International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 2, pp 58-63, February-2020

Optimizing RDF Clusters using ACO

[1]Rasha A. Bin-Thalab, [2]Seham A. Bamatraf

[1,2]Assist. Prof., Department of Computer Engineering, College of Engineering & Petroleum, Hadhramout University, Mukalla, Yemen

***Abstract* - Resource Description framework (RDF) in semantics web faces several challenges in terms of rapid increase in its volume and continuous change. This paper presents a new clustering methodology for semantic web data by utilizing ant colony optimization algorithm. The methodology has two pre-processing steps to extract RDF instances and compute a distance matrix between these instances. Next, ACO is implemented to find clusters based on ants discovering the shortest path. The algorithm also uses two objective functions, compactness and separation, to evaluate the discovered clusters. The experiments are conducted on the proposed methodology and showed promised results for clustering quality.**

Keywords: RDF, Semantic web, Clustering, ACO.

I. INTRODUCTION

In the semantic web community[1], the RDF (Resource Description Framework) is one of the basic components for creating semantic data websites. RDF is a type of graph database model which stores semantic facts for publishing data and web exchange. The RDF statements[2]composed of triples which split statements into three components: the subject, predicate, and the object. This makes RDF triples the preferred choice for the management of highly interlinked data. Practically, RDF documents can contain more than one statement. RDF faces several challenges such as data scaling in size[3, 4], variety data structures, and semantic of data[5]. This emerged the problem of how to efficiently browsing RDF data and speed up data access.

In order to query these graphs in such typically large network graph; you have to classify the matches of a particular query in a typically large graph modeled as a target graph. The query diagram can vary considerably from its matches in the target graph in structure and node labels semantic), due to a lack of noise and fixed schema in the target graph which creates challenges with graph querying. However, real RDF graph are complex and ambiguity, with schemes sometimes missing standardized. Graph node labels also hold rich semantics including id, rul[6], personal information. In this case, a match with regard to the label and topological equality may not inevitably be isomorphic to the query graph.

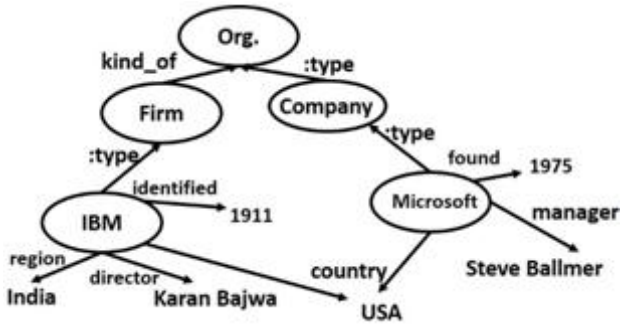

Figure 1: Portion of real-world RDF graph – Dbpedia

Figure 1 shows a real-world graph fragment (portion).Here the structure of RDF has been represented using a directed labelled graph include nodes and edges. Nodes represent the subjects or objects like 'Microsoft' and 'IBM'. Edges, on the other hand, represent the predicates of properties of nodes like as manager and identified. Note that some nodes and labels are similar in meaning like 'Company' and 'Firm', 'manager' and 'director' respectively.

Several methods proposed to enhance retrieving of RDF concentrated on enhanced keyword search algorithms[7]. Other methods handled these large graphs using machine learning techniques such as classification and clustering. Classification used when the data is predefined. However, RDF data are typically not presented with predefined labels when dealing with real applications. Clustering, on the other hand, could be more suitable for categorized RDF data [8-12] since no predefined labels are required. Typically, clustering aims to identify commonality in the features of data which are used for predicting groups of new data. The goal of grouping is to achieve high performance and accuracy in grouping and retrieve similar elements together.

Although, there are different types of clustering including hierarchical, distance, and distribution based, these approaches have limitations due to outliers and predefined number of clusters. Furthermore, clustering can be achieved using Swarm Intelligent (SI) methods[13]. SI is an artificial intelligent method aims to simulate a collaborative behavior of a real, autonomous, and self-organized system. The advantages of swarm intelligence can be described into two points[14]. First, the instability of one or more individuals can't affect the final

ISSN (online): 2581-3048
Volume 4, Issue 2, pp 58-63, February-2020

solution for the entire system. I.e. no central control. (This is useful for outlier problem). Second, individuals communicate in a network environment in a cooperative manner. So, there is no direct communication between individuals, instead, they are interacted implicitly. This is very useful for scalable systems. In this paper, we focused on ACO algorithm which is one of the major paths in SI studies[15]. These algorithms can be very helpful to overcome clustering problems effectively.

