MST-Pathfinder Network Scaling

Description

Network scaling algorithms such as the Pathfinder algorithm are used to prune many different kinds of networks, including citation networks, random networks, and social networks. However, this algorithm suffers from run time problems for large networks and online processing due to its $O(n^4)$ time complexity. Hence another alternative, MST-Pathfinder algorithm, is used which prunes the original network to get its PFNET(infinity,n-1) in just $O(n^2 * \log n)$ time.

The underlying idea comes from the fact that the union (superposition) of all the Minimum Spanning Trees extracted from a given network is equivalent to the PFNET resulting from the Pathfinder algorithm parameterized by a specific set of values (r = infinity and q = n-1), those usually considered in many different applications.

Pros & Cons

The algorithm has a space complexity of $3 * n^2 + n$, which is very less compared to the Original Pathfinder approach's complexity of $2 * (n^3 - n^2)$. Also its time complexity is $O(n^2 * \log(n))$ compared to $O(n^4)$ of the Original Pathfinder approach. According to an experimental setup, this efficiency enables MST based approach to scale the network with 10,000 nodes and 100 million edges in approximately 129 seconds compared to more than 222 hours using Original approach.

However, since it only provides an optimal solution the user cannot control the parameters,

1. ***q*** which constrains the number of indirect proximities examined in generating the network. The q parameter is an integer value between 2 and n-1, inclusive where n is the number of nodes or items. Value of q is fixed to n-1.
2. ***r*** defines the metric used for computing the distance of paths. The r parameter is a real number between 1 and infinity, inclusive. Value of r is fixed to infinity.

This, however, is mitigated by the fact that the networks resulting from citation, cocitation, or term co-occurrence analysis are usually dense when the categories are used as the unit for each node. Due to this fact, and especially in the case of vast scientific domains with a high number of entities (categories in our case) in the network, Pathfinder is usually parameterized to r = infinity and q = n-1, obtain an schematic representation of the most outstanding existing information by means of a network showing just the most salient links.

One of the drawbacks of this algorithm is that it cannot be applied on directed networks, because the sorting process deals with undirected links. Scaling of Directed networks using Pathfinder algorithm can be implemented in the future.

Implementation Details

Once the input file is validated the algorithm is started. It works as follows,

1. When reading through the nodes, initiate clusters (one for each node).
2. Next, when reading through the edges create tuples containing unique edge id, node 1, node 2 & edge weight. Depending upon users choice of how edge weight should be represented, normalize the edge weight i.e. If they chose similarity, normalize edge weights to be *Dissimilarity* based by inverting the edge weight.
3. Sort edge tuples based on their weight.
4. Iterate over the edge tuple i.e. for each edge e(u, v) do,
    a. For all the edges e(u', v') remaining in the edge list having edge weight equal to the current edge e(u, v) do,
        i. Remove e(u', v') from the edge list
        ii. If nodes u' & v' are not in the same cluster then,
            1. Add e(u', v') to the final edges list, which holds the edges which will belong to the final scaled network.
            2. Merge the clusters belonging to u' & v' such that smaller cluster is merged with the bigger cluster & smaller cluster is deleted.
5. After all the edges are considered output the final edges list to a new .NWB file. This will contain the nodes from the original network and only those edges which are present in the final edges list.

Usage Hints

The user has to provide 3 inputs; the file storing the information about the network, the column name that represents the edge weight and how the edge weight should be considered (Dissimilarity or Similarity). The provided network should not contain edge weights of less than equal to 0 else a warning is generated and default edge weight is assumed for that edge. The algorithm assumes the edge weights to be a measure of dissimilarity i.e. More the weight, more is the dissimilarity. If the input edge weights are a measure of Similarity i.e. More the weight, less is the similarity then the user can select *Similarity* option from the drop-down box. Only in the case of user selecting *Similarity* the edge weights will be inverted. For example, in a co-authorship network, more number of articles authored by 2 subjects equates to a stronger relationship. In this case the user should select *Similarity* option. The output will be provided as a .NWB file with a pruned network.

Links

- Source Code

Acknowledgments

The original Pathfinder Network Scaling algorithm authored by Roger Schvaneveldt et al. was adapted with a MST based approach by *Arnaud Quirin et al.* is implemented, integrated and documented by Chintan Tank. Also thanks to Micah Linnemeier and Russell Duhon for providing guidance during design and implementation. For the description I acknowledge Wikipedia.

References

1. Schvaneveldt, Roger., Durso, Francis., Dearholt, Donald. Graph theoretic foundations of pathfinder networks. In *Computers & mathematics with applications*, volume 15, pages 337-345, 1988. Link
2. Schvaneveldt, Roger., Durso, Francis., Dearholt, Donald. Network structures in Proximity Data. 1989. Link
3. Schvaneveldt, Roger. Pathfinder associative networks : Studies in Knowledge organization. 1990. Link
4. Guerrero-Bote, Vicente., Zapico-Alonso, Felipe., Espinosa-Calvo, Maria Eugenia., Crisostomo, Rocio Gomez., Moya-Anegnon, Felix. Binary Pathfinder: An improvement to the Pathfinder algorithm. In *Information processing & management*, volume 42, pages 1484-1490, 2006. Link
5. Quirin, Arnaud., Cordon, Oscar., Guerrero-Bote, Vicente., Vargas-Quesada, Benjamin., Moya-Anegnon, Felix. A quick MST-based algorithm to obtain Pathfinder networks ((infinity), n - 1). In *Journal of the American Society for Information Science and Technology*, volume 59, pages 1912-1924, 2008. Link
6. Quirin, Arnaud., Cordon, Oscar., Santamaria, J., Vargas-Quesada, Benjamin., Moya-Anegnon, Felix. A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. In *Information processing & management*, volume 44, pages 1611-1623, 2008. Link

See Also