# 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

### 1.2.1 数据挖掘技术的发展

数据挖掘作为一个新兴的多学科交叉应用领域，正在各行各业的决策支持活动中扮演着越来越重要的角色，随着信息技术的迅速发展，各行各业都积累了海量异构的数据资料，这些数据往往隐含着各种各样有用的信息，仅仅依靠数据库的查询检索机制和统计学方法很难获得这些信息，迫切需要将这些数据转化成有用的信息和知识，从而达到为决策服务的目的。数据挖掘分析得到的信息和知识现在已经得到了广泛的应用，例如商务管理、生产控制、市场分析、工程设计和科学探索等。数据挖掘是一个多学科领域，它融合了数据库技术、人工智能、机器学习、统计学、只是工程、信息检索等最新技术的研究成果。

随着海量数据的增加，大数据时代的到来，导致人们对数据的研究和利用越来越多，其中数据挖掘技术的不断进步和发展也给整个世界信息的发展带来了许多成果，在科学领域方面，先进的现代化科学观测仪器的使用造成每天都要产生巨量的数据，如各种同步卫星每小时传回地球的遥感图像数据就达50千兆字节。天文学上有一个很著名的应用系统—SKICAT，这是第一个相当陈宫的数据挖掘应用，也是人工智能技术在天文学和空间科学上第一批成功的应用之一。目前，科学家已利用SKICAT发现了16个新的极其遥远的类星体。在市场营销方面，条形码技术在商业上的普遍使用使得很多行业每天都积累了大量数据，从市场营销来说，通过数据分析了解客户购物行为的一些特征，对提高竞争力及促进销售是有很大帮助的。在金融投资方面，目前国内有很多进行股票分析的软件，并且定期有专家进行股票交易预测。数据挖掘技术还可以应用在甄别诈骗上，进行诈骗甄别主要是通过总结正常行为和诈骗行为之间的关系，得到诈骗行为的一些特征，这样当某项业务符合这些特征时，可以向决策人员提出警告。这方面比较陈宫的系统有FALCON和FAIS系统。在Web应用上，世界上最强大的搜索引擎Google相比其他很多搜索引擎，它的搜索结果更让人满意，其中Google使用的搜索算法主要是PageRank算法，在2001年9月被授予美国专利，Google的PageRank是根据网站的外部链接和内部链接的数量和质量两相衡量网站的价值。

现今，数据挖掘的发展趋势主要在一下几个方面：

数据挖掘语言的标准化：语言的标准化对于数据挖掘系统的开发和数据挖掘技术的普遍使用是至关重要的。其可改进多个数据挖掘系统和功能间的互操作，促进其在企业和社会中的使用。

数据挖掘的可视化：可视化要求已经成为数据挖掘系统中必不可少的技术。可以再发现知识的过程中进行很好的人机交互。数据的可视化起到了推动人们主动进行知识发现的作用。

分布式数据挖掘。分布式技术的到来为日益正常的数据提供了有力的支持，而分布式数据挖掘中将分布式技术和数据挖掘技术的结合，也使对分离数据库的可协作数据挖掘工作开发了一个重要领域。

数据挖掘与数据库系统和Web数据库系统的集成：数据库系统和Web数据库已经成为信息处理系统的主流。数据挖掘系统的理想体系结构是与数据库和数据仓库系统的紧耦合。

挖掘复杂数据类型的新方法：挖掘复杂数据类型是数据挖掘的重要前沿研究课题，也有人称复杂类型的数据挖掘是“下一代数据挖掘”。伴随着数据的增多，需要处理的数据类型也变得越来越发扎，例如数据流、时间序列、时间空间、多媒体和文本数据、虽然现在在很多复杂数据类型的挖掘方面取得了一些进展，但是在应用需求和可用技术之间仍然存在较大的距离。

数据挖掘中的隐私保护和信息安全：随着信息技术的发展，越来越多的数据涌入了网络，其中包括大量电子形式的个人信息，而挖掘技术的发展和科技的更新，在相反的一面上也使大量的个人信息收到了威胁，因此保护隐私的数据挖掘方法愈显重要。

