising in terms of computational time as compared to particle swarm optimization and ant colony optimization.

In **[20]** proposed an algorithm for Multi join query optimization problem which is NP- Complete problem. The proposed algorithm is the combination of two metaheuristics i.e. tabu search and cuckoo search for reducing the response time of query processor.

C. Distributed Database Queries Optimization

Distributed database query optimization is NP-Hard problem. Distributed database (DDB) is defined as collection of data which resides on different database servers having processing capabilities and connected to each other via reliable communication network. For managing distributed database management system (DDBMS) software is used. DDBMS manages the distributed database and facilitate distribution transparent to users. Data fragmentation is supported by DDBMS. Database is divided into fragments (pieces) for physical storage. Fragments can be vertical (subsets of attributes) or horizontal (subsets of tuples). If fragments are placed on different sites then database is called partitioned. The problem of determining optimal join order of database from different sites is known as distributed database queries optimization.

In **[22]** ant colony optimization metaheuristic is combined with dynamic programming for generating good query execution plans quickly and it uses very little memory. This proposed algorithm i.e. DP-ACO is easily adapted to existing query optimizers. Reason being DP-based algorithms is commonly used by query optimizers. Time and space complexity of this algorithm is controlled via SSL (search space limit) parameter. This algorithm outperforms as compared to dynamic programming and genetic algorithm.

In **[21]** combination of particle swarm optimization metaheuristics and evolutionary paradigm i.e. mutation operator are used for distributed database query optimization. Particles represent the solution in the search space and the best global value is used to decide new location of the particle. It treats every particle as an alternative in query plan and from different solution, best solution is selected on the basis of cost (CPU + disk I/O time). When compared with iterative dynamic programming and genetic algorithm, PSO performs well.

III. Data mining Problems resolved with Metaheuristics

Internet and computing resources are ubiquitous in all areas due to enormous growth of storage space and computing power. Due to which every field and sector is generating high volume and multidimensional data.

The digital universe will grow by factor of 300 i.e. from 300 exabytes to 40 trillion gigabytes **[23]**. Over internet, terabytes of new data created everyday in various forms. This enormous data is stored in databases and lately in data warehouses, which contains mountains of raw data. This raw data is useless if it is not transformed into valuable information **[24].**

This need gave birth to data mining; therefore data mining refers to extracting knowledge from large amount of data. It is also termed as knowledge discovery from data i.e. KDD. In Data mining intelligent methods are applied on raw data in order to extract data patterns. Data mining consists of set of modules like characterization, association, correlation analysis, cluster analysis, outlier analysis and evolution analysis **[4].**

With the help of data mining, companies now able to focus on the most useful information from the data warehouses. Data mining is evolving day by day due to evolution of new techniques, which are available for gathering required information every time.**[25]**

Data mining have different problems which are NP-Hard and NP-Complete and these can be solved with the help of metaheuristics. Some of them are discussed below:

A. Cluster Analysis or Clustering

Clustering is defined as a process of dividing similar objects into a number of clusters **[26]**. In the context of machine learning it is classified as unsupervised learning. Objects in clusters are similar to each other in some sense but objects in different clusters are not similar **[27]**. Clusters are used to organize, navigate and summarize the data **[29].** Cluster analysis is used in many applications like data analysis, document clustering, pattern recognition and many more. Clustering algorithms are categorized as partition clustering, hierarchical clustering, density based clustering and graph based clustering **[30].** Many swarm intelligence algorithms have been proposed, which are used for data clustering like genetic algorithms, artificial bee colony algorithm, cuckoo search, particle swarm optimization **[26].**

In **[30]** cuckoo search optimization algorithm is used for overcoming the disadvantage of renowned partioning algorithm i.e. K-means. Problems with k-means algorithm are it may get trapped in local optimal solutions and initially it have to define number of clusters. For forming cluster it use Euclidean distance for calculating the similarity between the objects. Cuckoo search is used to reach the centroids of the clusters and nests are represented by matrix with k rows and m columns. Rows represent centroids and m represents objects. Simulation results show that, the proposed algorithm is useful for data clustering problems.

