Hierarchical Approach in Clustering to Euclidean Traveling Salesman Problem

Abdulah Fajar[1,2], Nanna Suryana Herman[2], Nur Azman Abu[2], and Sahrin Shahib[2]

[1] Faculty of Engineering, Universitas Widyatama, Jl Cikutra 204A Bandung 40125 Indonesia
abd.fajar@gmail.com

[2] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal Melaka Malaysia
{nura,nsuryana,shahrinsahib}@utem.edu.my

Abstract. There has been growing interest in studying combinatorial optimization problems by clustering strategy, with a special emphasis on the traveling salesman problem (TSP). TSP naturally arises as a sub problem in much transportation, manufacturing and logistics application, this problem has caught much attention of mathematicians and computer scientists. A clustering approach will decompose TSP into sub graph and form cluster, so it may reduce problem size into smaller problem. Impact of hierarchical approach will be investigated to produce a better clustering strategy that fit into Euclidean TSP. Clustering strategy to Euclidean TSP consist of two main step, there are; clustering and tour construction. The significant of this research is clustering approach solution result has error less than 10% compare to best known solution (TSPLIB) and there is improvement to a hierarchical clustering algorithm in order to fit in such Euclidean TSP solution method.

Keywords: Hierarchical Approach, Clustering, Dendogram, Tour Construction Euclidean TSP.

1 Introduction

Nature inspired problem solving is become popular approach in last decade. This work inspired by human problem solving and particle movement such as Brownian Motions. Human beings possess the natural ability of clustering objects. This is a capability to organize the data which is received everyday into cluster, so that they may draw important conclusions. However making computer to solve clustering problem is quite difficult and demanding the attentions of computer scientist and engineers all over the world till now.

The task of computerized data clustering problem has been approached from diverse domains of knowledge like graph theory, statistics, artificial neural network and so on. The most popular approach in this direction has been the formulations of clustering as an optimization problem. There has been growing interest in studying combinatorial optimization problems by clustering strategy, with a special emphasis on the traveling salesman problem (TSP) [2][3][4]. TSP naturally arises as a sub problem

G. Shen and X. Huang (Eds.): ECWAC 2011, Part I, CCIS 143, pp. 192–198, 2011.

in much transportation, manufacturing and logistics application, this problem has caught much attention of mathematicians and computer scientists.

Traveling salesman problem (TSP) is a combinatorial optimization task of finding the shortest tour of n cities given the intercity costs. In a more formal way the goal is to find the least weight Hamiltonian cycle in a complete graph G. The TSP problem is a '*NP-hard*' problem [5]. An *NP-hard* problem is extremely unlikely to have a polynomial algorithm to solve it optimally although it is not proved that the worst case exponential solution running times are unavoidable. Good approximation algorithms can produce solutions that are only a few percent longer than an optimal solution and the time of solving the problem is a low-order polynomial function of the number of cities. Some approximating algorithms produce tours whose lengths are close to that of the shortest tour, but the time complexity is substantially higher than linear.

The primary objective of this research is to investigate impact of hierarchical approach to produce a better hierarchical clustering strategy that fit into such Euclidean TSP solution method. The current research in hierarchical clustering to solve Euclidean TSP mostly producing their best algorithm with the running time near $O(n^2)$ while this approach is expected to be at most near $O(n^3)$ and average final TSP solution is more than 10% compare to the best known solution.

The rest of this paper will describe about motivation of this study regarding Euclidean TSP in section 2, objective this study will be described in section 3. Literature background will be described in section 4 and section 5 describes methodology. Preliminary result is presented in section 6 also final section discussion and conclusion written in section 7 and 8.

2 Literature Background

A clustering algorithm is expected to discover the natural grouping that exists in a set of pattern. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis.

Among the most popular hierarchical clustering algorithms, BIRCH [6] can typically find a good clustering in a single scan of the data and improve the quality further in a few additional scans. It is also the first clustering algorithm that handles noise effectively. CURE [7] represents each cluster by a certain number of points selected from a well-scattered sample and then shrinking them toward the cluster centroid by a specified fraction. It uses a combination of random sampling and partition clustering to handle large databases. ROCK [8] is a robust clustering algorithm for Boolean and categorical data. It introduces two new concepts: a point's neighbors and links, and to measure the similarity/proximity between a pair of data points.