To this end, this paper aims to achieve RDF Data clustering based upon ACO approach. Therefore, RDF graph has to be adopted for clustering. We summarized the proposed approach in the following steps:

1. Extracting entities from RDF sub trees with one level.
2. Construct distance matrix using sequence alignments embedding with wordnet tool.
3. Apply ACO to problem to find the shortest path in the distance matrix.
4. Extract clusters depending on ranked distances of the shortest path.

The rest of the paper is organized as follows. Section II presents an overview of related work in literature. Next, section III describes a background of ACO algorithm. Next, section IV introduced the methodology followed in our proposed method to find RDF clusters. Then, section V describes the experiments and results obtained. Section VI concludes and suggests future works.

II. RELATED WORK

Main clustering approaches are categorized as follows: partitioned-based, hierarchical-based, density-based, grid-based and model-based methods. Also, there exists other clustering methods which incorporate the concepts of different clustering approaches[16]. However, existing clustering algorithms are actually subjected to practical constraints including high dimensionality and low interpretability. Each clustering method also has its own drawback. For example, k-means[17] depends on initial instances which may lead to local optimum, and the number of clustering should be defined. Outliers also represent a real challenge for clustering methods.

Work on clustering using ant colony strategies is also represented a good trend for swarm intelligence[18].There are two different methods of clusters in ant colonies: one is called foraging model which is based on the ACO[19], which is inspired by ant colony behavior, and determines the shortest route between their nest and a food-source; the other is called piling model which is based on ant colonial activity as their bodies are clustered and their larvae are sorted[20].In this paper we focused on foraging model.

The foraging model is heavy studied in literature. Tsai with his colleague[21]proposed an effective clustering method for large databases. The paper introduced a virtual ringing model for ants to visit cities decreasingly to get the best local options, then the ant picks a route by using a tournament search technique. Their simulation results showed better performance than fast self-organizing map FSOM combine with K-means and genetic k-means. Next, Runkler[19] utilized ACO for discrete problems. He demonstrated how to optimize objective function for clustering models like hard c-means (HCM) and fuzzy c-means (FCM) with certain modifications.

However, the topic of RDF clustering with swarm intelligence have been discussed earlier by Sebastian's master thesis[13]. The thesis introduced, applied and tested swarm-based techniques in order to build and maintain semantic neighborhoods and retrieve RDF triples efficiently. The technique is based on the LINDA space analysis method. SwarmLinda[22] combines LINDA for swarm methods.

Later, another work[23]presented and evaluated a distributed RDF storage that uses swarm algorithms to cluster similar RDF triples based on a configurable similarity measure.

To our best knowledge, in the RDF cluster strategies, the ant colony algorithm needs to take adequate account and requires further study.

III. BACKGROUND

ACO has been one of the best and most efficient in the history. The ant colony algorithm was proposed by an Italian scientist named Marco Dorigo as a part of his PhD[24]. The first ant inspired algorithm was called the ant system[25]. These days we use improved versions of this algorithm which called ant colony optimization ACO[26].

Normally, ant colonies have the properties of self-adaptation, self-management, robustness, and achieve parallel computation with no prior knowledge[16]. Finding a food is an optimization task where organisms hard to achieve the maximum amount of food source by consuming the minimum amount of energy. In an ant colony, this can be achieved by finding the shortest path from the nest (colony) to any food source. In nature and solve these problems, ants produced chemicals called pheromone. There are many types of pheromones for different purposes in an ant colony of which one of them is used to mark the path towards food sources. Most of the ants are also blind so that’s the only way that they

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 4, Issue 2, pp 58-63, February-2020

can communicate. Ants are more likely to choose a path with higher pheromone (stronger). This means they make decisions based on probabilities. The higher the pheromones level the higher probability of choosing the path. Below we will describe the mathematical model for ACO algorithm.