In **[26]** data clustering is done with improved cuckoo search algorithm. Improvements were done on movement and randomization of cuckoo bird. Improved cuckoo search is proposed, due to convergence and stability is not ideal due to heavy tail property of levy flights. In this improved algorithm, each egg in nest represents a cluster center. This improved cuckoo search

algorithm is compared with CS, PSO, K-means and simulation results show that ICS find optimal clustering effects, best convergence and good stability.

In **[28],** two variant of particle swarm optimization algorithm is used for data clustering. In first variant, simple particle swarm optimization algorithm is used for clustering data into predefined number of clusters. In second variant, K-means algorithm is used to seed the initial swarm and after the termination of k-means algorithm, its output is used as one of the particle in swarm whereas rest of the swarm is randomly initialized. After that particle swarm optimization algorithm is applied for clustering of data. After working on six different datasets and comparison with k-means algorithm, it was prove that particle swarm optimization algorithm and its variant performs well.

In **[27],** genetic algorithm is used for clustering of data by determining the fixed number of centers in cluster. Population size is P and this process is repeated for each P chromosome in population. Initially its fitness value is calculated, proportional selection strategy is opted for selecting the chromosome. With the help of single point crossover, two new children generated and after that mutation is performed. After this genetic algorithm clustering is compared with k-means algorithm and its performance is significantly superior to k-means algorithm.

B. Classification

When categories and collection of data is given then classification is used to find the correct category for the given objects. It is a kind of mapping function which is defined as $f : D \rightarrow C$, where *D,* is defined as object space on the set of classes *C*. **[31]**

In **[35]** a survey is done on particle swarm classification (PSC) which is an outcome of particle swarm optimization algorithm. Literature of PSC is studied in which classification is done by two types of PSO i.e. plain PSO and hybrid PSO. It mentioned regarding two issues i.e. high dimensionality of data and mixed attribute data which encountered in classification. After that PSC is compared with five benchmark datasets (FLUO dataset, LET dataset, OPTIDIGIT dataset, adult dataset and CMC dataset) and it is proved that PSC is efficiently applied to classification problem.

In **[36]** particle swarm optimization is applied to inventory classification problem. In this an algorithm is developed which can be applied to single objective as well as multi objective problems optimized according to different performance measures such as cost, inventory correlation maximization or inventory turnover ratio. The algorithm is able to determine the number of groups for classification and classify the item into groups according to given criteria.

Classification is done on sorted list of items. List of items is sorted according to weight scores and after that cut-off points are determined. Items that lie between cut-off points (that are adjacent to each other) are classified into same group. After comparison of PSO classification algorithm with ABC classification scheme, no grouping, classification by supplier and placing all items in single group; the proposed algorithm outperformed.
In **[37],** a hybrid PSO/ACO algorithm is designed for automatically classify well drilling operations using mud logging. By using pheromone based mechanism of ACO, PSO more effectively cope with categorical attributes. This hybrid algorithm effectively classifies if-then rules using continuous and nominal attributes. The stages of the drilling operation were learnt by the PSO/ACO algorithm based on a classification elaborated by a Petroleum Engineering expert. PSO/ACO is then compared with other classification algorithms like artificial immune system (bio-inspired), decision tree and rule induction algorithm and the proposed algorithm performed well.

C. Outlier Detection

D.M. Hawkins defines the outliers as an observation that deviates too much from rest of the observations **[32].** It is also defined as points in the huge datasets which do not conform to normal behavior. The process of extracting such distinguishable patterns from dataset is termed as outlier detection **[33].** These distinguish data point's reveals useful information in number of based services, climatology and many more **[32].** Outliers can be classified on the basis of data instances i.e. point outliers and collective outliers, vector outliers, graph outliers, trajectory outliers and sequence outliers etc. **[34].** A simple example of outliers in 2d data set is shown in fig 2.