### 1.2.2 MapReduce技术的发展

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

数据挖掘是从海量数据中挖掘出有趣的模式和知识的过程。其中数据源可能会包括数据仓库、数据库、Web存储、内存存储或动态的进入系统的数据。这里存在大量的数据挖掘模式，包括特征化与区分，频繁模式、关联和相关性挖掘，分类与回归，聚类分析，离群点分析。一般而言，这些任务可以分为两类：描述性和预测性。描述性挖掘任务刻画目标数据中数据的一般性质。预测性挖掘任务在当前数据上进行归纳，以便做出预测。

类/概念性描述：特征化与区分。数据特征化是目标类数据的一般特性或特征的汇总。通常通过查询来收集对应于用户指定类的数据。将数据汇总和特征化有一些有效的方法。基于统计度量和图的简单数据汇总，基于数据立方体的OLAP上卷操作可以用来执行用户控制的、沿着指定维的数据汇总。面向属性的归纳技术可以用爱进行数据的泛化和特征化，而不必一步步地与用户交互。数据特征化的输出可以用多种形式提供，例如饼图、条图、曲线、多维数据立方体和包括交叉表在内的多维表。数据区分是将目标类数据对象的一般特性与一个或多个对比类对象的一般特性进行比较。目标类和对比类可以由用户指定，而对应的数据对象可以通过数据库查询检索。

挖掘频繁模式、关联和相关性。频繁模式是在数据中频繁出现的模式。存在多种类型的频繁模式，包括频繁项藉、频繁子序列和频繁子结构。频繁项藉一般是指频繁的在事务数据集中一起出现的商品的集合，子结构可能涉及不同的结构形式，可以与项集或子序列结合在一起，如果一个子结构频繁的出现，则称它为结构模式。挖掘频繁模式导致发现数据中有趣的关联和相关性。通常，一个关联规则被认为是无趣的而被丢弃，如果它不能同时满足最小支持度阈值和最小置信度阈值。还可以做进一步分析，发现相关联的属性-值对之间的有趣的统计相关性。

用于预测分析的分类与回归。分类是这样的过程，它找出描述和区分数据类或概念的模型，以便能够使用模型预测类标号未知的对象的类标号。导出模型是基于对训练数据集的分析。该模型用来预测类标号未知的对象的类标号。分类预测类别标号，而回归建立连续值函数模型。也就是说，回归用来预测缺失的或难以获得的数值数据值，而不是（离散的）类标号。术语预测可以指数值预测和类标号预测。尽管还存在其他方法，但是回归分析是一种最常使用的数值预测的统计学方法。

聚类分析。不像分类和回归分析标记类的数据集，聚类分析数据对象，而不考虑类标号。在许多情况下，开始并不存在标记类的数据。可以使用聚类产生数据组群的类标号。对象根据最大化类内相似性、最小化类间相似性的原则进行聚类或分组。也就是说，对象的簇这样形成，使得相比之下在同一个簇中的对象具有很高的相似性，而与其他簇中的对象很不相似。所形成的每个簇都可以看做一个对象类，由它可以导出规则。聚类也便于分类法形成，即将观测组织成类分层结构，把类似的事件组织在一起。

离群点分析。数据集中可能包含一些数据对象，他们与数据的一般行为或模型不一致。这些数据对象是离群点。大部分数据挖掘方法都将离群点视为噪声或异常而丢弃。然而，在一些应用中，罕见的事件可能比正常出现的事件更令人感兴趣。离群点数据分析称作离群点分析或异常检测。