Algorithm (Pseudo code and formula)

ACO Meta-heuristic

```
Initialize pheromone and parameters
(number of ants, number of iterations,
alpha, beta)
      Loop for each iteration
            Loop for each ant
                  Build solution
                  Daemon action
            Update pheromone
      End Loop
```

a) Build solution

Given ant walks on a graph. The goal of ant is to search for an optimal path in the graph based. Here's the mathematical model used to calculate the probability for all the edges connected to the current node and is a number in the interval of 0 and 1.

$$p^{k}{}_{i,j} = \frac{[\tau_{i,j}{}^{\alpha}]\times[\eta_{i,j}{}^{\beta}]}{\sum[\tau_{i,j}{}^{\alpha}]\times[\eta_{i,j}{}^{\beta}]},$$

Where:

$p^{k}{}_{i,j}$ Represents the probability of ant k to move from node i to node$j$

$\tau_{i,j}$ Represents the amount of pheromone that an ant deposits.

$\eta_{i,j}$ Indicates the quality of edge connected node i to node j (typically *1/dij* where *d* is the cost between nodes)

αParameter constant to regulate the impact of τ

βParameter constant to regulate the impact of η

b) Update pheromone

The quantity of pheromone is updated according to the equation:

$$\tau_{i,j} = (1-\rho)\tau_{i,j} + \sum_{k=1}^{n}\Delta\tau_{i,j},$$

Where:

ρ A constant parameter defines the evaporation rate of pheromone

$\Delta\tau_{i,j}$ The amount of pheromone deposited, typically given by:

$$\Delta\tau_{i,j} = \begin{Bmatrix} \frac{1}{L_k} & if\ ant\ k\ travels\ on\ edge\ i,j \\ 0 & otherwise \end{Bmatrix},$$

Where $L_k$is the length of the path found by ant k. The length is divided by 1 because we search for the shortest path.

IV. METHODOLOGY

For first time readers, ACO-CRDF is a clustering algorithm for RDF data which is used to find similar clusters for a set of n elements based on ACO. Number of ant's 'k' (the number of clusters) is provided as an input from the user.

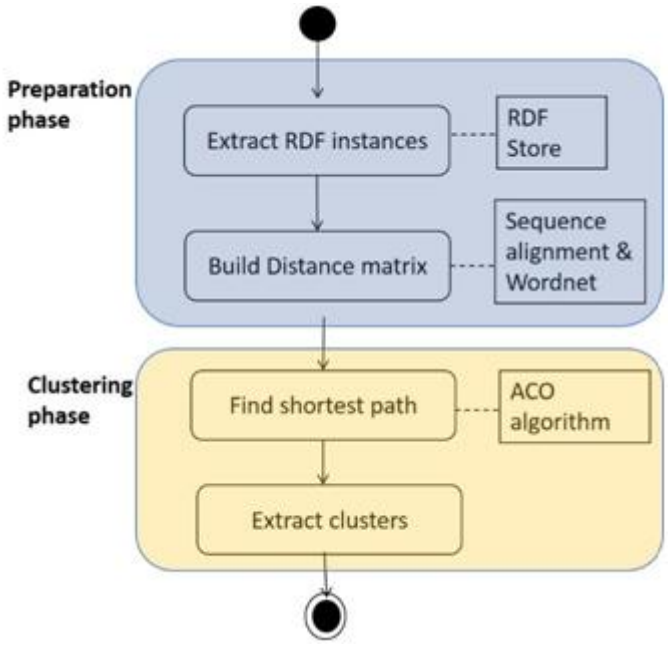

Figure 2: Activity diagram of proposed method

The activity diagram of our proposed method is shown in figure. The methodology (proposed approach) divided into two phases: preparation and clustering. Each phase will be described below.

a) Preparation phase

This phase aims to prepare data to be ready used by ACO algorithm in next phase. So, we need to quantify the instances and distances among them. To this end, we perform two steps as follows.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 4, Issue 2, pp 58-63, February-2020