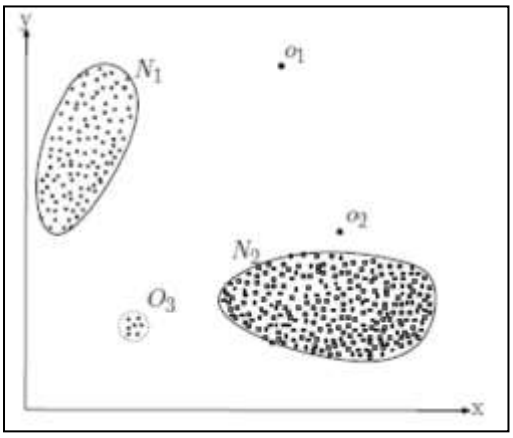

Fig 2: A simple example of outliers in a 2-dimensional data set.

In **[50]** outlier detection problem is considered as optimization problem. Particle swarm optimization algorithm automatically detect outliers in the datasets by evolving the best values for the parameters and then the results are compared with local outlier factor (LOF), which is a distance based outlier detection method.

In **[33],** a hybrid model of grey wolf optimization algorithm and k- nearest neighbor is used. Grey wolf belongs to canidea family and wolfs are considered to be at topmost level in the food chain. They are predators which live in pack of 5-12 on average; it is interesting to study their behavior for survival. GWO passes the location of wolves to K-nearest neighbor algorithm to find optimized results. This approach works well with supervised datasets. This algorithm detects outliers of different datasets in different domains.

In **[34]** two algorithms are proposed for clustering and for outlier detection. In this cuckoo search is combined with k-means algorithm and with support vector data description for clustering and for global outlier detection respectively. The performance of algorithm was measured by introducing significant noise to the original data and the labels of randomly selected data was flipped with opposite labels.

IV.Big Data Optimization Problem

Big data is 3V's i. e. volume, velocity and variety. Data which is too large (volume), too fast (velocity) and having variety of data including structure, semi structured and unstructured data (variety). Some other authors define two other V's i.e. value and veracity where value is defined as worth of the managed information and veracity is accuracy of the information on which decisions are made **[38].**

Big data is different from other data in respect of its characteristics like high dimensions, dynamical data and multi objective problems. The performance of many algorithms decreases as the dimension of data increased exponentially. With the increase in scale of data problems will become more difficult to solve. To handle high dimensional data, it will be transformed into low dimensional space via projection operation. Dynamical data includes videos, audios, real time traffic, and data which rapidly changes over time. Multi objective optimization problem in big data involves two or more objectives to deal with **[14].**

Other challenges with big data include its storage, data capture, searching, analysis, sharing and visualization. In analysis process of big data incompleteness of data, inconsistencies of data, scalability and its security are different challenges **[48].**

HACE theorem describes big data characteristics. These characteristics are very challenging for analysis process or for finding useful knowledge. Big data have heterogeneous and diverse dimensionality. Sources of big data are autonomous which are distributed and having decentralized control **[49]. M**etaheuristics are used to solve different applications in different areas which are described under.

A. **Survey**

In [**39]** optimization is performed on bioinformatics data specifically on protein sequences with swarm algorithm (PSO). Protein sequences are studied because of high volume data created due to genome wide sequencing projects and it is necessary for new drugs, for development of synthetic bio-chemicals and in improvement of plant capability from maceration process (process which is used for extracting plant material). It uses 2-tier hybrid approach to optimize data i.e. filter phase is followed by wrapper phase. Filter phase is used to remove the irrelevant and redundant features from the data due to over fitting issues. After this step data is further refined, to contain the most optimum features in wrapper phase. Wrapper phase employs particle swarm optimization algorithm which is wrapped with support vector machine algorithm. After this the performance of algorithm is measured on the basis of classification accuracy by 10-fold cross validation on the reduced feature subsets and PSO wrapper + filter outperforms as compared to other variants (PSO with all features, PSO with filter).