## 2.2 聚类分析

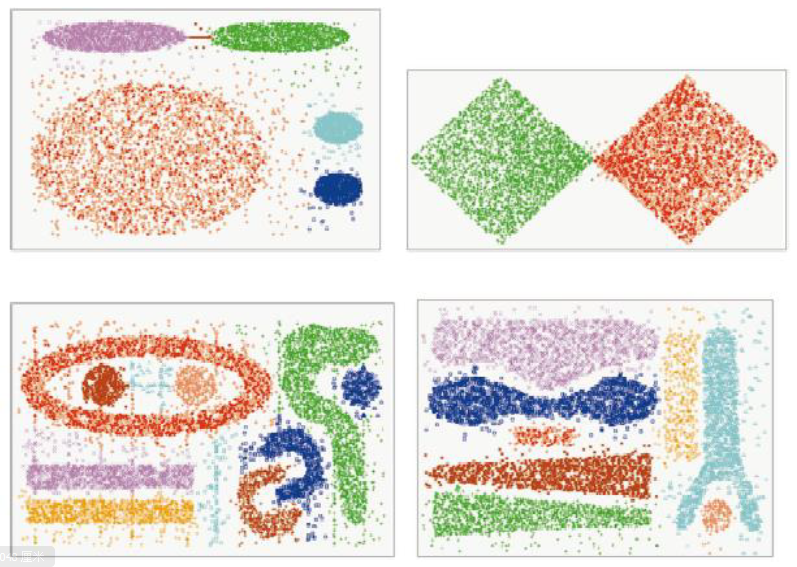
### 2.2.1 聚类分析的概念

聚类分析简称聚类，是一个把数据对象（或观测）划分成子集的过程。每个自己是一个簇，使得簇中的对象彼此相似，但与其他簇中的对象不相似。由聚类分析产生的簇的集合称作一个聚类。在这种语境下，在相同的数据集上，不同的聚类方法可能产生不同的聚类。划分不是通过人，而是通过聚类算法进行。聚类是有用的，因为它可能导致数据内事先未知的群组的发现。

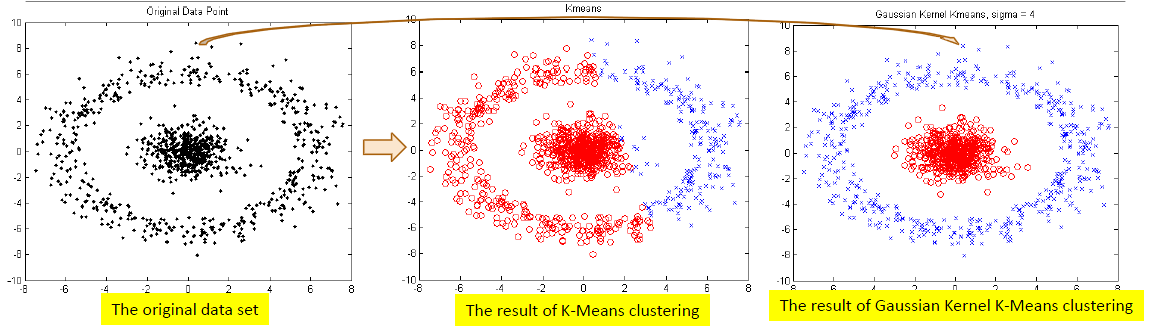
作为一种数据挖掘功能，聚类分析也可以作为一种独立的工具，用来洞察数据的分部，观察每个簇的特征，将进一步分析集中在特定的簇集合上。另外，聚类分析可以作为其他算法的预处理步骤，之后这些算法将在检测到的簇和选择的属性或特征上进行操作。

在某些应用中，聚类又称作数据分割，因为它根据数据的相似性把大型数据集合划分成组。聚类还可以用于离群点检测，其中离群点（“远离”任何簇的值）可能比普通情况更值得注意。

