Visualization of Large-Scale Weighted Clustered Graph: A Genetic Approach

Jiayu Zhou and **Youfang Lin** and **Xi Wang**
School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, P. R. China
zhoujiayu@computer.org, yflin@bjtu.edu.cn, starchiwang@gmail.com

Abstract

In this paper, a bottom-up hierarchical genetic algorithm is proposed to visualize clustered data into a planar graph. To achieve global optimization by accelerating local optimization process, we introduce subgraph rotating and scaling processes into the genetic algorithm. Compared with existing methods, the proposed approach is more feasible and promising, with more accurate graph layout and more satisfiable computationally efficient performance, as demonstrated by the experimental results.

Introduction

The soaring IT development inspires the generation of large scale data and usually the data are needed to process and analyze further, and even visualize. For example, the social network analysis (SNA), needs to visualize weighted relationships among a group of actors according to clustered relationship data which are obtained from various data sources. The geometric shapes and measures, such as cluster densities, of resulting graphs can be used for SNA processes, for instance, to find out the important roles or links in an organization. Visualizing chemical structures is another example.

Many algorithms for graph visualization have been proposed, and an extensive survey was done by Eades and Tamassia. Heuristic approaches were introduced to graphic drawing in earlier researches (Eades 1984), and of which, genetic algorithm (GA) show its excellent flexibility to complex constrains, but at the same time encounters serious performance issues, which made it hard to be applied to large-scale applications (Vrajitoru 2007). The visualization of structured information has been studied by Sugiyama and Misue (1991), and recently force-directed approaches have been introduced to visualize clustered data as well (Romain, David, and Mary 2007). Those approaches, however, limit flexibility to complex constrains, compared with GA algorithm.

To improve the performance of the GA algorithm, a new approach is presented here to draw large-scale weighted clustered planar graphs by hierarchical applying GA to clustered data, through which we can take advantage of the flexibility of GA and simultaneously achieve computational efficiency. Experimental results have proved that our method can achieve a high efficiency and accurately visualize the relationships in terms of correlation we defined.

Hierarchical Genetic Visualization

The idea of hierarchical GA (HGA) is to perform heuristic local optimization processes on sub-graphs in clustered data and to approximate global optimization through rotating and scaling processes.

Bottom-Up Drawing

Given a clustered data set, we firstly apply traditional GA to ground-level clusters (i.e. leaf nodes of the hierarchal clustering tree) to allocate positions of simple nodes. Those nodes are then abstracted into centroids, which represent centers of lower-level clusters.

Secondly, all centroids will be processed by rotating and scaling, which are described in the next section. The relationships and attributes between/of sub-nodes are abstracted and then mapped into features between/of their centriods. Finally, GA is hierarchically applied into those centroids, as if they are simple nodes.

Sub-Graph Rotating and Scaling

In order to approximate global optimization through applying hierarchal local optimization processes, we introduce two processes in visualization, rotating and scaling, both of which are incarnated as parts of our genetic algorithm.

Rotating is used to decide an angle to what extend a sub-graph should bias from its initial state. By rotating the sub-graph, we can find out a very angle that maximizes a fitting function in the given shape of sub-graph calculated. In a large scale graph, the rotating angle of a sub-graph is defined as the result of fitness functions, which is calculated among all the nodes of this sub-graph and all the centroids, rather than all the nodes, of other sub-graphs in the same cluster hierarchy. This approximation, using the extracted information of sub-nodes, could dramatically improve the computational efficiency. Similarly, scaling process adjusts the size of sub-graphs to the extend that maximizes the result of fitness function.

Empirical study in the experiment section shows that the approximation assumption of global optimization approximation is reasonable.

Hierarchical Genetic Algorithm

HGA is primarily different from traditional GA approaches in the design of coding and fitness function.

Coding. By integrating rotating and scaling functions into GA, the coding becomes a 4-tuple (x, y, r, s) to represent a graph node or a cluster centroid, where (x, y), r, s stand for the planar position, rotation angle and scaling ratio respectively. Especially, when applying GA to a cluster, a coded structure of the centroid will be added.

