VOL. 10, NO. 5, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

BIG DATA ANALYSIS BASED ON MATHEMATICAL MODEL: A COMPREHENSIVE SURVEY

Vijaylakshmi S. and Priyadarshini J.
School of Computing Sciences and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, India
E-Mail: vijayalakshhmi.s2014@vit.ac.in

ABSTRACT

Increasing web services day by day and huge volume of data is also increasing exponentially. Processing a large amount of data efficiently can be a substantial problem. Currently, the method for processing a large amount of data comprises adopting parallel computing. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. The challenges comprise analysis, capture, creation, search, sharing, storage, transfer, visualization, and privacy violations. With pervasive sensors continuously collecting and storing enormous amounts of information leads to data flood. Learning from these large volumes of data is expected to bring significant science and engineering advances along with improvements in quality of life. However, with such a big blessing come big challenges. Billions of Internet users and machine-to-machine connections are producing a huge volume of data growth. Utilizing big data requires transforming information infrastructure into a more flexible, distributed, and open environment. In this paper, a survey has been prepared about the techniques available for optimization in big data with the presence of swarm intelligence. Using mathematical model based algorithm for optimization (Swarm Intelligence) in big data will yield better performance while handling of dynamic data in the non-stationary environments and dynamic environments.

Keywords: big data, mathematical model, non-deterministic environment, swarm intelligent, dynamic data.

1. INTRODUCTION

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, it is represented by petabytes and exabytes of data, much of which cannot be integrated easily. Big data deal with large volume of data, the wide variety of types of data and the velocity at which the data must be processed.

Bigdata takes increase in time leads to increase in cost into a traditional relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, data with extended metadata is aggregated in a data lake and Machine Learning and Artificial Intelligence (AI) programs use complex algorithms to look for repeatable patterns.

Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a platform like Hadoop to store large data sets across a distributed cluster and MapReduce to coordinate, combine and process data from multiple sources.

2. BACKGROUND MATERIAL

The Big data can be classified into two categories as shown in Figure-1. First is the stationary environment (Deterministic approach) and second is the non-stationary environment (Non-Deterministic approach). Stationary environments are deterministic system. It fixes the sampling rate for optimization or retrieval. If time gets elapsed the optimization will be low accuracy.

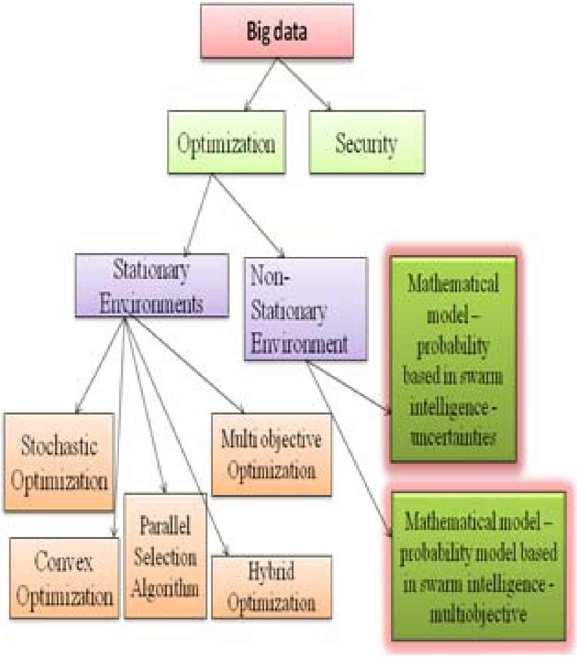
Big data
Optimization
Security
Stationary Environments
Non-Stationary Environment
Mathematical model – probability based in swarm intelligence - uncertainties
Stochastic Optimization
Multi objective Optimization
Parallel Selection Algorithm
Convex Optimization
Hybrid Optimization
Mathematical model – probability model based in swarm intelligence - multiobjective

Figure-1. Flow of optimization in Big Data.

Non-stationary environments are non-deterministic nature. i.e., it has infinite sampling rate for optimization or retrieval of data. So it is necessary to have non-stationary environments for big data analysis. Deterministic environments further divided into five types of optimization algorithms. They are stochastic optimization, convex optimization, Parallel selection algorithm, Hybrid optimization and Multi objective optimization. The problems which have been taken survey are related to non-deterministic environments.