### 2.2.2 基于层次的聚类算法

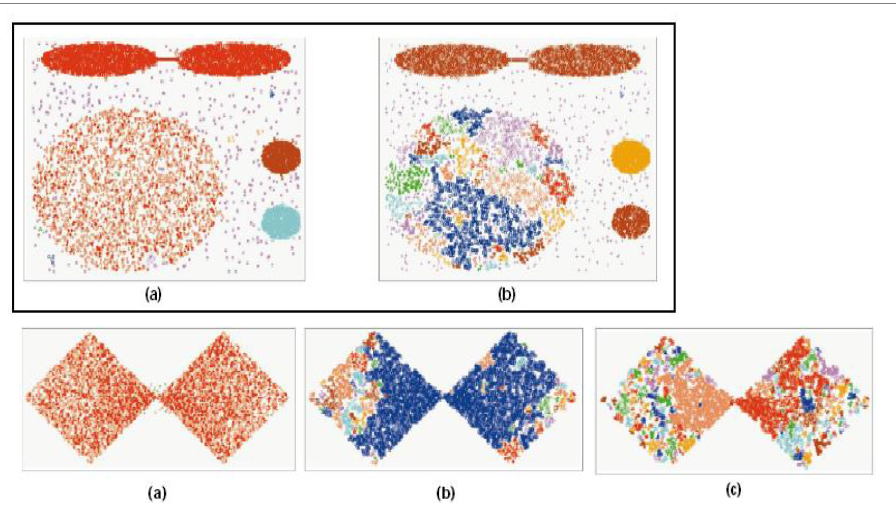
层次聚类方法将数据对象组成层次结构或簇的“树”。该方法主要有两种路径：凝聚和分裂，也可以理解为自底向上法（bottom-up）和自顶向下法（top-down）。凝聚法就是一开始每个个体（object）都是一个类，然后根据相似性度量寻找同类，最后形成一个“类”。分裂法就是反过来，一开始所有个体都属于一个“类”，然后根据相似性度量排除异己，最后每个个体都成为一个“类”。这两种路径本质上没有孰优孰劣之分，只是在实际应用的时候要根据数据特点以及自己想得到的“类”的个数，来考虑是自上而下更快还是自下而上更快。至于根据相似性度量判断“类”的方法就是最短距离法、最长距离法、中间距离法、类平均法等等（其中类平均法往往被认为是最常用也最好用的方法，一方面因为其良好的单调性，另一方面因为其空间扩张/浓缩的程度适中）。基于层次的方法中比较新的算法有BIRCH（Balanced Iterative Reducing and Clustering Using Hierarchies）主要是在数据体量很大的时候使用，而且数据类型是数字型；ROCK（A Hierarchical Clustering Algorithm for Categorical Attributes）主要用在categorical的数据类型上；Chameleon（A Hierarchical Clustering Algorithm Using Dynamic Modeling）里用到的相似性度量是kNN（k-nearest-neighbor）算法，并以此构建一个graph，Chameleon的聚类效果被认为非常强大，比BIRCH好用，但运算复杂度很高，O(n^2)。以下是Chameleon的聚类效果图，其中一个颜色代表一类，可以看出来是可以处理非常复杂的形状的。

### 2.2.3 基于划分的聚类算法

其原理简单来说就是，想象你有一堆散点需要聚类，想要的聚类效果就是“类内的点都足够近，类间的点都足够远”。首先你要确定这堆散点最后聚成几类，然后挑选几个点作为初始中心点，再然后依据预先定好的启发式算法（heuristic algorithms）给数据点做迭代重置（iterative relocation），直到最后到达“类内的点都足够近，类间的点都足够远”的目标效果。也正是根据所谓的“启发式算法”，形成了k-means算法及其变体包括k-medoids、k-modes、k-medians、kernel k-means等算法。k-means对初始值的设置很敏感，所以有了k-means++、intelligent k-means、genetic k-means；k-means对噪声和离群值非常敏感，所以有了k-medoids和k-medians；k-means只用于数值类型数据，不适用于类别类型数据，所以k-modes；k-means不能解决非凸（non-convex）数据，所以有了kernel k-means。另外，很多教程都告诉我们基于划分的聚类多适用于中等体量的数据集，但我们也不知道“中等”到底有多“中”，所以不妨理解成，数据集越大，越有可能陷入局部最小。下图显示的就是面对非凸，k-means和kernel k-means的不同效果。

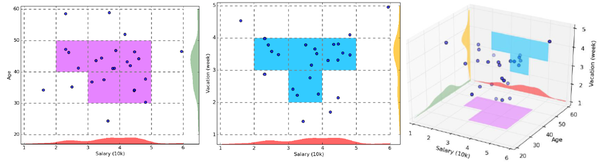
### 2.2.4 基于密度的聚类算法

由于k-means解决不了这种不规则形状的聚类。于是就有了基于密度的聚类算法来系统解决这个问题。该方法同时也对噪声数据的处理比较好。其原理简单说画圈儿，其中要定义两个参数，一个是圈儿的最大半径，一个是一个圈儿里最少应容纳几个点。最后在一个圈里的，就是一个类。DBSCAN（Density-Based Spatial Clustering of Applications with Noise）就是其中的典型，可惜参数设置也是个问题，对这两个参数的设置非常敏感。DBSCAN的扩展叫OPTICS（Ordering Points To Identify Clustering Structure）通过优先对高密度（high density）进行搜索，然后根据高密度的特点设置参数，改善了DBSCAN的不足。下图是表现了DBSCAN对参数设置的敏感。



