IJCTA, 9(12), 2016, pp. 5681-5685

Spider Monkey Optimization for Data Manipulation in Big Data

***K. Sriprasadh **S.Mageshkumar ***Dr. S.Sivasubramanian**

Abstract : Big Data is that the largest pool of information, unremarkably these pool of information square measure hold on within the cloud, a cloud act as massive an outsized database wherever a large range of information will be accommodated. The secure manipulating those knowledge is that the biggest task. Security is that the largest concern over the work, there square measure varied algorithms and manipulations square measure provided for the secure manipulation. Here a new approach for knowledge improvement is planned supported characters of spider monkeys. Spider monkeys are referred to as fission- fusion social organization based mostly animals. The animals that follow fission –fusion social systems, they separate themselves from larger to smaller teams supported the inadequacy or accessibility of food. The planned algorithmic program will be thought-about for {the knowledge|the info|the information} segregation and clump of the information within the massive data, consistent with nature of the information.

Keywords : Spider M onkey Optimization, Big data, Data Manipulation.

1. INTRODUCTION

Big data large and fast developing sector for data manipulation big data comprises many techniques for handling data like data mapping , column-oriented databases , hadoop, Hive, Pig which used to handle the data. Even though there are many algorithms and technologies are there big data require a new optimal technique to cluster its own data according to the Volume of data,

2. BIG DAT A AND STORAGE ISSUES

At root, the key requirements of big data storage are that it can handle very large amounts of data and keep scaling to keep up with growth, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools [2]. Data is increasing exponentially, but storage technologies aren't necessarily keeping up. Hard drive capacity isn't keeping pace with the current level data increases. Predictions say that global data will increase by a factor of 50 by 2020, but hard drives are only predicted to grow 15-fold in that same timeframe. Not to mention, hard drive *costs* aren't decreasing quickly enough, either. The cost of digital storage media used to decrease exponentially according to Kryder's Law. The cost of disk space tended to drop about 40 percent each year. But recently storage space costs have not gone down as quickly as in past years. This trend combined with the big data surge means companies and institutions struggle to be able to purchase or lease enough space to keep up with demand for storage space. And right now, there are no good solutions to this problem. There are some stop gap measures available, but each has its own problems.

1. HDD vs SSD solid state drives tend to own a bigger capability than disks ,however SSDs area unit still far pricier than disc drives (HDDs), therefore a complete shift between the 2 storage media is impractical and cost –prohibitive for currently

* Research Scholar Department of Computer Science and Engineering srisaiprasadhhh@gmail.com

** Research Scholar Department of Computer Science and Engineering mageshkumars@yahoo.com

*** Head of the Department Department of Information Technology drsivamdu2011@gmail.com

2. Deletion de duplication is generally erasing pointless information. It may work — however first it will require better calculations for figuring which information is "valuable" and which is most certainly not. In any case, this methodology appears to be encouraging, and individuals are as of now discussing "keen information" rather than enormous information.
3. Cloud storage Each organizations have the alternative source for outsourcing a few or the greater part of their information to the cloud, which could settle the capacity issue for those organizations. Be that as it may, this doesn't address the master plan — the general pattern of exponentialinformation development, and the natural logistical issues. Cloud supplier server farms still need to buy and keep up adequate stockpiling assets to handle the expansion in information figuring needs [2,3].
4. Multiple storage spots Another test with enormous information is that it is regularly put away in various spots. This can make it hard to confirm that all aspects of the information is being gotten to because of a specific inquiry. It is prescribed for utilizing measures to guarantee clients get access of the information they have to do their jobs–and nothing more. In any case, if these principles are set mistakenly or present improper confinements, the examination conveyed will probably be off base. In the event that these worries appear to be sensational, they are no more so than securities required for any delicate information store. The distinction is that so much consideration has been given to the execution and accumulation of huge information that the dangers of access are once in a while disregarded. Utilizing the correct shields can go far toward upgrading both the short-and long haul estimation of even the biggest accumulations of data [3].Solution for these problems data allocation in the proper storage space can be considered as the time being solution. For allocating the data various algorithms are considered, in that Spider monkey optimization algorithm can be used for allocating data in the proper space, according to the data type and data density.

