1. 引言

本章主要通过论文课题的背景进行介绍，对选题进行分析，阐述国内外对聚类算法并行化的研究现状及其在各个领域的应用发展，分析和证明了本课题的研究意义和应用价值，根据以上内容引出本文的结构安排和章节介绍。

1.1 课题研究背景及意义

随着计算机硬件的高速发展以及Web 2.0时代的来临，网络终端每时每刻都产生着浩如烟海的数据，然而我们却缺乏对其充分的理解和应用，传统的数据分析方法已经不能满足海量数据分析和处理的要求。于是，数据挖掘技术应运而生。数据挖掘，比较公认的定义，指的是从大量不完全的、有噪声的、模糊的、随机的数据中，提取出隐含的、事先不为人所知道的、却又是潜在有用的知识和信息的过程[1]。它是一种在海量数据中寻找规则或者模式的过程，是一个新兴的并且具有广阔应用前景的研究领域。

聚类分析是数据挖掘法技术中重要组成部分，可以有效的分析数据并从中发现有用的信息。聚类分析是指根据数据中对象及其之间的关系，将数据对象分组。其目标是，使组内的对象相互之间是相似的（相关的），而不同组中的对象是不同的（不相关的）。组内的相似性（同质性）越大，组间差别越大，聚类效果就越好[2]。它广泛应用于多个领域，如文本聚类、模式识别、人工智能、市场分析、医疗卫生、图像分析和信息检索。

由于数据挖掘是从海量数据中提取有用信息，处理效率问题成了对海量数据处理的瓶颈之一，传统的单机串行算法效率较低；由于部分聚类算法中蕴涵并行性，所以为了解决处理效率问题，将并行化的程序设计思想（并行处理）引入聚类算法，同时降低算法的复杂度，使用集群系统进行并行计算，从而有效的缩短聚类的时间。

Hadoop[3]是一个开源的分布式云计算平台，能够实现对大量的数据集高效、可靠、可伸缩的分布式并行处理。而Hadoop中的MapReduce 编程模式是将已有单机算法实现分布式的关键，通过实现 MapReduce编程模式，我们就可以方便的把已有的算法移植到Hadoop平台实现算法的并行化。

当前，MapReduce在数据挖掘领域被广泛应用，出现了很多基于MapReduce平台的聚类算法。然而随着数据量的进一步增加，实际应用需求的差异，以及实际项目中数据集的不同，针对数据挖掘中的诸多问题，除了研究新的聚类算法以外，针对具体应用需求对现有聚类算法进行改进并移植到 Hadoop 平台上进行分布式实现，从而提高对大规模数据集处理的扩展性，也