Ant Colony Algorithm for Clustering through of Cliques

Julio Cesar Ponce Gallegos[a] , Felipe Padilla Díaz[a], Carlos Alberto Ochoa Ortiz Zezzatti[b], Alejandro Padilla Díaz[a], Eunice Esther Ponce de León [a] y Fatima Sayuri Quezada Aguilera[a]

[a] Departamento de Ciencias de la Computacion, Universidad Autónoma de Aguascalientes, Av. Universidad #940, Aguascalientes, Ags., México
{jcponce, fpadilla, meza, apadilla, eponce, fsquezada}@correo.uaa.mx
[b] Instituto de Ingeniería y Tecnología, Universidad Autónoma de Ciudad Juárez, C.J.,Chihuahua. México
megamax8@hotmail.com

Abstract. This work show an Ant Colony Algorithm (ACO) for Clustering, a technique of group very used in the Data Mining (DM) is the clustering, this algorithm works in the search of maximal cliques which represent groups (clusters). For this was used the algorithm base of Ant Colony for the problem of maximum clique which already was implemented and was made a modification to the algorithm so that worked like clustering algorithm, in the work the algorithm is described and are the experimental results in this first approach.

Keywords: Data Mining, Clustering, Ant Colony Optimization, Maximal Clique.

1 Introduction

The Knowledge Discovery (KD) and Data Mining are powerful tools of data analysis, and it is predicted that they are possible to be turned in the more frequently used analytical tools in the future. The terms “Knowledge Discovery” and “Data Mining” are used to describe the extraction non-trivial or implicit, previously unknown and potentially useful of the data information [10]. The Knowledge Discovery is a concept that describes the process of the search in great volumes of data of patrons who can be considered knowledge on the data. The most well-known branch of the Knowledge Discovery is the Data Mining. The Data Mining, consists in the extraction of the hidden information in great data bases, is a new and potential technology. The Data Mining is a process of Knowledge Discovery in great and complex data sets, refer the extraction process or “miner” of you seed amounts of data [6]. On the other hand, the Data Mining can be used to predict a result for a given organization [5]. The algorithms of clustering in the Data Mining are equivalent to the task of identifying groups of files that are similar between they but different with the rest [9]. The Data Mining is a multidisciplinary field with many techniques. With which it is possible to create a model that describes the data is being used.

The maximum clique problem is a problem of combinatory optimization that is classified within the NP-Hard problems which are difficult to solve. Due to their complexity the exact conventional techniques (exhaustive) take long time to provide a solution, therefore it is necessary to develop heuristic algorithms they solve that it reaching a solution near the optimal in a reasonable time. This problem has real applications eg: Codes Theory, Errors Diagnosis, Computer Vision, Clustering Analysis, Information Retrieval, Learning Automatic, Data Mining, among others. Therefore it is important to use new heuristic and/or meta-heuristics techniques to try to solve this problem, which obtain better results in a polinomial time.
They have been used different heuristic to try to solve this problem, eg: Local Search, Genetic Algorithms, Taboo Search and Ant Colony Optimization Algorithms (ACO) [7]. The Ant Colony Optimization Algorithms are a bio-inspired meta-heuristic based on the behavior of the natural ants, in as they establish the most suitable way between the anthill and a food source [2], these have a great variety of applications between which is the Data Mining.

2 Description of the Clique Problem

Given to a graph nondirected any G= (V, E), in which V= {1.2,…, n} is the set of the vertices of the graph and E is the set of edges. Clique is a set C of vertices where all pair of vertices of C is connected with an edge in G, that is to say C is a complete subgraph. Clique is partial if this form leaves from another clique, of another form this is maximal. The goal of the algorithm is to find all the maximal cliques. An example of clique can be observed graphically in fig 1.