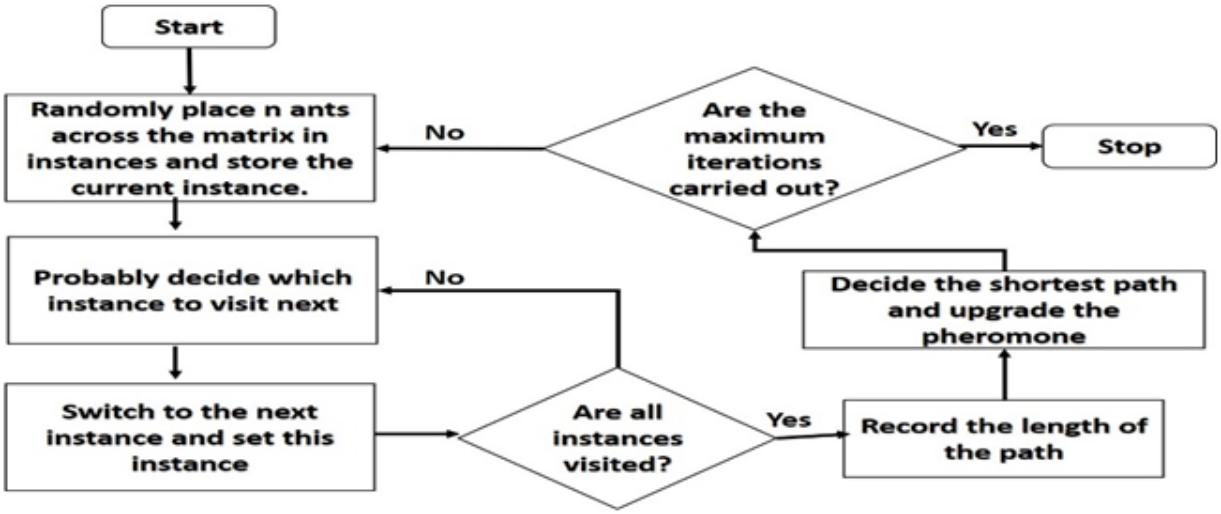

Figure 3: A flowchart to find the shortest path from distance matrix

i) Extract instances

In this step, the RDF triples is processed to extract the instances from RDF graph. Here we chose RDF sub trees which are represented by a parent with its children predicates with one level. For example, from figure 1, we can extract two instances:

Inst1: *Microsoft, type, found, manager, country*

Inst2: *IBM, type, identified, country, director, region*

ii) Construct distance matrix

Clustering of individuals requires conversion into numerical quantities to represent similarity of instances. A number of similarity measures have been discussed to evaluate their suitability in distributed systems. In this paper we used the sequence alignment combined with wordnet as a metric to compute the similarity between RDF instances[9]. This is a good metric to represent the similarities based on both structure and semantic relationships. Needleman-Wunsch [27](NM-W) suggested a solution to compute the similarity score between two sequences. The solution breaks the problem into two parts. First, it generates all possible of the two sequences, each alignment includes a unique combination of match, mismatch, single insertions and deletions. The second step is to use a scoring system to score these trial alignments to find the best one. Every match in a trial alignment is given a score of 1. Every mismatch is given a zero. An individual gap adds a penalty score these numbers are then added across the alignment to obtain a total. The alignment with the highest possible score is defined as the optimal alignment. At the end, the algorithm globally aligns the two sequences end to end. However, in this proposed system, we use wordnet[28] tool to evaluate the similarity between words within the instances sequences. For example, the two instances; Inst1 and Inst2, are arranged to each other to compute their similarity score using NM-W algorithm. The output of this step is a distance matrix for all instances in the RDF document.

b) Clustering phase

This phase includes two sub-stages:

1. Find the shortest path between RDF instances in the distance matrix using ACO strategy. Figure3 shows a flowchart of the ACO algorithm.
2. Partition the RDF instance into clusters according to their distances between them. The distance in the shortest path are arranged in descendent order, where longest distances are taken from one group to another and the shortest within the same group. Figure shows the flowchart for partitioning.