Fitness Function. Given a clustered data set C and the corresponding graph $G = \{V, E, W\}$, where V is the set of vertices, E is the set of edges, and W denotes the set of weights, then the nodes with closer distance have greater weights, and is thought to be good G, which leads to a minimum deviation of d_i/w_i (where $d_i \in D$, $w_i \in W$, D is the set of distances based on which we draw edges on the graph.), in this way, nodes within same cluster are closer than those outside the cluster and thus the hierarchal relations can be presented in the graph. Because scaling and rotating are used here, a normalize process is required (we used the Sigmod in our experiment), yielding the following correlation relationship:

$$Cr(C) = \sigma(Normalize(\frac{D}{W}))$$

For any cluster C_i, we use fitness function to find the optimal set of node positions $\mathcal{P}_{i*}$ in the drawing space $\mathcal{S}$.

$$\mathcal{P}_{i*} = argmin_{\mathcal{P}_i \in \mathcal{S}}(w_1 \times Cr(C_i) + w_2 \times Cr(SR(C_i)))$$

where function SR performs linear transformation to the cluster in order to find the optimal scaling ratio and rotating angle. Weight vector (w_1, w_2) specifies effects of each kind of constraints, depending on different applications.

Experiment Result

To demonstrate the performance of the proposed hierarchical genetic algorithm (HGA), we compared it with Planar GA (PGA), Force-Directed Algorithm(FDA), which was introduced by Romain, David, and Mary (2007), and random drawing in terms of correlation and compared with PGA in efficiency.

Results are shown in Figure 1. In the correlation experiment, HGA with RS ranks second, dropping to 0.65 as nodes reaches 5000, comparing to poor 0.3 without RS, which proves that RS has significantly improved the correlation, although second to the remarkable correlation of PGA, which is above 0.9. Not taking account of correlation, FDA performs as poor as that of random. As the number of nodes increasing, correlations of the top 3 algorithms keep decreasing. It is naturally harder to compromise a high global optimization when the graph is getting larger.

Acceleration rates of computation time of HGA with and without RS are prone to be linear in large-scale application, comparing to an exponential increment of PGA. It takes

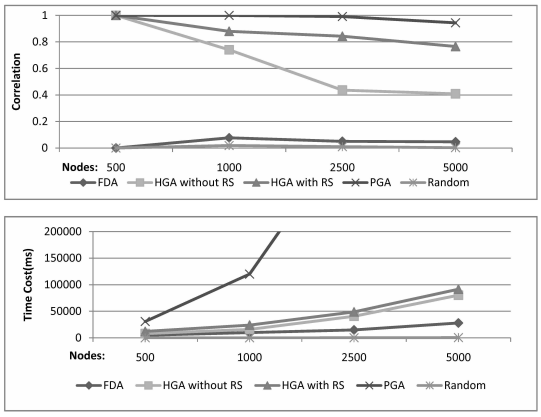

Figure 1: Average results of 100 times experiments on correlation (above) and efficiency

91.56 sec. to draw a 5000-node clustered graph with HGA including an addition 11.22 sec. to perform RS, indicating that the time cost of RS is acceptable. The PGA, on the other hand, spends over 1007.75 sec. Besides, it can be found that the HGA is very sensitive to depth of a certain cluster.

Conclusion

In this paper, a novel approach is proposed to visualize large-scale weighted clustered graph by a hierarchical genetic algorithm. Integrating rotating and scaling into the GA algorithm, the methods could approximate global optimization. It is the contribution of this study to apply HGA to large-scale visualization. Experimental results show that the method draws clustered graph remarkably faster than traditional genetic algorithms, and can possesses a practical high correlation at the same time.

References

Eades, P., and Tamassia, R. 1989. Algorithms for automatic graph drawing: An annotated bibliography. Technical Report CS-89-09, Dept. Comput. Sci., Brown Univ.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium* 42:149.

Romain, B.; David, A.; and Mary, P. 2007. How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm. In *Proc. of the eleventh Int. Conf. Information Visualization (IV'07)*.

Sugiyama, K., and Misue, K. 1991. Visualization of structure information: automatica drawing of compound digraphs. *IEEE Trans. on System, Man and Cybernetics* 21:876.

Vrajitoru, D. 2007. Hybrid multiobjective optimization genetic algorithms for graph drawing. In *Proc. of the Genetic and Evolutionary Computation Conference (GECCO'07)*.