Haxhimusa *et. al* formulated a pyramid algorithm[1] for E-TSP motivated by the failure to identify an existing algorithm that could provide a good fit to the subjects data. More recently, hierarchical (pyramid) algorithms have been used to model mental mechanisms involved in other types of visual problems. The main aspects of the models are (multi resolution) pyramid architecture and a coarse to fine process of successive tour approximations. This algorithm was motivated by the properties of the human visual system has main idea to use agglomerative (bottom up) processes to reduce the size of the input by clustering the nodes into small group and top-down refinement to find an approximate solution, as same as Graham do in [3]. The size of the input (number of vertices in the graph) is reduced so that an optimal solution can be found by the combinatorial search. A pyramid is used to reduce the size of the input in the bottom-up processes.

In bottom-up clustering, cities are close neighbors are put into the same cluster using greedy approach. These clustered cities are considered as a single city at the reduced resolution. Graph pyramid strategy use MST using Bouruvka algorithm. MST is used as natural lower bound for TSP solution. In the case TSP with triangle inequality which in the case for the Euclidean TSP, MST can be used as to prove the upper bound.

Graph pyramid solution give good approximation while is compared to TSPLIB. Solution error for this solution tends to linear around 12%. Time performance is quadratic for this solution especially while number of cities more than 600.

3 Hierarchical Approach in Clustering

A clustering algorithm is expected to discover the natural grouping that exists in a set of pattern. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram, as shown in Figure 1. Figure 1 shows dendogram which illustrates the fusions or divisions made at each successive stage of analysis.

Among the most popular hierarchical clustering algorithms, BIRCH [6] can typically find a good clustering in a single scan of the data and improve the quality further in a few additional scans. It is also the first clustering algorithm that handles noise effectively. CURE [7] represents each cluster by a certain number of points selected from a well-scattered sample and then shrinking them toward the cluster centroid by a specified fraction. It uses a combination of random sampling and partition clustering to handle large databases. ROCK [8] is a robust clustering algorithm for Boolean and categorical data. It introduces two new concepts: a point's neighbors and links, and to measure the similarity/proximity between a pair of data points.

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, C_n, C_{n-1}, ..., C_1. The first C_n consists of n single object 'clusters', the last C_1, consists of single group containing all n cases. At each particular stage the method joins together the two clusters which are closest together (most similar). Differences between methods arise because of the different ways of defining distance (or similarity) between clusters.

Number of clusters that is generated in hierarchical clustering is automatically. In this work, it is defined a threshold R [2], where

$$R = \sqrt{\frac{\text{Area of Cluster } A}{\text{The number of Vertices } n}} \tag{1}$$

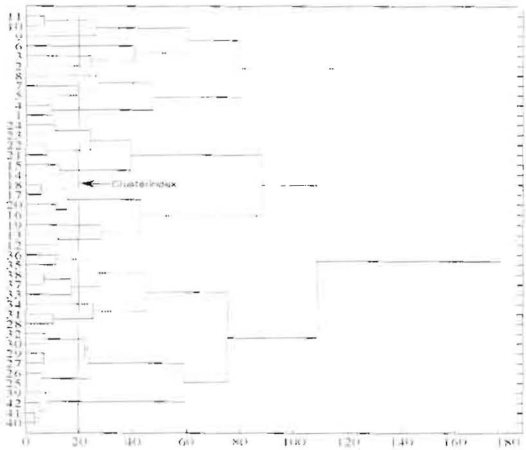

Fig. 1. Dendgram of Complete Linkage Method for Dantzig42.tsp

This threshold will act as guidance to split one cluster to several and result basic clusters. Basic cluster have small size number of nodes. This threshold also prevent clusters overlapped one to others.

Clustering algorithm should contains function to compute linkage distance for a set of clusters C and the linkage distance function will vary depend on linkage method. Here is several technique of linkage between the clusters used in this study:

1. **Nearest linkage:** The distance between two clusters is the minimum distance

$$d(G_r, G_S) = \min\{W(V_r, V_S)\} \tag{2}$$

2. **Complete linkage:** The opposite of nearest linkage, maximum distance

$$d(G_r, G_S) = \max\{W(V_r, V_S)\} \tag{3}$$

3. **Average linkage:** The distance between two clusters is mean distance between all possible pairs of nodes in two clusters.

$$d(G_r, G_S) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(V_{r_i}, V_{s_j}) \tag{4}$$