3. SPIDER MONKEY OPTIM IZATION ALGORITHM

Food searching technique of spider monkeys had motivated J.C.Bansal et al a new population based meta-heuristics. This can be applied for data clustering, rendering, segregation and data manipulation the original data is been modified for the data analysis process [5]. This process consists of seven phases

1. Data Population Initialization
2. Base Data Phase (BDP)
3. Global Data Phase (GDP)
4. Base Data Identifying phase
5. Base Data Decision phase
6. Global Data Decision phase

1. **Data Population Initialization :** The data for analysis is taken into consideration at first and weighted then initialized as D. Initial data is denoted by a Dn-dimensional vector SMOn. The derived data and Meta data is numbered as $n = 123\ldots,N$ .Every data is originated from a data which has derived and its own Meta data, In using SMO, the data is represented in the equation as

$$SMO_{ij} = SMO\ min_j + øx\ (SMO_{max\ j} - SMO_{min\ j})\ Ø \in (0, 1) \quad (1)$$

Here $SMO\ min_j$ and $SMO\ max_j$ indicate lower and upper bounds of SMO in jth direction correspondingly.

2. **Base Data Phase :** The subsequent phase is Base data phase, based on the data collection or basic data and meta data. SMO finds the present location of base data and compares fitness of new location and current location based on data weight and applies A^* optimal selection. The i^{th} data position and k^{th}data position is updated using equation

$$SMO_{newij} = SMO_{ij} + rand[0,1]\ x\ (LLkj - SMO_{ij}) + rand\ [-1,1]x(SMO_{rj} - SMO_{ij}) \quad (2)$$

Where SMO_{ij} denote i^{th} SMO in j^{th} dimension, LL correspond to the k^{th} local group data location in j^{th} dimension. SMO_{rj} is the r^{th} SMO which is arbitrarily selected from k^{th} group of data such that $r \neq i$ in j^{th} dimension.

3. **Global Data Phase :** The Global Data Phase (GDP) starts just after finding the BDP based on the data grouping under the nature of the data and the data of the local data altered according to the Eq.(3)

$$SMO_{newij} = SMO_{ij} + rand[0,1]X(GDj - SMO_{ij}) + rand[1\text{-}1]\ X(SMO_{rj} - SMO_{ij}) \quad (3)$$

Where GD_j stands for the global Data's position in j^{th} dimension and $j \in \{1,2,\ldots.D\}$ denotes a randomly selected index. The SMOi updates their locations with the help of probabilities *pi's*. Probability of a particular solution calculated using its fitness. There are number of different methods for computing fitness and probability, here pi computed using Eq.(4)

$$Pi = 0.9\ x\ fitness_i + 0.1\ fitness_{max} \quad \text{Equation (4)}$$

4. **Global Data Identification (GDI) Phase :** Now global data is located in the database with the help of A^* algorithm approach. Highly fitted data is mapped with global data. It also perform a check of mapped data with position of global data, whether it is properly mapped or not and modify global map count accordingly.
5. **Local Data Identification Phase :** Now the derived data is mapped according to its location with help of A* approaches. Highly fitted data is considered as current data within a clustered data as derived data. It also perform a check that the position of derived data is mapped or not and count of relevant data are made.
6. **Local Data Decision (LDD) phase :** In this phase the data taken for the manipulation position is verified for mapping according to the position of the data. In case of no change it randomly initializes positions of LD. Position of LD may be decided with help of Eq. (5)

$$SMO_{newij} = SMO_{ij} + rand\ [0,1]\ X\ (GL_j - SM_{ij}) + rand\ [0,1]X(SM_{ij} - LL_{kj}) \quad (5)$$

7. **Global Data Identification (GDI) Phase :** Local data are grouped under the global data according to weight of threshold and data synthesis, the limit of data is maintained according to the threshold level. Then data are sorted based on data type (based on time and data quality).The sub groups are created based on the data field aspects and named as maximum number of group (MG). Local data are decided for newly created using LDI process.

   The SMO algorithm has four control parameters named base data limit, global data limit, ma number of group and perturbation rate. If the N is meant as group then the maximum of group is then N/10. Base data limit should be in the range of D*N, with the dimension –D; Global data limit should be in range of [N/2,2N] and perturbation rate should be in range [1.0, 09]

   Based on these phases the data are arranged in the big data for the manipulation. Here the data is ordered based on the user preference and availability of the data. Most wanted data are grouped under data population initiation, where the relevant data count is numbered and mapped. The weight of the data also is manipulated in this phase [5],[6]. In base data phase the root of the data is been identified and named as base data, The data relevant to base data meta data is mapped in sequence and in order based on the weight, preference and rank.

