Some Clustering Methods on Dissimilarity or Similarity Matrices: Uncovering Clusters in WEB Content, Structure and Usage

Yves Lechevallier

INRIA-Paris-Rocquencourt
78153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr

Abstract Clustering is one of the most popular techniques in knowledge acquisition and it is applied in various fields including data mining and statistical data analysis. Clustering involves organizing a set of individuals into clusters in such a way that individuals within a given cluster have a high degree of similarity, while individuals belonging to different clusters have a high degree of dissimilarity.

The definition of *homogeneous cluster* depends on a particular algorithm: this is indeed a structure, which, in the absence of prior knowledge about the multidimensional shape of the data, may be a reasonable starting point towards the discovery of richer and more complex structures.

We propose an clustering method for partitioning a set of objects where the relation between two objects is described by a dissimilarity or similarity measures. The clustering criterion, based on the sum of weighted dissimilarities between the objects belonging to the same class, measures the homogeneity of the cluster. The mathematical properties of these weighted distances and to implement the corresponding algorithms which optimize the clustering criterion and an empirical framework to their evaluation will be studied The advantage of this approach is that the clustering algorithm recognizes different shapes and sizes of clusters.

Clustering is a valuable technique for analyzing the Web. We propose to study clustering approaches in Content and Structure Document Mining and Usage mining. The analysis of a web site based on its usage data is an important task as it provides insight into the organization of the site and its adequacy regarding user needs. We thus defined an approach for discovering the profiles of visitor groups.