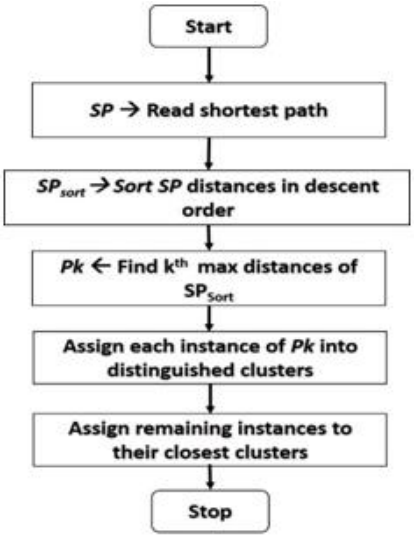

Figure 4: A flowchart to find clusters from the shortest path

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 2, pp 58-63, February-2020

V. EXPERIMENTS

All test data in this paper is downloaded from Drug Bank[29] website. All algorithms are programmed and run on a laptop (CPU: intel core i5-8250U 1.60GHz; memory: 8GB; software: Python). The parameters are listed as below. Initialize pheromone trails $\tau ij(0)$ = 1, number of ants = 10, iteration number 10, decay= 0.95, α = 1, β = 1. Two parameters are measured to evaluate clustering quality. One parameter is the inter-clustering, which is defined as the distance between elements between clusters. The bigger the distance is, the more quality. The second is the intra-clustering to indicate the distance among elements within the same cluster. Here the minimum distance the more quality.

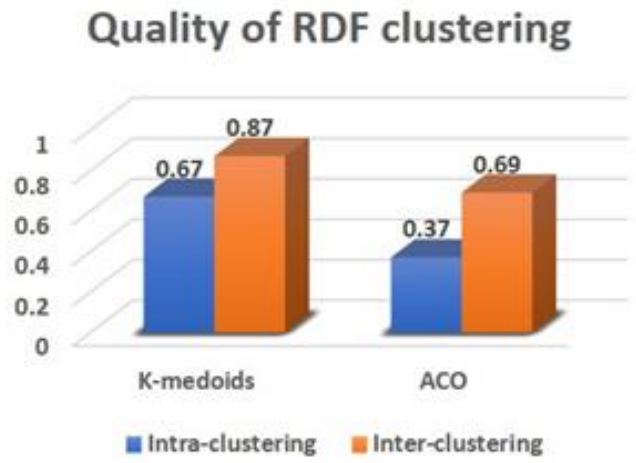

Figure 5: Clustering quality of Drug Bank dataset

As shown in Figure 5, ACO improved the quality of clustering for inter and intra clustering over k-medoids[9]. This is due to the ability of ants to adapting and organizing themselves to get the desired target.

VI. CONCLUSION

This paper addressed the RDF-clustering problem in order to improve data retrieval. The proposed method firstly prepared the data by searching for RDF entity instances and then measured syntactically and semantically the distances between these instances. Next, ACO algorithm was used to determine the shortest path between instances and display the possible cluster hierarchies. Experiments showed good quality of clustering relative to other clustering approaches. Here, we used compactness and separation as an evaluation function for clustering quality. Unfortunately, this single metric evaluation cannot be fitted and robust for all datasets. In future work, we recommend building a framework for multi-objective function to enhance clustering performance.

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, vol. 5, pp. 34-43, 2001.

[2] RDF, Working, and Group, "Resource Description Framework (RDF)." vol. 2019: W3C Semantic web, 2014.

[3] A.Pugliese, O. Udrea, and V. S. Subrahmanian, "Scaling RDF with Time," in Proceedings of the 17th international conference on World Wide Web Beijing, China: ACM, 2008, pp. 605-614.

[4] V. Castellana, J. Weaver, A. Morari, A. Tumeo, D. Haglin, J. Feo, and O. Villa, "Scaling RDF Triple Stores in Size and Performance. Modeling SPARQL Queries as Graph Homomorphism Routines," Handbook of Statistics, pp. 339-362, 2015.

[5] J. Hjelm, Creating the Semantic Web with RDF: Professional Developer's Guide vol. 1: Wiley, 2001.

[6] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," RFC, vol. 3986, pp. 1-61, 2002.

[7] D. Dosso, "Keyword Search on RDF Datasets," in ECIR 2019: Advances in Information Retrieval, Cham, 2019, pp. 332-336.

[8] G. Aluç, M. T. Özsu, and K. Daudjee, "Building self-clustering RDF databases using Tunable-LSH," The VLDB Journal, vol. 28, pp. 173–195, 2019.