4. **Median Linkage:** The distance between two clusters is their weighted centroids

$$d(G_r, G_S) = \left\| \tilde{V}_r - \tilde{V}_s \right\|_2 \tag{5}$$

Where $\tilde{V}_r$ is the weighted centroid of r. If r was created from clusters p and q, then $\tilde{V}_r$ is defined recursively as:

$$\tilde{V} = \frac{1}{2(\tilde{V}_p + \tilde{V}_q)} \tag{6}$$

5. **Centroid linkage:** The distance between two clusters is their centroid as calculated by arithmetic mean

$$d(G_r, G_s) = \left\| \tilde{V}_r - \tilde{V}_s \right\|_2 \tag{7}$$

Where $\tilde{x}$ is the centroid of r, by arithmetic mean:

$$\tilde{V}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} V_{r_i} \tag{8}$$

6. **Ward's linkage (Sum Square Method):** The distance between two clusters, is how much the sum squares will increase when merge them

$$d(G_r, G_s) = \frac{n_r + n_s}{n_r . n_s} \left\| \tilde{V}_r - \tilde{V}_s \right\|^2 \tag{9}$$

Where $\tilde{V}_r$ center of cluster r and n_r is is number of points in cluster r.

4 Clustering Result and Impact on Euclidean TSP

This research use TSP Lib [9] (70 sample problem with node size between 29 and 3038) and randomized sample data to test the algorithm. All linkage distance function show uniform result except nearest linkage. Nearest linkage result a chained effect clusters and this effect tend to worst result on Euclidean TSP, while other methods generate compact size clusters. Number of nodes per clusters between 2 – 4 nodes in these linkage distance function, see figure 2 below.

Ward's and completes linkage distance function are promising method when compared to TSP Lib best known result. The result on average is less than 7%. Tour construction that is used in this research is cheapest insertion method and result refined by 2-opt exchange move. There are 75% samples which has consumed up to 50% concorde running time from TSP Lib using 2.6 GHz Processor. Longer running time is consumed by TSP Lib sample problem about 14%. Figure 3 show distribution of running time consumption level.

The research has adopt and develop several algorithm and each algorithm has complexity started from $O(n^2)$ and the worst case is $O(n^3)$. Clustering process is divided into two main algorithms, linkage and clusters generation. Complexity of clustering process at the worst case is $O(n^3)$. Tour construction and improvement has $O(n^2 \log n)$ time complexity. The overall of the algorithm complexity for this work has $O(n^3)$.

The clustering process has identified the best method to cluster the Euclidean TSP into smaller with compact size clusters. Complete and Ward's Linkage has shown as the best method of all linkage method. Both linkage methods have good result in running time and percentage of gap compared to TSP Lib. Some method has shown potential result and the drawback lay on running time in tour construction. Several method result overlap clusters and big size cluster such as average and centroid linkage method. Overlap and big size cluster have the impact to tour construction and running time. Bigger cluster size result longer running time and quality of tour construction, as well as overlap clusters.