Chia-WeiLee *et al*., (2014) proposed a Dynamic Data Placement Policy (DDP) for map tasks of data locality to allocate data blocks. The Hadoop default data placement strategy is assumed to be applied in a homogeneous environment. In a homogeneous cluster, the Hadoop strategy has been used. The execution time of a job compare with worst-case can make full use of the resources of each node. However, in a heterogeneous environment, produces load imbalance creates the necessity to spend additional overhead. The proposed DDP algorithm is based on the different computing capacities of nodes to allocate data blocks, thereby improving data locality and reducing the additional overhead. They compared the execution time of the DDP with the Hadoop default policy. They used two applications such as WordCount and Grep, regarding WordCount, the DDP execution time is 24.7% and regarding Grep, the execution time of DDP if 32.1% is improved. The author claimed that this approach is applicable for simple applications only.

XiaoyongLi *et al*., (2014) proposed an efficient pruning mechanism for preprocessing with grid summary. Furthermore, many strategies for optimizing the queries based on a feedback mechanism. Extensive experimental results with real data and synthetic data have verified the effectiveness and efficiency of the proposals. Skyline is also referred as multi-objective optimization problem. This paper did not address querying the skylines over complex distributed uncertain data streams.

Zainudin Zukhri *et al*., (2014) proposed, a hybrid optimization algorithm based on GA (Genetic Algorithm) and ACO (Ant Colony Optimization) to solve TSP (Travelling Salesman Problem) and then evaluated for both random data and sample data from the library of TSP. Evolutional process of the GA together with instinct of ant colony in finding the shortest route to seek food are fully combined and formulated as new optimization method called GACO (Genetic Ant Colony Optimization). Experimental studies demonstrate that with small amount of data, it shows insignificancy. But on the big data, it can improve the performance over both GA and ACO. In this case, the solution of the proposed hybridization method has been significantly improved. However, this work only focused on the how to combine GA and ACO procedurally for small amount of data and not for big data.

K. Govinda and S. Maragatham (2013) presented a method for optimizing the storage for big data with the help of open source technology. They used Deflate algorithm to compress the data. Deflate algorithm is a combination of LZ77 algorithm and Huffman coding which is applicable for simple applications. Furthermore, many other methods as well as algorithms should be used to store enormous data.

The SQP (Sequential Quadratic Programming) algorithm proposed by Phillip E *et al.,* (2005), was solving the non-linear programs with large numbers of conditions and variables. Then on linear functions are smooth and first derivatives are available. The algorithm minimizes a sequence of augmented Lagrangian functions, using a QP sub problem at each stage to predict the set of active constraints and to produce a search direction in both the primal and the dual variables. Convergence is assured from arbitrary starting points. They showed that the QP solver for definite QP Hessians and additional techniques are needed to handle even more degrees of freedom for indefinite QP solver.

Julien Mairal (2013) proposed a stochastic majorization-minimization algorithm that gracefully scales to millions of training samples. It has strong theoretical properties and some practical value in the context of machine learning. They derived from their framework several new algorithms, which have shown to match or outperform the state of the art for solving large-scale convex problems, and to open up new possibilities for non-convex ones. They did not address a crucial issue to deal with badly conditioned datasets.

Shi Cheng. *et al*., (2013) analyzed the difficulty of big data analytics problem. Big data analytics are divided into four components: handling large amount of data, handling high dimensional data, handling dynamical data, and multi-objective optimization. Most real world big data problems can be modelled as a large scale, dynamical, and multi-objective problems. They did not survey what have been done in the past, but to suggest the potential of swarm intelligence in big data analytics. Big data involves high dimensional problems and a large amount of data. Swarm intelligence studies the collective behaviours in a group of individuals. Moreover, they showed significant achievements on solving large scale, dynamical, and multi-objective problems. They did not handle the effective methods to solve big data analytics problems using swarm intelligence.

Konstantinos Slavakis *et al*., (2014) presented Principal Component Analysis (PCA), Dictionary Learning (DL), Compressive Sampling (CS), and subspace clustering. It offers scalable architectures and optimization algorithms for decentralized and online learning problems, while revealing fundamental insights into the various analytic and implementation tradeoffs involved. Extensions of the encompassing models to timely data-sketching, tensor- and kernel-based learning tasks are also provided. Finally, the close connections of the presented framework with several big data tasks, such as network visualization, decentralized and dynamic estimation, prediction, and imputation of network link load traffic, as well as imputation in tensor-based medical imaging are highlighted. They showed that the optimization techniques are for stationary environments on big data.

