RESEARCH ARTICLE OPEN ACCESS

# ANT-BASED ALGORITHM AND GRID COMPUTING BY BIG DATA CLUSTERING

DERANGULA RADHIKA
Department of CSE, Annamacharya Institute of Technology and Sciences, Rajampet, A.P, India

**Abstract:**

**In present days, the entire business world run towards to data analytics. There consists of huge data called "BIG DATA". Big data has the power to dramatically change the way institutes and organizations use their data. Transforming the massive amounts of data into knowledge will leverage the organizations performance to the maximum. With the spreading prevalence of Big Data, many advances have recently been made in this field. With the spreading prevalence of Big Data, many advances have been recently made in this field.**

***Keywords:*** **Big Data, Grid Computing, Ant-based Algorithm, Clustering.**

## INTRODUCTION:

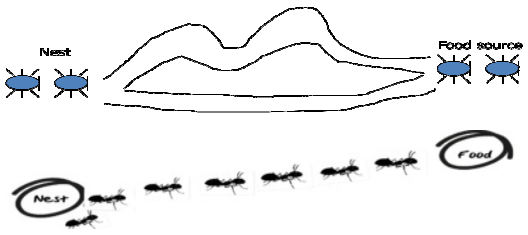
With the spreading prevalence of Big Data, many advances have recently been made in the fields because Data size has increased dramatically with the advent of today's technology in many sectors such as manufacturing, business, science and web application. The term 'BIG DATA' is defined by Garlasu et al. (2013) as data that involve great volume, cannot be structured into regular database tables and are produced with great velocity. Sources of big data can be classified into human-generated data and machine-generated data, through the use of Map/Reduce and Hadoop. The functions are designed to work with a list of inputs. The map function produces an output for each item in the list while the reduce function produces a single output for the entire list. Hadoop is a software library framework for developing highly scalable distributed computing applications. Classification and clustering of big data is important to uncover useful information and knowledge. The algorithms scale poorly in terms of computation time as the size of the data gets large and are impractical without modification when the data exceeds the size of memory. Thus grid technology to overcome the hardware limitation in term of storage space, processing power and memory capacity.

## ANT COLONY OPTIMIZATION:

Ant Colony Optimisation – a way to solve optimisation problems based on the way that ants indirectly communicate directions to each other, have ability to discover the shortest route from the nest to the food source by a chemical substance called pheromones. The ants do not have an advanced vision system but they have the ability to communicate with the environment. In artificial ants, pheromone is deposited along the path. The

evaporation property in artificial ants is a powerful mechanism to update the route information, will most likely select the route with richer pheromones. The mechanism in ANT COLONY OPTIMIZATION is to find the shortest path. The first variant of ACO is an Ant System where the pheromone trail is updated only after all ants have constructed their solutions and the pheromone quantity deposit by each ant is calculated based on the solution quality, the ant system called the Elitist strategy for Ant System (EAS). Rank-Based Ant System (ASrank) is another improvement over ant system In ASrank, each ant deposits an amount of pheromone that decreases with its rank.

**EXAMPLE ON TSP:** The travelling salesman problem (TSP) plays an important role in ant colony optimization because it was the first problem to be attacked by, which the ant colony metaphor is easily adapted, NP-hard problems in compatorial optimization. Stigmergy is indirect communication via interaction with the environment. Individual ants lay pheromone trails while travelling from the nest, to the nest or possibly in both directions. The pheromone trail gradually evaporates over time. But pheromone trail strength accumulate with multiple ants using path.

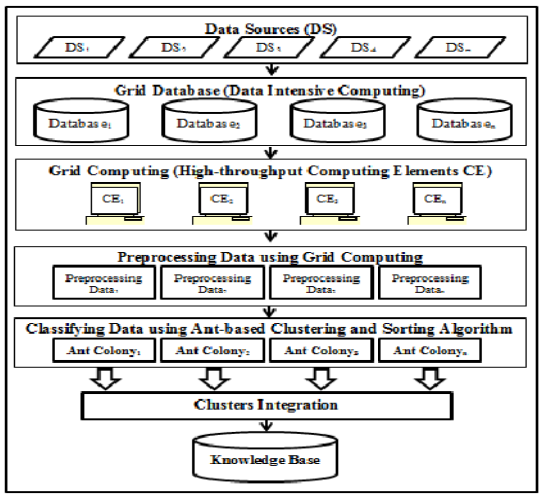



**Ant colony optimization**

Ant colony optimization algorithm has the ability to hybrid with other heuristic & meta heuristic algorithms to solve different types of NP-hard problems. Ant-based clustering algorithm proposed by Deneubourg et al. (1990) has been modifiedby Lumer and Faieta (1994) to be applied in numerical data analysis and data mining.

**FRAMEWORK FOR BIG DATA MANIPULATION:**

Grid technology is used for data storage and data processing while ant-based algorithm is utilized for data clustering in the proposed framework for big data manipulation as depicted in Figure. The flow chart starts with various data sources sending data to the databases. complicated process can be divided into multiple tasks and process them by the computing elements in the grid computing environment. The next layer is high-throughput computing elements, benefit of using unused processor cycles.




**Grid Technology Approach for Big Data**

The preprocessing data stage performs data cleaning, data representation and data scaling to address outliers, missing values, inconsistent values and duplicate data. It provides a powerful nature-inspired heuristics for solving the clustering problems.

The final layer in the framework will integrate the clusters from all computing elements and save them in knowledge base.