In **[40]** communication network routing problem is optimized with the help of ant colony optimization metaheuristics. Communication network is classified as packet switched and circuit switched network. In this routing problem of packet switched network is solved with the help of improved ant colony optimization algorithm (IACO). This network is studied because traffic is very high in large networks and nodes get added or removed from the network due to which structure of the network change. In ACO routing problems path is evaluated on the basis of communication delay and number of hops. Therefore path which is shorter gets more pheromone and having higher probability for being selected for transfer of data packets. But problem with simple ACO path with higher pheromone value and shorter delay may become unavailable due to node movements or link disconnections. IACO uses a new pheromone updating rule that integrates local feature, global feature, mutation operator and 2-opt exchange for network routing problem. Source node (an ant) finds route to all the nodes and those nodes which have already been visited and violated the capacity constraints are kept in infeasible node list. Mutation operator randomly mutates the traversed link and creates new solution that is not far from source node. Where as in 2-opt exchange, all possible pairwise exchanges of node locations visited by packets are tested to see if an overall improvement in

the objective function can be achieved. For stimulate the natural evaporation of pheromone initially the pheromone is reduced from all edges so that no path becomes dominant. This improved ant colony optimization is compared with simple ant colony optimization and the results prove that IACO performed better in terms of hops and communication delay.

In **[38]** discusses two metaheuristics i.e. genetic algorithm and particle swarm optimization for optimizing the large data present in railway engineering. It also provides a big data analysis framework, which has various stages like feature extraction, dimension reduction, model training, prediction & confidence estimation and the last stage is evaluation. It also provides a summary of metaheuristics that work in different areas for optimization.

In **[41],** a brief survey of particle swarm optimization algorithm in wireless sensor network is presented. In wireless sensor network many optimization issues exists such as node localization, node deployment, clustering and data aggregation which was solved with the help of PSO. WSN is used in many applications like surveillance, environment monitoring, and disaster management, healthcare and so on. PSO preferred over other alternatives because of its simplicity, high quality solutions, quick convergence and availability for different variants like real, integer and binary domains. For optimal deployment of stationary node it briefly explain PSO- Voronoi and PSO Traffic algorithm. For positioning of mobile nodes (dynamic deployment) and base station it explains particle swarm genetic optimization (PSGO), virtual force co-evolutionary PSO (*VFCPSO*) and PSO Multibase in two-tier WSN respectively. For node localization i.e. determination of location of target nodes, it summarizes three different algorithms i.e. PSO- Iterative, PSO-Beaconless, PSO-4 Beacon. For energy aware clustering in WSN it provides PSO Clustering, PSO-C, MST-PSO algorithms for clustering in wireless sensor networks. Last issue in this paper is data aggregation (a process of combining the data produced from different sources). PSO-Opt Allocation algorithm is discussed whose objective is to reduce the energy expenditure and meanwhile keeping fusion error probability under threshold value. ABC-PSO is used for determine optimal local thresholds and BMPSO is used for optimal sensor configuration.

In **[42],** a hybrid ant colony optimization (ACO) / particle swarm optimization (PSO) algorithm is proposed based on swarming agent. In this paper quality of service (QOS) constraints (for example delay, bandwidth, loss rate, delay jitter) for multicast routing problem is solved using HACOPSO. Multicast is defined as delivery of packets from one source to many destinations. In this algorithm ACO is used to locally guide the agent's movements with the help of pheromone and PSO is used to obtain the global maximum values of attributes with the help of random interaction between agents. The agents are able to sense the pheromone that is close to its

current position while pheromone that is far away is not known to the agent. The proposed algorithm is then compared with PSO-TREE and TGBACA, with the simulation results it was prove that HACOPSO try to minimize tree cost and construct the multicast patterns more sensibly.