[9] S. Bamatraf and R. Bin-Thalab, "Clustering RDF data using K-medoids," in International Conference of Intelligent Computing and Engineering, Mukalla, 2019, p. 8.

[10] S. Eddamiri, E. M. Zemmouri, and A. Benghabrit, "An improved RDF data Clustering Algorithm," in Second International Conference on Intelligent Computing in Data Sciences (ICDS 2018) Fez-Morocco: Elsevier B. V., 2018.

[11] S. Giannini, "RDF Data Clustering," in BIS 2013 Workshop, LNBIP 160, 2013, pp. 220 - 231.

[12] G. A. Grimnes, P. Edwards, and A. Preece, "Instance Based Clustering of Semantic Web Resources," in ESWC 2008: The Semantic Web: Research and Applications Berlin, Heidelberg, 2008, pp. 303-317.

[13] S. Koske, "Swarm Approaches For Semantic Triple Clustering And Retrieval In Distributed RDF Spaces," in FACHBEREICH MATHEMATIK UND INFORMATIK SERIE B • INFORMATIK. vol. M.Sc. Berlin: Freie Universitate Berlin, 2009, p. 146.

[14] J. Yang and J. Yang, "Intelligence Optimization Algorithms: A Survey," Journal of Advancements in Computing Technology, vol. 3, pp. 144-152, 2011.

ISSN (online): 2581-3048
Volume 4, Issue 2, pp 58-63, February-2020

[15] G. Li and K. Xia, "An Improved Data Mining Technique Combined Apriori Algorithm with Ant Colony Algorithm and its Application," International Journal of Digital Content Technology and its Applications, vol. 5, pp. 241-249, 2011.

[16] G. Zhe, L. Dan, A. Baoyu, O. Yangxi, C. Wei, N. Xinxin, and X. Yang, "An Analysis of Ant Colony Clustering Methods: Models, Algorithms and Applications," International Journal of Advancements in Computing Technology (IJACT), vol. 3, 2011.

[17] L. Kaufman and P. J. Rousseeuw, "Clustering by means of Medoids, in Statistical Data Analysis Based on the L1 - Norm and Related Methods," in Y. Dodge North- Holland, 1987, pp. 405–416.

[18] lin, nkaya, K. Sinan, gil, E. Nur, and zdemirel, "Ant Colony Optimization based clustering methodology," Appl. Soft Comput., vol. 28, pp. 301-311, 2015.

[19] T. Runkler, "Ant colony optimization of clustering models," International Journal of Intelligent Systems, vol. 2, pp. 1233-1261, 2005.

[20] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chr, and tien, "The dynamics of collective sorting robot-like ants and ant-like robots," in Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats Paris, France: MIT Press, 1990.

[21] C.-F. Tsai, C.-W. Tsai, and H.-C. Wu, "ACODF: a novel data clustering approach for data mining in large databases," Journal of System Software, vol. 73, pp. 133-145, 2004.

[22] D. Graff, "Implementation and Evaluation of a SwarmLinda System," Masterarbeit. FU Berli, Berlin 2008.

[23] M. Harasic, A. Augustin, P. Obermeier, and R. Tolksdorf, "RDFSwarms: selforganized distributed RDF triple store," in Proceedings of the 2010 ACM Symposium on Applied Computin Sierre, Switzerland: ACM, 2010, pp. 1339-1340.

[24] M. Dorigo, "Optimization, Learning and Natural Algorithms." vol. PhD thesis: Politecnico di Milano, 1992.

[25] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 26, pp. 29-41, 1996.

[26] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem." IEEE Trans. Evolutionary Computation, vol. 1, pp. 53-66, 1997.

[27] S. B.Needleman and C. D.Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology (jmb), vol. 48, pp. 443-453, 1970.

[28] T. Pedersen, S. Patwardhan, and J. A Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts," in Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 38-41.

[29] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Guo, and D. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.." vol. 2019: Nucleic Acids Res, 2011.

Citation of this Article:

Rasha A. Bin-Thalab, Seham A. Bamatraf, “Optimizing RDF Clusters using ACO” Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 4, Issue 2, pp 58-63, February 2020.

****