   Under global data phase all data are clustered in various group and sub group under a multiple domains based on nature of the data. Now base data is linked with the global data, various data are ranked and ordered based on the usage of the data this ranked data position is changeable according to the user usage of various data in this SMO algorithm mapping is done between each and every data element. This SMO algorithm avoids various storage issues like duplication of data, deletion of data and multiple storage spots.

   Based on these phases the data are arranged in the big data for the manipulation. Here the data is ordered based on the user preference and availability ofhe data. Most wanted data are grouped under data population initiation, where the relevant data count is numbered and mapped. The weight of the data also is manipulated in this phase [5], [6], [7]. In base data phase the root of the data is been identified and named as base data, The data relevant to base data meta data is mapped in sequence and

in order based on the weight, preference and rank. based on nature of the data. Now base data is linked with the global data, various data are ranked and ordered based on the usage of the data this ranked data position is changeable according to the user usage of various data in this SMO algorithm mapping is done between each and every data element. This SMO algorithm avoids various storage issues like duplication of data, deletion of data and multiple storage spots. [6,7]

4. AVOIDING DUPLICATION OF DATA AND SMO ALGORITHM APPROACH

All kinds of data are considered in this approach by that repetition of data is maximum avoided in major databases. As base data is been fixed and related data are followed or threaded from base data, the root of the data can be identified.

In SMO algorithm while processing \base data phase and in global data phase the irrelevant and duplicating of data is sorted out and each data are numbered and ranked. The numbering are made based on appearance of data, rank is made based on user usage of the data.

Numbering of the data is based on the position of the data it is considered if the data is in first position it is denoted as n1, if the data repeated or duplicated it is numbered as n1d similarly all data are numbered as sequenced according to grammar. B y that the data are framed properly.

The duplicated data n1d are identified and analyzed manually and verified to the need of repetition. If it is required the data are left over in the large data base.

5. RANKING OF DATA BASED ON SMO

Normally ranking is applied over data based on the user usage of the data and comments about the data. Relevancy is also checked while ranking is applied. The data are ranked in the position $R_1, R_2, R_3 \ldots R_n$, this ranking is varied according to the user considerations, comments and relevancy.Under global data phase all data are clustered in various group and sub group under a multiple domains.

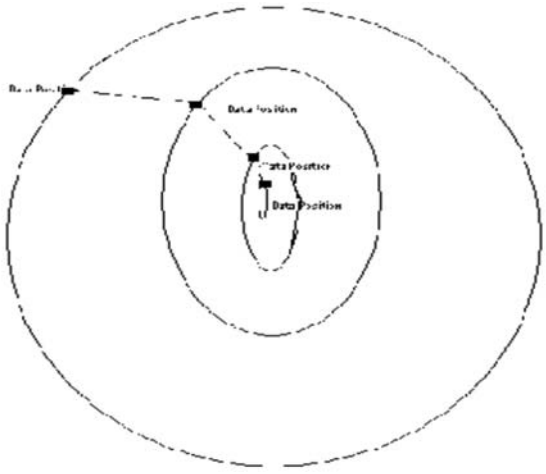

Fig. 1. Ranking of data

Normally updated data will be ranked 1 as it is updated. In SMO data is positioned based on rank ,top ranked data is positioned in the outer circle and related data to the top ranked data is been chained with the data which have been placed in the inner circle till base data.

6. CONCLUSION AND FUTURE ENHANCEM ENT

Using SMO technique in data manipulation in big data simplifies the process based on different phase. Sorting of data using SMO it is a optimized way of approach. In SMO can be applied to some other data oriented manipulation issues in big data.

7. REFERENCE

1. Big Data Now: 2014 Edition Current Perspectiv from O'Reilly Media Publisher: O'ReillyReleased:
2. Big Data Security Challenges and Solutions by Guilermo Lafuente November 10,2014 [3]http://www.contegix.co m/b ig-data-co mes- with-big- problems/#sthash.S7kFIalK.dpuf
4. http://www.datasciencecentral.com
5. Nature –Inspired Metaheuristic Algorithm Xin –She Yang University of Cambridge united Kingdom.
6. Spider Monkey Optimization for numerical optimization, Jagadish Chand Bansal,Harish sharma, Shipmi Singh jadon, Maurice clerc. Memetic Computing March 2014, Volume 6, Issue 1, pp 31-47
7. Fitness Based Position Update in Spider Monkey Optimization algorithm DOI: 10.1016/j.procs.2015.08.504, Conference: 2015 International Conference on Soft Computing and Software Engineering (SCSE'15), At University of California, Berkeley, California, USA, and Volume: 62.