# 引言

本章主要对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即将到来的AI/VR时代，网络客户端和服务器每时每刻都在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。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。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，也非常的有效且相对方便，成为当前研究的重要方向，具有十分重大的意义。

## 1.2 国内外发展现状

MapReduce作为云计算的核心技术之一，为并行系统的数据处理提供了一个简单、优雅的解决方案[4]。近几年来，随着数据规模的急速扩大，MapReduce受到了较多的关注，获得了较大的发展，虽然还没有形成成熟的、系统化的理论体系[5]，但机器学习领域中的许多算法都已在MapReduce框架下实现了并行化，并获得了较好的执行效率。

论文[6]针对数据密集型计算环境下数据具有海量、分布、异构、高速变化等特点，分析传统的基于密度的分布式聚类(Density Base Distributed Clustering，DBDC)算法，借助MapReduce编程模型，提出一种新的分布式聚类算法，采用局部和全局的方式处理海量、异构数据，解决具有以上特点的数据密集型计算环境下数据的分析挖掘问题。论文[7]提出了一种基于开源云计算平台Hadoop的网络热点话题发现方案。该方案采用MapReduce分布式并行计算架构处理海量、复杂数据，通过将命名实体词作为文本的特征项，并采用标题和正文的双向量表示文本。实验结果表明随着参与并行计算节点数的增加，话题聚类所用的时间显著下降，因而网络热点话题发现的速度得到明显提高。论文[8]利用MapReduce并行计算模型对Jarvis-Patrick(JP)聚类算法进行并行化设计，用以解决大规模海量文本聚类分析中具有的文本稀疏度和维度高的问题。以搜狗语料库为基础，使用Hadoop平台对算法进行了具体实现，实验结果表明并行运行的Jarvis-Patrick算法与单机环境相比，在处理海量文本数据时拥有更高的运行效率。论文[9]将MapReduce并行计算模型应用于朴素贝叶斯，K-modes和ECLAT算法中。这些算法分别是数据挖掘领域中常用的用以分类，聚类和挖掘频繁项目集的算法。实验结果显示，在确保算法相同准确率的条件下，利用MapReduce模型对算法进行并行化改造可以显著提升算法的执行速度。论文[10]提出一种新的适合于分布式计算的最小生成树算法，结合适合的相似度度量，设计了一种用于解决海量数据分析的分布式聚类算法，并给出了基于MapReduce编程模型的分布式实现。论文[11]提出了一种关系型数据库ChunkDB。该数据库采用了分块结构，是一种分布式数据库，主要用于解决关系型数据库系统与MapReduee并行计算框架之间存在的兼容问题,实验结果表明MapReduce与ChunkDB数据库的结合可以提供高效率的并行化查询。论文[12]设计了一种基于MapReduce计算框架的DeepWeb并行处理算法，将MapReduce并行计算框架应用于DeepWeb爬虫模型。在由虚拟机搭建的计算机集群上对算法进行并行实现，实验结果表明该算法提高了搜索发现数据源的性能，优化了各种资源的使用，证明了将并行处理方法应用于DeepWeb爬虫的有效性。

## 1.3 本文章节安排

本文将分为六个章节对课题进行阐述，各章的内容安排如下：

第一章：引言。

第二章：课题采用关键技术。

第三章：基于曲线相似度的供热过程评价。

第四章：聚类算法的并行化分析。

第五章：实验及结果。

第六章：供热评价系统。

第七章：总结和展望。

# 相关技术介绍

## 2.1 数据挖掘

## 2.2 聚类算法

## 2.3 Hadoop

## 2.4 MapReduce

# 数据挖掘相关应用研究

## 3.1 数据挖掘在供热方面的应用

## 3.2 数据挖掘在其他方面的应用

# 基于曲线相似度的供热行为评价方法

# 聚类算法并行化分析

# 实验及结果

## 6.1 曲线相似度实验

## 6.2 聚类并行化实验

# 总结与展望

## 7.1 总结

## 7.2 展望