Jin, Y., Sendhoff (2009) proposed an enhanced version of Multiobjective evolutionary algorithms (MOEAs) to evolutionary design optimization characterized by considerations at four levels, namely, the system property level, temporal level, spatial level and process level. They did not address self-organization, self-repair and scalability which plays a central role in optimization.

Rong Hu *et al*., (2014) proposed a clustering-based collaborative filtering approach with agglomerative hierarchical clustering (AHC). The rating similarities between services within the same cluster are computed. The execution time is reduced, they used simple applications.

This survey classifies the literature on optimization in stationary environments with simple applications. There search status on different problem types is reviewed and summarized in Table-1.

2.1 Solution prepared for drawbacks of existing system

A. Handling dynamic data

The big data, such as the web usage data of Internet, real time traffic information, rapidly changes over time. The analytical algorithms needed to process these data quickly. The dynamic problems, sometimes termed as non-stationary environments, or uncertain environments, dynamically change over time. Swarm intelligence has been widely applied to solve stationary and dynamical optimization problems. Swarm intelligence often has to solve optimization problems in the presence of a wide range of uncertainties. Generally, uncertainties in optimized problems can be divided into the following categories.

The fitness function or the processed data is noisy. The design variables and/or the environmental parameters may change after optimization, and the quality of the obtained optimal solution should be robust against environmental changes or deviations from the optimal point. The fitness function is approximated [9], such as surrogate-based fitness evaluations, which means that the fitness function suffers from approximation errors.

The optimum in the problem space may change over time. The algorithm should be able to track the optimum continuously. The target of optimization may change over time. The demand of optimization may adjust to the dynamical environment, for example, there should be a balance between the computing efficiency and the computational cost for different computing loads.

In all these cases, additional measures must be taken so that swarm intelligence algorithms are still able to solve satisfactorily dynamic problems.

B. Multi objective optimization

Different sources of data are integrated in the big data research, and in most of the big data analytics problems, more than one objective need to be satisfied at the same time. According to the number of objectives, optimization problems can be divided as single objective and multiobjective problems. For the multiobjective problems, the traditional mathematical programming techniques have to perform a series of separate runs to satisfy different objectives. Multiobjective Optimization refers to optimization problems that involve two or more objectives, and a set of solutions is sought instead of one. A general multiobjective optimization problem can be described as a vector function **f** that maps a tuples of n parameters (decision variables) to a tuple of k objectives. Unlike the single objective optimization, the multiobjective problems have many or infinite solutions. The optimization goal of an MOP consists of three objectives:

- The distance of the resulting nondominated solutions to the true optimal Pareto front should be minimized
- A good (in most cases uniform) distribution of the obtained solutions is needed
- The spread of the obtained nondominated solutions should be maximized, i.e., for each objective a wide

Range of values should be covered by the non-dominated solutions. In a multiobjective optimization problem, find the set of optimal trade off solutions known as the Pareto optimal set which is defined with respect to the concept of non-dominated points in the objective space.

Swarm intelligence methods can effectively solve the multiobjective problems. Several new techniques are combined in the swarm intelligence techniques to solve multiobjective problems with more than ten objectives, in which almost every solution is Pareto non-dominated in the problems. These techniques include objective decomposition, objective reduction and clustering in the objective space.

3. SUMMARY

There are some optimization algorithms are available in big data which is applicable for deterministic environments and which did not address during uncertain condition using swarm intelligence. The existing optimization algorithms such as stochastic optimization, convex optimization, Parallel selection algorithm, Hybrid optimization and Multi objective optimization are mainly deals with the deterministic environments of dynamic changes. The Table-1.1 shows the summary of existing work.

4. RESEARCH DIRECTIONS

Only with a limited number of optimization is present but this will be apposite for deterministic environment (few parameters are verified).There is no optimization based on automata is available. Handlings of dynamic data in the non-stationary environments and dynamic environments based on the automata model have the possibility for giving the better solution for the above mentioned problems for non- Deterministic environments.

5. CONCLUSIONS

Generally there is an inadequate number of optimization techniques are available for Big Data. It is major research issues, better to develop supplementary number of optimization based on automata model. It helps to optimize the system effectively. This paid special attention to the Deterministic optimization technique. Furthermore, it is better to develop a new optimization for uncertain conditions and dynamic changes over time in big data for non-deterministic environment. So handling of dynamic data in the non-stationary environments and dynamic environments based on the automata model will have the possibility of giving the solution for optimizing the big data in the NDS (Non- Deterministic system) environment.