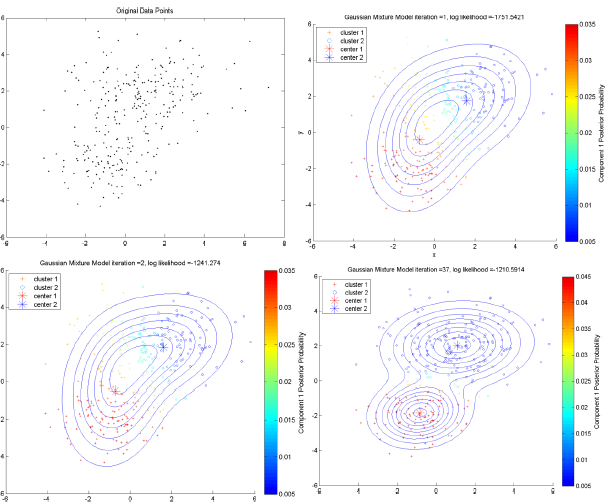
### 2.2.5 基于网格的聚类算法

基于网格的聚类算法原理就是将数据空间划分为网格单元，将数据对象集映射到网格单元中，并计算每个单元的密度。根据预设的阈值判断每个网格单元是否为高密度单元，由邻近的稠密单元组形成“类”。该类方法的优点就是执行效率高，因为其速度与数据对象的个数无关，而只依赖于数据空间中每个维上单元的个数。但缺点也是不少，比如对参数敏感、无法处理不规则分布的数据、维数灾难等。STING（STatistical INformation Grid）和CLIQUE（Clustering In QUEst）是该类方法中的代表性算法。下图是CLIQUE的一个例子：



### 2.2.6 基于模型的聚类算法

基于模型的聚类算法主要是指基于概率模型的方法和基于神经网络模型的方法，尤其以基于概率模型的方法居多。这里的概率模型主要指概率生成模型（generative Model），同一“类”的数据属于同一种概率分布。这中方法的优点就是对“类”的划分不那么“坚硬”，而是以概率形式表现，每一类的特征也可以用参数来表达；但缺点就是执行效率不高，特别是分布数量很多并且数据量很少的时候。其中最典型、也最常用的方法就是高斯混合模型（GMM，Gaussian Mixture Models）。基于神经网络模型的方法主要就是指SOM（Self Organized Maps）了，也是我们所知的唯一一个非监督学习的神经网络了。下图表现的就是GMM的一个例子，其中用到EM算法来做最大似然估计。



## 2.3 MapReduce

MapReduce是一个编程模型，也是一个处理和生成超大数据集的算法模型的相关实现。用户首先创建一个Map函数处理一个基于key/value pair的数据集合，输出中间的基于key/value pair的数据集合；然后再创建一个Reduce函数用来合并所有的具有相同中间key值得中间value值。现实世界中有很多满足上述处理模型的例子，也就是说MapReduce模型的提出，是基于现实世界中能满足这种处理模型的实际情况的存储。

其原理是：利用一个输入key/value pair集合来产生一个输出的key/value pair集合。MapReduce库的用户用两个函数表达这个计算：Map和Reduce。用户自定义的Map函数接受一个输入的key/value pair值，然后产生一个中间key/value pair值得集合。MapReduce库把所有具有相同中间key值I得中间value值结合在一起后传递给reduce函数。

用户自定义的Map函数接受一个中间key的值I和相关的一个value值得集合。Reduce函数合并这些value值，形成一个较小的value值得集合。一般的，每次Reduce函数调用只产生0或1个输出value值。通常我们通过一个迭代器把中间value值提供给Reduce函数，这样我们就可以处理无法全部放入内存中的大量的value值的集合。

# 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

随着供热行业自动化水平的不断提高，“煤改气”和“十三五”规划节能减排的进行，我国的城市供热过程基本上实现了自动控制，供热品质得到改善，能源利用率得到提高。很多学者已经利用数据挖掘等技术对锅炉系统历史数据分析。孙群丽等对锅炉运行数据进行关联规则挖掘，提供了几组在不同负荷及外部条件下的最优运行方式与参数控制[1]；路海昌等通过对时间序列进行相空间重构，建立了基于支持向量回归的时序数据预测模型，从而实现对锅炉输出参数的预测[2]；岳晓忠采用后向反馈BP神经网络理论和关联规则算法的数据挖掘方法，对锅炉实时运行数据进行分析，从而建立锅炉运行模型[3]。以上的分析都是如何去优化控制锅炉，但实际操作人员由于各种原因较难掌握这些方法的使用，而供热公司也没有合适的方法对现场人员供热过程量化分析，即缺乏切实可靠的量化评价方法，难以满足管理者实时掌握考评状况和调整运行策略的需求[4]。因此研究供热过程的评价方法对推进节能减排、降低运行成本都具有重要意义。