In [**43],** a velocity varying PSO feature selection is proposed for dynamic streams (text, images etc.). These data streams (text, images etc.) are high dimensional and having high velocity and for classification of these data streams an algorithm is proposed. The algorithm is wrapper based feature selection model which retains its accuracy by choosing higher fitness values for this a random feature subset. To speed up the searching process multiple agents work in parallel. This new algorithm is implemented in client server system. MongoDB platform is opted for data storage because data streams are unstructured in nature and this platform provides easier and faster integration of data in specific application. In this algorithm SVM based classification in PSO algorithm is used on big data and for graphical user interface HTML, Javascript, JSP is used. Through simulation results it was proved that this proposed algorithm works well for five different datasets from UCI repository.

In **[44],** an incremental genetic algorithm is proposed which builds the rule classification model by evolving fine granular dataset which reduces the learning cost and making it scalable to large datasets. Genetic algorithm is stochastic search method which is widely used for rules classification. For classification problems genetic algorithm calculates the fitness values for all records in the training set. This step contributes to learning cost in every system. This proposed algorithm works in two pass. In first pass it divides the large data into p partition and after that genetic algorithm is applied on each partition for creating rule subset. After this all the rule subsets from all partitions are combined to create a candidate rule set (C). In second pass support and confidence of candidate rules are computed using whole data set and only those rules are selected in final rule set whose support and confidence value is greater than threshold value. Partitioning is done, to reduce learning time because if two partitions are similar only one of them is evaluated. After this, simulation is performed with simple genetic algorithm and incremental genetic algorithm on dataset obtained from UCI machine repository and incremental genetic algorithm outperforms.

In **[45],** genetic algorithm is used for image clustering problems. Image data sets have high volume due to which traditional methods failed to handle. Therefore, genetic algorithm is applied to automatically determine the number of cluster centroids during classification. It uses two component fitness functions; First function, within a cluster measures variance using distance metric and second function measures the similarity of the labeled attributes. It also

pre-processes the data using summarization method for input data. After that a grid constructed and input data is applied to the grid. On the basis of location of representative point the weight is calculated. Simulation is done on two data sets; one is artificial datasets created manually and contains groups of color points and second is real group of image datasets fetched from Google to identify different color space. K-means algorithm produces poor results at every run as compared to genetic algorithm.

In **[46],** an agent based metaheuristic is proposed which integrates evolutionary multi agent system (EMAS) with ant colony optimization. This algorithm is used to solve discrete optimization problem i.e. QAP. This paper also describes ant system and three different methods for pheromone deposition i.e. ant density, ant quantity and ant cycle. Ant cycle creates the best solutions as compared to other two methods. In EMAS the main components are death and reproduction. Action of death is elimination of agent and reproduction means production of agent. The system is implemented on AgE platform developed at AGH University of Science & Technology and two algorithms were tested i.e. original ant system (AS) and min-max ant system (MMAS). MMAS outperforms in solving QAP problem as compared to AS.

In **[47],** ant colony optimization algorithm is used to solve portfolio optimization problem. Portfolio optimization problem is defined as a process of selecting the proportion of different assets to be held in portfolio, in a way to make portfolio better. In this paper, portfolio optimization problem is considered to be multi objective because the one objective is to maximize the portfolio returns and at the same time other objective is to minimize the portfolio risks. This paper also describes the ACO method and its application to portfolio optimization problems. This algorithm searches for the minimum cost which meets the restriction of problems. It selects the weight of each asset in the portfolio to maximize the return and to minimize the risk. The simulation was conducted on MATLAB and ACO method was compared with GA method. ACO was capable to identify maximum returns and minimum risk of portfolio.

V. CONCLUSION

Big data has attracted researchers due to its distinguishable characteristics like volume, velocity and variety. Researchers focus on different optimization issues related to big data like route optimization, storage optimization, location optimization and many more.

