Semi-Supervised Hierarchical Clustering Given a Reference Tree of Labeled Documents

Authors : Ying Zhao, Xingyan Bin

Abstract : Semi-supervised clustering algorithms have been shown effective to improve clustering process with even limited supervision. However, semi-supervised hierarchical clustering remains challenging due to the complexities of expressing constraints for agglomerative clustering algorithms. This paper proposes novel semi-supervised agglomerative clustering algorithms to build a hierarchy based on a known reference tree. We prove that by enforcing distance constraints defined by a reference tree during the process of hierarchical clustering, the resultant tree is guaranteed to be consistent with the reference tree. We also propose a framework that allows the hierarchical tree generation be aware of levels of levels of the agglomerative tree under creation, so that metric weights can be learned and adopted at each level in a recursive fashion. The experimental evaluation shows that the additional cost of our contraint-based semi-supervised hierarchical clustering algorithm (HAC) is negligible, and our combined semi-supervised HAC algorithm outperforms the state-of-the-art algorithms on real-world datasets. The experiments also show that our proposed methods can improve clustering performance even with a small number of unevenly distributed labeled data.

Keywords : semi-supervised clustering, hierarchical agglomerative clustering, reference trees, distance constraints

Conference Title : ICIKM 2015 : International Conference on Information and Knowledge Management

Conference Location : Zurich, Switzerland

Conference Dates : January 13-14, 2015

Open Science Index, Information and Communication Engineering Vol:9, No:01, 2015 publications.waset.org/abstracts/19478.pdf