本文首先解释目标温度曲线的由来并分析目标温度曲线与实际出水温度曲线之间不同的多种可能情况，然后将其分解为趋势变化、平移和伸缩三个属性相似度并给出各个属性的相应计算方法，并最后融合为一个一致性度量来评价供热过程。通过实验分析，证明了此评价方法的有效性，为相关管理人员量化管理提供了一种参考依据，避免仅以燃气、水、电等能耗来衡量供热行为，而是需要区分不同的供热过程，在满足供热户室内温度的情况下尽可能节能。

1 目标温度曲线与出水温度曲线分析

1.1 目标温度曲线

早期锅炉系统管理人员根据运行经验，根据室外温度会制定一个出水温度标准作为供热锅炉出水温度的参考，通常与室外温度为线性相关。在此基础上一段连续时间的出水温度就构成了目标温度曲线。随着节能减排的规划以及物联网技术的应用，目标温度的定义需考虑各方面因素确定，主要包括天气条件（室外温度、风速、日照），供热用户不同时间段的需求，回水温度，出水提前量等。其中天气条件将直接影响供热用户的采暖需求，回水温度反映了热量的利用情况，而供热公司也需要根据供热用户的作息规律、生活习惯、上班或在家的情况进行适度的调控，尽量节约能源，出水提前量指热水热交换后到用户家里的时间。因此是一个比较复杂的过程，需要考虑较多因素，主要因素如公式（1）所示。

T(t)=f(T\_t 〖,W〗\_t 〖,S〗\_t 〖,O〗\_t 〖,B〗\_t,E\_t ) (1)

式中t为时间，T\_t 〖,W〗\_t 〖,S〗\_t 〖,O〗\_t 〖,B〗\_t,E\_t分别为t时间的室外温度、风速、日照、偏移量、回水温度、出水提前时间。

目标温度曲线主要是根据专家经验综合以上因素并参考供热用户建筑计算或对大量历史采集数据分析预测得到。

1.2 目标温度曲线与出水温度曲线相似度

将目标温度曲线（A）与出水温度曲线（B）进行一致性分析，可以分析出锅炉房供热过程是否严格按照要求供热，及满足节能要求，为锅炉管理人员对操作人员量化管理提供参考。

两条曲线的一致性存在多种情况，如图1所示。其中图1-(a)中曲线B1基本与曲线A相同，而曲线B2的在12时后较曲线A的趋势有明显差异，表明曲线B2基本没有按照曲线A进行相应的调整，即两条曲线反映的供热调节趋势不一致；图1-(b)表现了曲线间的横向平移问题，在供热数据上，横坐标的平移代表了出水温度与目标温度调整的提前和延时情况，纵坐标的平移代表了出水温度与目标温度的温差情况，即供热温度相比目标温度偏高或偏低。其中曲线B1在横坐标上较曲线A整体提前，而曲线B2在横坐标上较曲线A整体延后，表明虽然趋势调整基本正确，但调节的及时性存在问题；图1-(c)表现了曲线间的纵向平移问题，曲线B1在纵坐标上较曲线A低了近2摄氏度，表明供热温度偏低，未能达到供热户供暖需求，而曲线B2在纵坐标上较曲线A高了近2摄氏度，表明供热温度偏高，造成了能源浪费；图1-(d)中，曲线B1的最大值与最小值间横坐标差距相较于曲线A较小，说明该降温时间延后，而升温时间提前，造成了能源的浪费，而曲线B2的最大值与最小值间横坐标差距大于曲线A，说明降温时间提前而升温时间延后，未能达到供热户需求，容易引发客诉；图1-(e)中，曲线B1与曲线A相比，其最大值与最小值间的纵坐标差距更小，在高温时未能达到供热户需求，在低温时浪费热量，而曲线B2最大值与最小值间的纵坐标差距更大，在高温时浪费热量，在低温时未能达到供热户需求。