This paper briefly describes the problems of database, data mining and big data specifically solved with the help of swarm intelligence metaheuristics and evolutionary metaheuristics like genetic algorithm, particle swarm optimization, ant colony optimization, cuckoo search, artificial bee colony metaheuristics and their variants.

With this survey it is concluded that large scale combinatorial problems and large scale learning problems i.e. knowledge discovery will be solved with swarm intelligence metaheuristics and evolutionary metaheuristics. With the help of this paper researchers will get to know about applications of metaheuristics in different domains.

VI.REFERENCES

1. Dinesh Thakur, *"Traditional fie Processing System:http://ecomputernotes.com/fundamental/what-is-a-database/traditional-file-processing-system"*, accessed on 26 October 2017
2. Maxkanaskar, *"Big Data Technology Series: https://maxkanaskar.wordpress.com/2014/04/06/",* April, 2014, accessed on 26 October 2017.
3. Madden Sam, *"From Databases to Big Data",* IEEE Computer Society, 2012
4. "Jiawei Han et al.," Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, July 2011
5. Najafabadi et al., *"Deep learning applications and challenges in big data analytics"*, Journal of Big Data, 2015
6. Khanolkar Rashmi, *"From Relational Database Management to Big Data: Solutions for Data Migration Testing"*, Cognizant Insights, September 2015
7. *"Computing: http://home.cern/about/computing"*, accessed on 26 October 2017
8. Zhou, Zhi-Hua, et al. "*Big data opportunities and challenges: Discussions from data analytics perspectives* ", Computational Intelligence Magazine, IEEE, 2014
9. Gendreau Michel etal. *, "Metaheuristics in Combinatorial Optimization",* Springer, November, 2005
10. Blum Christian etal. *," Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison",* ACM, September, 2003
11. G. Laporte etal. *,"Metaheuristics: A bibliography"*, Annals of Operations Research, 1996.
12. **Jarraya Bilel etal. ,"***Metaheuristic Optimization Backgrounds: A Literature Review",* International Journal of Contemporary Business Studies,December, 2012
13. F Roythlauf, *" Optimization problems",* Design of Modern Heuristics, Natural Computing Series, Springer, 2011
14. Chang Shi etal. ,"Swarm intelligence in Big Data Analytics", Springer, 2013
15. *"Speeding Up Your Website Database: https://www.smashingmagazine.com/2011/03/speeding-up-your-websites-database/"*, accessed on October 30, 2017
16. Alsaedi A.K.Z. etal. *," An Efficient Multi Join Query Optimization for DBMS Using Swarm Intelligent approach",* IEEE Explore, April 2015