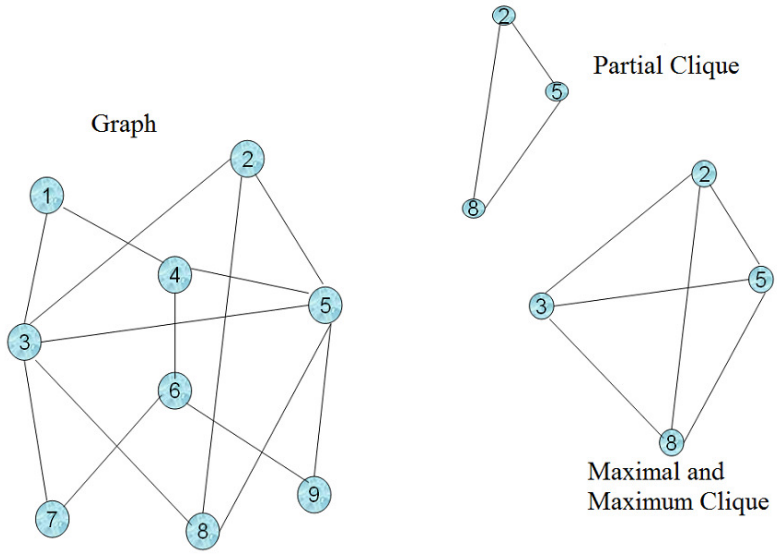

Fig. 1. Example of cliques contained in a graph.

3 **Ant Clique Algorithms (ACA)**

The ACO Algorithms have been used to solve different problems from combinatory optimization [3,8]. The main idea of the ACO is to model the problems to look for the way of minimum cost in a graph. The ants cross the graph in search of good ways (solutions). Each ant is an agent who has a simple behavior so that not always she finds quality ways when this is alone. The ants find better ways as a result of the global cooperation between the colony. This cooperation is realized of an indirect way when depositing the pheromone, a substance that is deposited by an ant in its route.

The general Ant Colony Algorithm for the maximum clique problem proposed by Fenet and Solnon [4] is showed in fig 2.

To initialize the pheromone signs
To place Ants Randomly
Repeat
  For k =1..nb Ants **do**:
    Build the clique (Solution) C_k
  Update the pheromone signs { $C_1, \ldots, C_{nbAnts}$ }
  If is the first iteration to keep the best Solution
  If not to compare if the best solution in the iteration is better than the previous, if it is thus to replace it
Until Reaching the Number of Cycles or Finding the optimum solution

Fig. 2. Pseudo code of Ant Clique Algorithm.

Initialize the pheromone: The ants communicate through the pheromone which is deposited in the edges of the graph. The pheromone concentration in the edge (v_i, v_j) is denoted by $\tau\ (v_i, v_j)$, the initial pheromone sign is denoted by c.

Construction of cliques with the ants: An initial vertex is selected randomly and iteratively it chooses vertices to add to clique. Of set of candidates (all the vertices that are connected with all vertices of the partial clique), to see fig 3.

Choose the first vertex randomly $v_f \in V$
$C \leftarrow \{ v_f \}$
Candidates $\leftarrow \{ v_i / (v_f, v_i) \in E \}$
While Candidate $\neq 0$ **do**
Choose a vertex $v_i \in$ Candidates with a probability p(v_i), see Ec. (2)
$C \leftarrow C \cup \{ v_i \}$
Candidates $\leftarrow$ Candidates $\cap \{ v_j / (v_i, v_j) \in E$
End While
Return C

Fig. 3. Construction of Clique.

The update of the pheromone sign uses the Ec. (1).

$$\tau_{ij}(t+n) = \rho\tau_{ij}(t) + \Delta\tau_{ij} \quad (1)$$

$$p(v_i) = \frac{\left[\tau c(v_i)\right]^{\alpha}}{\sum v\ j\ candidates \left[\tau c(v_j)\right]^{\alpha}} \quad (2)$$

4 Proposed Algorithm

The proposed algorithm is based on the algorithm in [7] the difference is in the part to choose the solution within the clique construction process which is showed in fig 4.

To initialize the pheromone signs
To place Ants Randomly
Repeat
  For k en 1..nb Ants **do**:
    Build the clique (Solution) C_k
  Update the pheromone signs $\{C_1, \ldots, C_{nbAnts}\}$
  If is the first iteration to keep in lists all the solutions without repeating no one
  Else only are added to the list the solutions that not exist in the list
Until Reaching the Number of Cycles or Finding the optimum solution

Fig. 4. Pseudo code of Ant Clustering Algorithm.

In the algorithm all the solutions must be kept since each of these represents maximal clique (cluster).