综合上述分析，两条曲线间主要存在趋势变化、平移和伸缩三个属性上不同。因此，出水温度曲线与目标温度曲线的一致性评价问题可以分解为对这三种属性相似度的计算问题。在供热数据中，曲线间的差异也代表了不同的供热行为：趋势变化主要反映两者在整体形态上的一致性，即操作人员能否按照要求进行合理的调控，横向时间的偏移体现了温度调整的提前或延后，纵向温度的偏移体现了供热温度的偏高或偏低；伸缩体现了曲线在升温和降温调整的不同。

2 供热效果评价

2.1 出水温度曲线与目标温度曲线的趋势变化

两条曲线趋势变化相同，可以认为锅炉系统能综合考虑各种因素对出水温度做出合理的调整，使供热户在不同的条件下得到最合适的热量。对两条曲线趋势的分析，可以认为是曲线的相似性问题[5]。

曲线的相似性测度一般有两种方法：距离测度法和相似性函数法[6]。相似性函数是用函数的方法来表征两曲线相似的程度，主要有夹角余弦和相关系数等方法，但由于供热数据受天气和地域影响较大，很难拟合成相应的曲线函数，而距离测度法主要有Euclidean距离、Minkowsky距离、Hausdorff距离、Fréchet距离等[7]，其中，Hausdorff距离和Fréchet距离主要用来计算两个点集间的相似性，但Hausdorff距离忽略了点集的时间序列问题，基于供热系统时序数据的特点，本文采用Fréchet距离作为曲线趋势变化属性相似度的度量，其优点在于充分考虑了曲线的连续性，非常适用于曲线间的相似性比较[8]。

Fréchet 距离由 M. Fréchet提出，描述了两质点分别沿着2条给定曲线以任意速度单向运动时，二者之间的最短距离。Axel Mosig和Michael Clausen曾将Fréchet 距离与变换群的交叉子集结合，应用到判别两条曲线的相似性上[9]，曹凯等引入Fréchet距离进行云规则推理，设计了一种智能地图匹配算法[10]。Eiter 和 Mannila在连续 Fréchet 距离的基础上提出了离散 Fréchet 距离[11]的定义 , 而朱洁等考虑了离散Fréchet距离的关键特征峰值点，减少了算法的复杂度并将其运用到了手写签名验证上[12]，收到了一定的效果。

离散Fréchet距离定义如下：

（1） 给定1个有n个至高点的多边形链P=〈P\_1,P\_2,P\_3,〖…,P〗\_n,〉，1个沿着P的k步，分割P的峰值点成为k个不相交的非空子集{P\_i }\_(i=1,…,k)，使得P\_i=〈P\_(n\_(i-1)+1),…,P\_(n\_i ) 〉和1=n\_0<n\_1<⋯<n\_k=n

（2） 给定2个多边形链A=〈a\_1,…,a\_m 〉，B=〈b\_1,…b\_n 〉，1个沿着A和B的组合步是1个沿着A的k步{A\_i }\_(i=1,…,k)和1个沿着B的k步{B\_i }\_(i=1,…,k)组成，使得对于1≤i≤k，A\_i,B\_i中有1个恰好包含1个至高点。

（3） 1个沿着链A和B的组合步W={(A\_i,B\_i )}的花费(cost)为：d\_F^W (A,B)=max┬i⁡max┬((a,b)∈A\_i×B\_i )⁡〖dist(a,b)〗

其中dist()为a,b间的欧氏距离，则链A和B间的离散Fréchet距离为：

F(A,B)=min┬W⁡〖d\_F^W (A,B)〗 (2)

2.2 出水温度曲线与目标温度曲线的平移问题

出水温度曲线与目标温度曲线的平移表现为两者横纵坐标的差异，在温度曲线中，横坐标的度量为时间，表现为出水温度调整的提前或延后问题；而纵坐标的度量为温度，体现在两条曲线的温差问题，具体为锅炉系统是否能按需达到预定温度以及在不需要较多热量时降低负荷节约能源。

