

Research on some semi-supervised clustering methods applied to the problem of clustering web server logs data

Nguyen Duc Ngoc

Abstract: This study investigates several semi-supervised clustering methods and their application to the problem of clustering web server log data. While traditional unsupervised clustering can reveal inherent structures in log data, it often lacks the precision needed for specific analytical goals. Semi-supervised clustering addresses this limitation by incorporating limited prior knowledge, such as a small set of labeled logs or pairwise constraints, to guide the clustering process. The research explores and evaluates the effectiveness of constraint-based and metric-based semi-supervised approaches, adapting them to the high-dimensional and noisy characteristics of web server logs. Experimental results on real-world log datasets demonstrate that the proposed semi-supervised methods significantly improve clustering accuracy and relevance compared to purely unsupervised baselines. The findings highlight the practical value of integrating domain knowledge into the clustering process for tasks such as user behavior profiling, anomaly detection, and session identification, offering a robust framework for more interpretable and actionable web log analysis.

Keywords: semi-supervised clustering, web server logs, data clustering, machine learning