5 Design of Experiments and Results

The ACO Algorithms depend of the α parameters that is the factor of weight (importance) of the pheromone, and ρ is the percentage of evaporation of the pheromone. If we decrease the value of α, the ants have less sensitivity to the pheromone sign, and if ρ is increased, the evaporation of the pheromone is slower. When the ability of exploration of the ants is increased, these can find better solutions but this implies more time. Taking into account these parameters the algorithm with the following values was executed: Ants number=100, initial concentration of the pheromone c =0.01, importance of the pheromone α =1, factor of evaporation of the pheromone ρ =0.99, maximum concentration that can take the pheromone, number of cycles that executes the Nc=100 algorithm, these values were chosen on the basis of

the results obtained when implementing a first algorithm at the beginning of the 2006 in which the ants within the graph in the vertices with greater degree were placed [7]. In order to carry out the design of experiments we took one from benchmark of the DIMACS [1] that is the used ones at the moment at international level for the problem of maximum clique. It was decided to solve the problem with the execution of software with the parameters before mentioned in the algorithm, in 1 of the 36 benchmarks of the DIMACS that is brock200_2. Which obtain the greater cluster without problem because it is designed to solve the problem of maximum clique, and the found amount of clusters depends on the number of iterations whereupon it is ran.

5 Conclusions and Future Work

In this paper is presented an algorithm based on Ant Colony with a Local Optimizer k-opt, which was used to obtain clusters in a graph taking into account the degree of the vertices from this, which can increase the probability of finding groups (cliques) greater, thus can be seen that it is possible to implement algorithms of Ant Colony to realize clustering in the area of the Data Mining. The proposed algorithm was executed in 1 benchmark of the DIMACS for the problem of maximum clique. This algorithm is a passage in this area since the majority of the proposed algorithms at the moment works clustering under a board of two dimensions, which limits the relations and the size of the applications. Future work: It is important to make a study of the behavior of the parameters and the algorithm, as well as to make a design of ampler experiment to determine which are the best values for the parameters, as well as to already use the algorithm in a real application like the social networks.

Referencias

[1] DIMACS Center for Discrete Mathematics and Theoretical Computer Science fttp://dimacs.rutgers.edu/pub/challenge/graph/benchmarks/

[2] M. Dorigo, V. Maniezzo, and A Colorni (1996) Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transactions on Systems, Man And Cybernetics –Part B: Cybernetics, 26:1, 29-41.

[3] M. Dorigo, G. Di Caro and L.M. Gambardella (1999) Ant algorithms for discrete optimization. Artificial Life, 5(2): 137–172.

[4] S. Fenet and C. Solnon (2003) Searching for Maximum Cliques with Ant Colony Optimization EvoWorkshops 2003, LNCS 2611, 236–245.

[5] J. Hernández, A. Ochoa, J. Muñoz & G. Burlak (2006). Detecting cheats in online student assessments using Data Mining, Proceedings of The 2006 International Conference on Data Mining (DMIN'2006), pp. 204-210, Las Vegas, USA, June 2006, Nevada City

[6] J. Ponce, A. Hernández, A. Ochoa, F. Padilla, A. Padilla, F. Álvarez y E. Ponce de León (2009), Data Mining in Web Application, in Book: Data Mining and Knowledge Discovery in Real Life Applications, Edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, Hard cover, 436 pages, January 2009, Publisher: IN-TECH

[7] J. Ponce, E. Ponce de León , F. Padilla, A. Padilla y A. Ochoa (2006) Algoritmo De Colonia De Hormigas Para El Problema Del Clique Máximo Con Un Optimizador Local K-Opt, Hífen, Uruguaiana, v. 30, n. 58, pag. 191, ISSN 0103-1155, Noviembre, Uruguaiana, Brasil.
[8] Stutzle T. and Hoos H.H. (2000) MAX–MIN Ant System. Journal of Future Generation Computer Systems, 16: 889–914.
[9] Varan, S. (2006). Crime Pattern Detection Using Data Mining, Oracle Corporation
[10] Wahlstrom K., & Roddick J. (2000). On the Impact of Knowledge Discovery and Data Mining, Proceedings of Australian Institute of Computer Ethics Conference (AiCE2000), Canberra, Australia, April 2000, Sydney City.