2.2.1时间差异

时间差异定义为两条曲线的n个同一维度上特征点间的时间差均值。而特征点的确定会对度量结果产生较大影响，考虑到供热锅炉数据的特点，在同一维度上很难找到成对的特征点[13]。这里先获得两条曲线的最值，然后按照最值将每条曲线单独划分为多个单调区间，对于单调增区间最小值为初始特征点对，否则最大值为初始特征点对，然后在各区间根据初始特征点对的类型计算下一个特征点对，取其时间的差值作为时间差异。依次计算出每个时间段的差异集合，最后将这些差异的均值作为差异度量。

D\_X=(∑\_(j=1)^k▒∑\_(i=1)^m▒(X\_(〖max〗\_ji)^A-X\_(〖max〗\_ji)^B ) )/n^' (3)

式中n^'为特征点的个数，m为每个单调区间的点数，A为目标温度，B为出水温度，则X\_(〖max〗\_ji)^A为目标温度曲线在第j个区间温度为第i大值的时间，供热控制较差时可在合理时间段取值。

2.2.2 温度差异

温度差异简称温差，主要表现在供热温度较低时是否能满足采暖需求，反之是否发生能源浪费。这里将分两方面考虑，曲线的最大值差和最小值差。

两条曲线最大值处的温差，能够判断供热效果，出水温度是否能按需达到采暖需求，为供热用户提供足够的热量。目标温度较高时通常意味着用户在家或者气象条件不能提供较多自然热量，所以需要锅炉系统高负荷运转提供充足热量，也可以一定程度上减少客诉。

两条曲线最小值处的温差，能够判断供热锅炉系统是否节约能源。目标温度较低时通常意味着，用户家中无人，或者室外温度等气象条件能提供较多自然热量，故锅炉系统需要降低运行负荷减少热量，节约能源降低运行成本。

综合上述两个方面将两条曲线的温差定义如下：

D\_Y=(Y\_min^A-Y\_min^B+Y\_max^A-Y\_max^B)/2 (4)

A为目标温度，B为出水温度， Y\_min^A为目标温度曲线最小值点的温度。

2.3 出水温度曲线与目标温度曲线的伸缩问题

出水温度曲线与目标温度曲线的伸缩问题，在其横坐标上表现为锅炉系统在时间上，是否按统一节奏对锅炉系统进行调控，对天气情况的变化是否做出时间一致的操作；在其纵坐标上表现为锅炉系统对温度控制的灵敏度，在高低温转换时可以及时达到预期温度。

借助离差标准化的思想，最值差可以完整的表现整体的数据跨度[14]，将横纵坐标的最值差比作为两条曲线的伸缩比，能较好的反映数据整体的特点，对其横向和纵向伸缩比的计算方式如式（5）和（6）所示.

E\_X=1-(X\_max^B-X\_min^B)/(X\_max^A-X\_min^A ) (5)

E\_Y=1-(Y\_max^B-Y\_min^B)/(Y\_max^A-Y\_min^A ) (6)

式中X\_max^A为目标温度曲线最大值的横坐标， Y\_min^B为出水温度曲线最小值的纵坐标。

2.4 评价结果

将三种属性相似度共五个度量加权融合为出水温度曲线和目标温度曲线的一致性度量：

sim(A,B)=(ω\_1 F)/ε\_F +(ω\_2 |D\_X |)/ε\_(D\_X ) +(ω\_3 |D\_Y |)/ε\_(D\_Y ) +(ω\_4 |E\_X |)/ε\_(E\_X ) +(ω\_5 |E\_Y |)/ε\_(E\_Y ) (7)

式中ω\_1、ω\_2、ω\_3、ω\_4、ω\_5分别为趋势、横向平移、纵向平移、横向伸缩、纵向伸缩属性相似度的权值，ω\_1+ω\_2+ω\_3+ω\_4+ω\_5=1，可通过数据统计及最小二乘法得出[15]。ε\_F 、ε\_(D\_X ) 、ε\_(D\_Y ) 、ε\_(E\_X ) 、ε\_(E\_Y ) 分别为五个属性相似度的阈值。

本文通过分析出水温度曲线与目标温度曲线间的一致性，分别给出曲线的趋势变化、平移、伸缩三种属性相似度的计算方式，并将其加权融合为一个评价结果，用来对锅炉供热过程进行评价。

# 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