**BIG DATA CLUSTERING USING ANT COLONY:**

The problem of finding the right number of clusters is considered as an NP-hard problem.Therefore, meta-heuristic algorithms can be applied as clustering algorithm in solving NP-hard problem.

**Ant-based Clustering Algorithm Pseudo-Code:**

The algorithm's basic principle focuses on agents where the agents represent the ants that randomly move around in their environment which is a squared grid with periodic boundary conditions. While ants wandering around in their environment, they pick up the data item that are either isolated or surrounded by dissimilar ones, will be transported and dropped by ants to form a group with a similar neighborhood items.

```
1: Begin
2: Initialization phase
3: Randomly scatter all data on the grid
4: While (termination condition not met) do
5:     Each ant randomly picks up one data item
6:     Each ant randomly placed on the grid
7:     For each ant (i=1, …., n) do
8:             While (ant[i] carries item)
9:                     ant[i]:= move randomly on the grid
10:                    if (ant[i] decide to drop item) do
11:                            ant[i]:= drop item
12:            End while
13:    End for
14: End while
15: End
```

Base on similarity and density of data items. He probability of picking an element increases with low density and decreases with the similarity of the element. Small clusters of data items grow by attracting ants to deposit more items. The probabilities for any ant to pick and drop an item in improving the quality of the

clustering :

$$P^*_{pick}(i) = \begin{cases} 1.0 & if\ f^*(i) \le 1.0 \\ \frac{1}{f^*(i)^2} & else \end{cases} \quad (1)$$

$$P^*_{drop}(i) = \begin{cases} 1.0 & if\ f^*(i) \ge 1.0 \\ f^*(i)^4 & else, \end{cases} \quad (2)$$

$$f^*(i) = \begin{cases} \frac{1}{\sigma^2}\sum_j\left(1 - \frac{d(i,j)}{\alpha}\right), & if\ \left(f^*(i) > 0 \wedge \forall j\left(1 - \frac{d(i,j)}{\alpha} > 0\right.\right. \\ 0, & otherwise \end{cases} \quad (3)$$

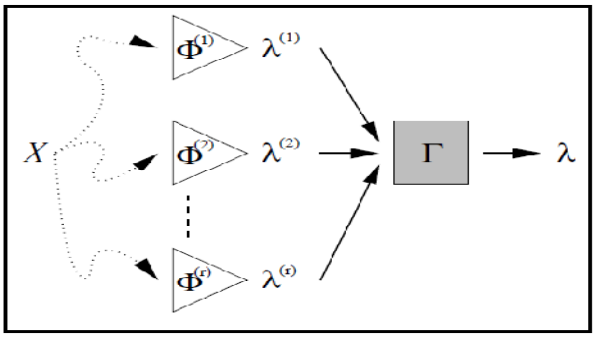
where f*(i), is a modified version of Lumer and Faieta (1994) neighborhood function given by:

The value of α is randomly selected from the interval [0, 1]. The modified function has two important features. First, similar to the original neighborhood function, the division by the neighborhood size penalizes empty grid cells which produce a tight clustering. Secondly, the additional constraint serves the purpose of heavily penalizing

high dissimilarities which significantly improve spatial separation between clusters.

**CLUSTER INTEGRATION:**

Three effective and efficient techniques to obtain high-quality combiners which are known as—consensus functions‖ will be applied. The first combiner focuses on the similarity measurement in the partitions and then re-clusters the objects. The second combiner is based on hyper graph partitioning and the third technique collapses groups of cluster into meta-cluster which then competes for each object to determine the combined clustering. Figure illustrates the cluster ensemble model where $X$ is denoted as a set of features, $\phi$ represents the clustering algorithm, $\lambda$ represents the cluster and $\Gamma$ is the consensus function. The approach consists of two parts.




**Cluster Ensemble Model**

The first part focuses on several independent and heterogeneous ant colonies. In the second part, a queen ant agent aggregates the output clusters from each ant colony using a hyper graph model. The advantage of using queen ant as an agent is that the computing of the new similarity matrix will be done centrally by the queen ant agent rather than letting all the colonies exchange information locally.

**CONCLUSION:**

A framework for the clustering of big data using grid computing and ant colony algorithmhas been proposed. The grid concept is to enable the storage of data in distributed databases across a wide geographical area while ant-based algorithm is for the clustering of big data.Ant-based algorithm has many advantages to be used in big data mining because it has the ability to scale with the size of the data set, prior knowledge. Big data analysis has most important areas is the data security.

**REFERENCES:**

1. Agneeswaran, V. S. (2012). Big-data theoretical, engineering and analytics perspective. In S.
2. Srinivasa & V. Bhatnagar (Eds.), *Big Data Analytics SE* – *2*Berlin, Germany: Springer-Verlag,7678, 8–15.
3. Brzezniak, M., Meyer, N., Flouris, M., Lachaiz, R. & Bilas, A. (2008). Analysis of grid storage element architectures: high-end fiber-channel vs emerging cluster-based networked storage.
4. In M. Brzezniak, N. Meyer, M. Flouris, R. Lachaiz & A. Bilas (Eds.), *Grid middleware and services SE* – 13, US: Springer, 187–201.
5. Bullnheimer, B., Hartl, R. F. & Strauss, C. (1999). A new rank-based version of the ant system: a computational study. *Central European for Operations Research and Economics,* 7(1), 25 – 38.