Clustering Methods for Large Databases: From the Past to the Future

Alexander Hinneburg, Daniel A. Keim
University of Halle-Wittenberg
Kurt-Mothes-Str. 1, D-06120 Halle, Germany
Phone (+49) 345 5524711 - Fax (+49) 345 5527009
{hinneburg, keim}@informatik.uni-halle.de

1. INTRODUCTION

Because of the fast technological progress, the amount of information which is stored in databases is rapidly increasing. In addition, new applications require the storage and retrieval of complex multimedia objects which are often represented by high-dimensional feature vectors. Finding the valuable information hidden in those databases is a difficult task. Cluster analysis is one of the basic techniques which is often applied in analyzing large data sets. Originating from the area of statistics, most cluster analysis algorithms have originally been developed for relatively small data sets. In the recent years, the clustering algorithms have been extended to efficiently work on large data sets, and some of them even allow the clustering of high-dimensional feature vectors. Many such methods use some kind of an index structure for an efficient retrieval of the required data; other approaches are based on preprocessing for a more efficient clustering.

The main goal of the tutorial is to provide an overview of the state-of-the-art in cluster discovery methods for large databases, covering well-known clustering methods from related fields such as statistics, pattern recognition, and machine learning, as well as database techniques which allow them to work efficiently on large databases. The target audience of the tutorial are researchers and practitioners, who are interested in the state-of-the art of cluster discovery methods and their applications to large databases. The tutorial especially addresses people from academia who are interested in developing new cluster discovery algorithms, and people from industry who want to apply cluster discovery methods in analyzing large databases.

SIGMOD '99 Philadelphia PA

The tutorial is structured as follows: First, we give a brief motivation for clustering from modern data mining applications. We discuss important design decisions and explain the interdependencies with the properties of data. In the second section, we introduce a large variety of clustering methods and classify them into three groups — model- and optimization-based methods, linkage- and density-based methods, and hybrid methods. A detailed comparison shows the strength and weaknesses of the existing techniques and reveals potentials for further improvements. In the next two section, we discuss database techniques which have been proposed to improve the effectiveness and efficiency of the cluster discover process. The four main categories of techniques which can be used for this purpose are hierarchical and incremental approaches, multidimensional indexing, sampling, and condensation-based approaches. The tutorial concludes with a discussion of open problems and future research issues.

2. OUTLINE

1. Introduction

2. Clustering Methods

2.1 Optimization-based Methods
2.2 Density-based Methods
2.3 Hybrid-based Methods

3. Techniques for Improving the Effectiveness

3.1 Hierarchical Techniques
3.2 Sampling Techniques

4. Techniques for Improving the Efficiency

4.1 Sampling Techniques
4.2 Focussed Optimization Techniques
4.3 Multidimensional Indexing Techniques
4.4 Condensation-based Techniques

5. Open Problems and Future Research Issues

3. TUTORIAL NOTES

http://www.informatik.uni-halle.de/~keim/SIGMOD99Tut.ps.gz