ABSTRACT

For the past several years, large graph-mining applications have been used extensively over a wide-range-of analyses. One such application is a Water Distribution System (WDS). WDS is networked with a large number of pipes and junctions which have both behavioural as well as topological complexity. Therefore, it becomes highly expensive to design such large WDSs. Hence, the work proposes to construct a reduced WDS network with minimum cost by applying a minimum spanning tree (MST) approach. But, applying traditional MST algorithms to large graphs becomes difficult because of their high topological complexity and time complexity of \(O(N^2\log N)\). Thereby the proposed approach explains a divide-and-conquer mechanism to find approximate MST that results to a refined MST with reduced topological structure. Thus the suggested approach shows the novelty of the proposed algorithm by decreasing the time-complexity to \(O(N^{1.5}\log N)\).

In large graph application like WDS, it becomes necessary to reduce the behavioural complexity by considering the node or junction properties and edge or pipe properties for identification of the crucial components in a community. Hence cluster analysis is performed on reduced and refined MST by considering both the vertex and edge properties. The reduced MST is clustered by considering the attributes of the node properties like node id, elevation, demand, head, pressure etc., and attributes of the link properties like link id, diameter, roughness, flow, velocity etc... Clustering results obtained based on the nodes or junctions are useful for identification of critical components of junctions whereas clustering results obtained based on the links or pipes are useful for identification of pipe problems which helps for an efficient maintenance of WDS with minimum time and cost. Thus the clustering approach reduces the leakages and thereby improving the water quality.
Experimental results of the proposed system shows its better performance, as well as fits the expected result for both hypothetical data and real time data.

**Keywords:** Large graph, graph partitioning, minimum spanning tree, water distribution system, clustering, k-means clustering, K-spanning clustering, subgraph, frequent subgraph.