Table-1. Summary of existing work.

S. No.	Author name	Algorithm proposed in existing system	Advantages of proposed system	Drawbacks of existing system	Solution
1	Chia-WeiLee *et al*.,	Dynamic Data placement policy (DDP) algorithm	Data locality improved and additional overhead reduced. The execution time of DDP 24.7% for Word Count and 23.5% for Grep, improved compared with Hadoop performance.	Only two simple applications are considered.	Supports general application
2	XiaoyongLi *et al*.,	Efficient pruning mechanism	Efficiency and effectiveness have been verified for real and synthetic data.	Did not have Skylines over complex distributed uncertain data streams.	Provides optimization in non-stationary environments of uncertain data.
3	Zainudin Zukhri *et al*.,	Hybrid optimization algorithm based on GA and ACO	Efficiency is increased with small amount of data	Simulation software used for simple applications	Implements for complex applications
4	K. Govinda and S. Maragatham	Deflate algorithm	Applicable for simple applications	implemented with simple applications	Implements for complex applications
5	Phillip. E et. al.,	Sequential quadratic programming algorithm	The algorithm minimizes a sequence of augmented Lagrangian functions, using a QP sub problem at each stage to predict the set of active constraints and to generate a search direction in both the primal and the dual variables	implemented with simple applications	Implements for complex applications
6	Julien Mairal	Stochastic proximal gradient method, stochastic proximal gradient Method	Effectiveness of this algorithm has been verified for solving large-scale structured matrix factorization problems.	This paper did not support for badly conditioned data sets.	Acceptable for all kinds of datasets.
7	Shi Cheng *et al*.,	Survey- handling large amount of data, handling high dimensional data, handling dynamical data, and multi-objective optimization	During certain conditions optimization is supported	During uncertain conditions, optimization are not supported with swarm intelligence	Supports in non-stationary environments using swarm intelligence

REFERENCES

[1] Chia-WeiLee, Kuang-YuHsieh, Sun-YuanHsieh and Hung-ChangHsiao, “A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments”, Journal of Big Data Research, vol. 1, pp. 14-22, 2014.

[2] XiaoyongLi, YijieWang, XiaolingLi, XiaoweiWang, JieYu, “DPS: An Efficient Approach for Skyline Queries over Distributed Uncertain Data”, Journal of Big Data Research, vol.1, pp. 23-36, 2014.

[3] Zainudin Zukhri and Irving Vitra Paputungan,” A hybrid optimization algorithm based on Genetic algorithm and ant colony Optimization”, International Journal of Artificial Intelligence and Applications (IJAIA), vol. 4, no. 5, pp. 63-75, 2013.

[4] K. Govinda and S. Maragatham, “Storage Optimization for Big Data”, International Journal of Applied Engineering Research, vol. 8, no. 19, pp. 2291-2292, 2013.

[5] Phillip E. Gill,Walter Murray,Michael A. Saunders.” SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization”, Journal of SIAM REVIEW of Society for Industrial and Applied Mathematics, vol. 47, no. 1, pp. 99-131, 2005.

[6] Julien Mairal, “Stochastic Majorization-Minimization Algorithms for Large-Scale Optimization”, proceedings of 26th International Conference on Neural Information Processing System, pp. 2283-2291, 2013.

[7] Shi Cheng, Yuhui Shi, Quande Qin and Ruibin Bai, "Swarm Intelligence in Big Data Analytics", Proceedings of 14th International Conference on Intelligent data Engineering and Automated Learning (IDEAL), vol. 8206, springer-verlag, pp. 417-426, 2013.

[8] Konstantinos Slavakis, Georgios B. Giannakis, and Gonzalo Mateos, "Modeling and Optimization for Big Data Analytics", Proceedings of Signal Processing Magazine, IEEE, pp. 18-31, 2014.

[9] Jin, Y., Sendhoff, B,"A systems approach to evolutionary multiobjective structural optimization and beyond", Journal on computational intelligence and magazine, vol. 4, no. 3, pp. 62-76, 2009.

[10] Rong Hu, Wanchun dou and Jianxun liu, "ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application", IEEE Transactions on Emerging Topics in Computing, vo.2, no.3, pp. 302-313, 2014.

[11] Chuanping Hu, Yunhuai Liu and Lan Chen, "Semantic Link Network-Based Model for Organizing Multimedia Big Data", IEEE Transactions on Emerging Topics in Computing, vol.2, no.3, pp. 376-387, 2014.