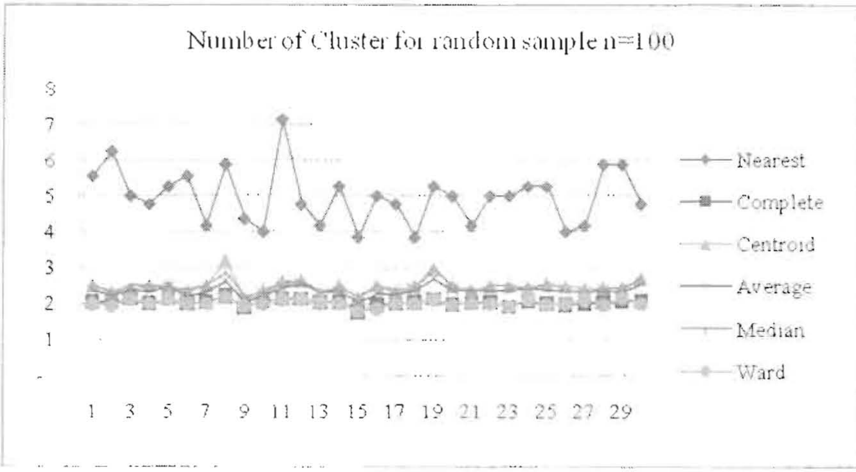

Fig. 2. Number of nodes per clusters fror random sample with n=10

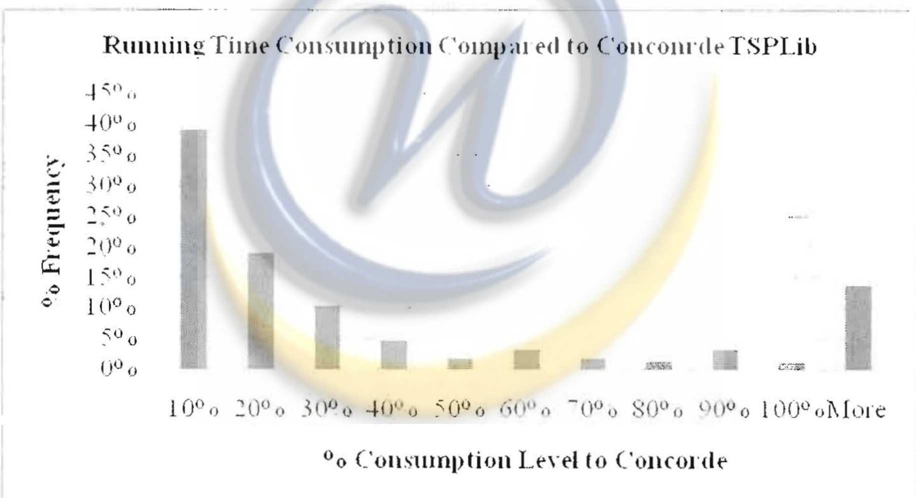

Fig. 3. Number of nodes per clusters fror random sample with n=10

5 Conclusion

There has been a long-term interest in the field of operations research in devising economical procedures for finding solutions that approximate the optimal, and many such heuristic techniques have been developed and compared. To achieve "reasonable" approximations to an optimal solution to within a few percentage points above the shortest path, such procedures generally need to perform on the order of n^3 calculations [10]. Based on explanation above, the primary objective of this development is to produce a better hierarchical clustering strategy that fit into such Euclidean TSP solution method. This approach has been successfully achived the objective i.e. algorithm most near $O(n^3)$ and average final TSP solution is about 6% compare to the best known solution.

This research has result contribution in term of Euclidean TSP and manufacturing problem. The contribution to Euclidean TSP is new solution approach using hierarchical clustering. The result of this research implemented to development automatics compact PCB Drilling machine that can be applied to Small Medium Industry. The algorithm is combined to this machine to establish that drilling order in shortest and effective ways. Time and cost saving will be resulted by this machine.

This research had resulted some new idea and improvement in clustering process and Euclidean TSP approach. The objectives of this research had been achieved with several notes. Outcome of this research had been produced for several areas such as clustering method, heuristics for Euclidean TSP and implementation to manufacturing problem.

References

1. Haxhimusa, Y., Krospatch, W.G., Pizlo, Z., Ion, A.: Approximate Graph Pyramid Solution of the E-TSP. Image and Vision Computing Journal (2008)
2. Abu, N.A., Sahib, S., Suryana, N.: A Novel Natural Approach to Euclidean TSP. In: Proceeding of The 3rd International Conference on Mathematics & Statistics (ICoMS) (2008)
3. Graham, S.M., Joshi, A., Pizlo, Z.: The Traveling Salesman Problem: a Hierarchical Model. Journal of Memory and Cognition 28(7), 1191–1204 (2000)
4. Vickers, D., Butavicius, M., Lee, M., Medvedev, A.: Human Performance on Visually Presented Traveling Salesman problems. Journal of Psychological Research 65, 34–45 (2001)
5. Punnen, A.P.: The Traveling Salesman Problem: Application, Formulations and Variations. In: Putin, G., Punnen, A.P. (eds.) The Traveling Salesman Problem and Its Variations, p. 22. Springer, Heidelberg (2007)
6. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, Montreal, Canada (1996)
7. Guha, S., Rastogi, R., Shim, K.: CURE: An Efficient Clustering Algorithm for Large Databases. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, Seatle, WA (1998)
8. Guha, S.: ROCK: A Robust Clustering Algorithm for Categorical Attributes. Journal of Information System 25(5), 345–366 (2000)
9. Reinelt, G.: TSPLIB–A Traveling Salesman Problem Library. Informs Journal on Computing 3, 376–384 (1991)
10. Golden, B., Bodin, L., Doyle, T., Stewart, W.: Approximate travelling salesman algorithms. Operations Research 28, 694–711 (1980)