17. Gomathi Ramalingam etal. *," A Novel Adaptive Cuckoo Search for Optimal Query Plan Generation"*, Hindawi Publishing Corporation (The Scientific World Journal), 2014
18. Dong Hongbin etal. *," Genetic Algorithms for Large Join Query Optimization",* ACM, 2007
19. Alsaedi A.K.Z. etal. *,"An Efficient Multi Join Query Optimization for Relational Database Management System Using Two Phase Artificial Bess Colony Algorithm",* Advances in Visual Informatics, Springer, 2015
20. Joshi Mukul etal. *,"Query Optimization: An Intelligent Hybrid Approach using Cuckoo and Tabu Search"*, International Journal of Intelligent Information Technologies, January, 2013
21. Dokeroglu Tansel etal. , *"Particle Swarm Intelligence as a Novel Heuristic for the Optimization of Distributed Database Queries"*, In proceedings of Application of Information and Communication Technologies (AICT), 6th International Conference , IEEE. June, 2013
22. Dokeroglu Tansel etal. , "*Dynamic Programming with Ant Colony Optimization Metaheuristic for Optimization of Distributed Database Queries"*, Springer, 2012
23. J. Gantz and D. Reinsel, "The digital universe in 2020." [Online]. Available: http://www.emc.com/leadership/digitaluniverse/2012iview/executive-summary-a-universe-of.htm.
24. Iglesia Beatriz etal.,"*The Use of Meta-Heuristic Algorithms for Data Mining"*, IEEE, 2005.
25. Verpma Dipti etal.," *Data Mining: Next Generation Challenges & Future Directions*", International Journal of Modeling and Optimization, October 2012
26. Zhao Jie etal.," *Clustering Using Improved Cuckoo Search Algorithm"*, Springer 2014
27. Maulik Ujjwal etal.," *Genetic algorithm-based clustering technique"*, Pattern Recognition Society. Elsevier, 2000
28. Merwe DW van der etal.," *Data Clustering using Particle Swarm Optimization*",
29. P. Manikandan etal.," *Data Clustering using Cuckoo Search Algorithm"*, Springer, 2014
30. Saida Ishak etal.,*"A New Algorithm for Data Clustering Based on CuckooSearch Optimization"*, Springer, 2014
31. Cheng Shi etal. *"Survey on data science with population-based algorithms",* Big Data Analytics, 2016
32. Xi Jingke, *"Outlier Detection Algorithms in Data Mining "*, Second International Symposium on Intelligent Information Technology Application, IEEE, 2008
33. Aswani Reema etal.," *A Novel Approach to Outlier Detection using Modified Grey Wolf Optimization and k-Nearest Neighbors Algorithm",* Indian Journal of Science and Technology, November 2016
34. Jayanthi P et al.," *An Enhanced Cuckoo Search Approach for Outlier Detection with Imperfect Data Labels*", International Journal of Advanced Engineering Technology, June, 2016

35. Nouaouria Nabila etal. , "*Particle Swarm Classification: A survey and positioning*", Elsevier, 2013
36. Tsai C Yang etal. , *"A multiple objective particle swarm optimization approach for inventory classification"*, Elsevier, 2008
37. Nunez Galvan Silvia etal., " *Metaheuristics in Big Data: An Approach to Railway Engineering*", IEEE, 2014.
38. Rahman Shuzlina," *Optimizing Big Data in Bioinformatics with Swarm Algorithms"*, IEEE, 2014.
39. Zaho Dongming etal., " *An Improved Ant Colony Optimization for Communication Network Routing Problem"*, Elsevier, 2010
40. Kulkarni V Raghavendra etal., *"Particle Swarm Opitmization in Wireless Sensor Networks: A Brief Survey*", IEEE Transactions on Systems, Man and Cybernatics, 2011
41. Patel Manoj, "A *Hybrid ACO/PSO based algorithm for QoS Multicast Routing Problem"*, Ain Shams Engineering Journal, 2014
42. Khamitkar Shubham etal.," *Data Stream Mining Big Data using Velocity Varying PSO Feature Selection",* International Journal of Innovative Research in Computer and Communication Engineering, 2016
43. Vivekanandan P etal. *" A Fast Genetic Algorithm for mining classification Rules in Large Datasets*", International Journal on Soft Computing (IJSC), November, 2010
44. Balakrishnan R etal. " An *Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problems*", International Journal of Computer Science Issues (IJCSI), January, 2012
45. Byrski Aleksander etal.," Agent-*Based Meta-Heuristic Approach to Discrete Optimization*", International Conference on Complex, Intelligent, and Software Intensive Systems, 2011
46. Sefiane Silmane etal., "*A Meta-Heuristic Ant Colony Optimization Method for Solving Portfolio Optimization*", International Journal of Financial Management, October, 2013
47. Chen Philip C.L. *"Data-intensive applications, challenges, techniques and technologies: A survey on Big Data"*, Elsevier, 2014
48. Xindong Wu etal.*," Data Mining with Big Data*", IEEE Transaction on Knowledge & Data Engineering, 2013
49. Mohemmed W. Amar," Particle swarm optimization for outlier detection", GECCO '10 Proceedings of the 12th annual conference on Genetic and evolutionary computation, July, 2010.