Chapter 12
Hierarchical Clustering in Social Sciences and Humanities: A Review

Nikos Koutsoupias
https://orcid.org/0000-0003-1664-5404
University of Macedonia, Greece

Marios Nosios
https://orcid.org/0009-0007-6161-3854
University of Macedonia, Greece

ABSTRACT

This chapter undertakes a systematic bibliometric analysis of hierarchical clustering applications within the social sciences and humanities, drawing upon publications indexed in the Scopus database from 1975 to 2024. The findings reveal the methodological versatility of clustering techniques in addressing complex research problems across a broad spectrum of academic disciplines. Thematic emphases on social network analysis and data management strategies underscore the centrality of clustering in elucidating intricate sociocultural systems. The United States, China, and the United Kingdom emerge as principal contributors, highlighting the global scope and interdisciplinary nature of this research domain. This analysis provides a comprehensive account of prevailing thematic trends, patterns of scholarly collaboration, and the intellectual architectures that underpin the field. It offers a robust foundation for both the methodological advancement of clustering techniques and their expanded application to pressing societal and academic challenges

DOI: 10.4018/979-8-3693-9400-7.ch012

INTRODUCTION

The rapid advancement of data collection technologies in the 21st century has led to an unprecedented proliferation of complex datasets. Researchers are increasingly faced with the challenge of analyzing large, multidimensional datasets to uncover underlying patterns, trends, and relationships. In response to these challenges, numerous clustering methodologies have been developed, with Hierarchical Clustering (HC) distinguished by its reliability and adaptability across a wide range of research contexts. This chapter builds upon a preliminary bibliometric analysis presented at the 12th Panhellenic Conference on Data Analytics and offers a systematic overview of the use and evolution of HC within the social sciences and humanities, providing insights into its methodological foundations, applications, limitations, and emerging trends.

As an unsupervised learning technique, HC is widely utilized in data analysis, particularly in disciplines dealing with large and complex datasets, to group data into clusters based on shared characteristics or similarities. Although this method was originally developed for the natural sciences, its ability to uncover complicated patterns and relationships has turned particularly popular within social sciences and humanities, in which datasets are often described by high heterogeneity and complexity (Jafarzadegan et al., 2019).

The assessment of similarity or distance between data points constitutes the foundation of HC, which is broadly classified into two primary types: agglomerative and divisive. Agglomerative Hierarchical Clustering (AHC) is the most common type following a bottom-up approach where each object starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy (Li et al., 2022; Murtagh & Contreras, 2011; Zhou et al., 2016). In contrast, Divisive Hierarchical Clustering (DHC) follows a top-down approach where all objects begin in a single cluster, which is recursively split into smaller ones. Both approaches involve calculating the distance between objects or clusters using a distance metric, such as Euclidean, or Manhattan distance, or cosine similarity (Roux, 2018). These distance metrics are crucial for quantifying the degree of similarity between data points, and the selection of an appropriate metric can profoundly impact the clustering results. Euclidean distance is commonly applied to numerical data, as it represents the straight-line distance between two points. In contrast, Manhattan distance is utilized when the absolute differences across dimensions are of importance, whereas cosine similarity is frequently employed in the analysis of text or vector data (Giordani et al., 2020).

The result of HC is typically represented as a tree-like diagram called a dendrogram, which visually depicts the nested clusters at different levels of granularity. An advantage of HC is that it does not require the user to specify the number of clusters in advance, unlike other clustering algorithms such as k-means. Instead,

EMG-Based Mobile Assessment System for Neck and Shoulder Fatigue
Pei Lun Lai, Hsiu-Sen Chiangand Qi-An Huang (2017). *International Journal of Big Data and Analytics in Healthcare (pp. 39-50).*
www.irma-international.org/article/emg-based-mobile-assessment-system-for-neck-and-shoulder-fatigue/204447