View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384　　分类号_____密级_____
学号：24320121152288　　UDC _____

厦门大学

硕 士 学 位 论 文

面向大数据的聚类方法及其应用研究

Research of Clustering Method and Its Application for Big Data

汪宜东

指 导 教 师： 吴 清 强 副 教 授
专 业 名 称： 软 件 工 程
论文提交日期： 2 0 1 5 年 4 月
论文答辩日期： 2 0 1 5 年 5 月
学位授予日期： 2 0 1 5 年 月

指 导 教 师：_____
答辩委员会主席：_____

2015 年 4 月

厦门大学博硕士论文摘要库

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学博硕士论文摘要库

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

厦门大学博硕士论文摘要库

摘要

近些年来，随着计算机科学与技术的快速发展，在很多行业中产生了越来越多的海量数据信息。聚类作为数据挖掘的一个非常受关注的分支学科，在这种情况下得到了长足的发展，一系列经典的聚类算法被研究者提出，但目前能应用于大数据聚类的算法不多，Apache Mahout 推出的聚类算法只有 5 种，其中有 4 种基于 Kmeans 算法开发的，Spark 官方推出的聚类算法目前只有 Kmeans。一些效果较好的聚类算法，它们的时间复杂度比较高，开发出适应大数据聚类的难度较大。传统的 Kmeans 可以用于大数据聚类，但其迭代过程涉及到多次的 HDFS 文件系统的读写操作也非常费时。

本文通过引入聚类特征树，获得微簇中心点集，利用 Maximin 算法选取初始聚类中心点集，提出了基于层次和划分的 BM2Kmeans 算法，同时利用微簇中心进行微簇融合，提出了基于层次和密度的 BMCMCluster 算法。前一种算法能够实现快速搜索到较好且稳定的初始聚类中心点集，从而实现高效的大数据聚类，但需要指定聚类类别数；后一种算法也能够实现快速、高效的大数据聚类，且聚类类别数不需指定，算法会通过微簇融合的方式形成大的聚簇，能够发现任意形状聚簇。

本文研究以油气勘探领域的数据为实验数据，对基于 Hadoop 平台实现的两种针对本文提出的聚类算法通过实验验证，并对实验结果进行分析，通过可视化的方式将这两种聚类算法的聚类结果表达出来。通过实验，可以看出基于层次和划分的集成聚类算法 BM2Kmeans 的聚类效果要优于传统的 Kmeans 大数据聚类算法，并对基于层次和密度的集成聚类算法 BMCMCluster 的聚类结果进行可视化展示。

关键词：大数据；集成聚类；聚类特征树

厦门大学博硕士论文摘要库

Abstract

In recent years, more and more huge amounts of data information is produced in many industries in the condition of the fast progress of computer science and technology. Cluster analysis technology is an important part of Data Mining, and of course, the development of cluster analysis is also fast and mature in this condition. a series of classical cluster algorithm is proposed, but cluster algorithms which are suitable for big data are scarce, Apache Mahout only implements five kinds of clustering algorithm, and there are four kinds of these algorithms based on Kmeans, Spark officially launched just Kmeans, some cluster algorithm have a good effect, but these time complexity is very high, which usually is difficult to implement to adapt to big data cluster. Kmeans cluster algorithm can be used for large data, but its iterative process is so long because of the I/O operation on HDFS.

This paper proposes BM2Kmeans algorithm based on hierarchy and classification, which gets micro cluster center set by cluster feature tree, and gets the initial cluster center set by the Maximin algorithm and micro cluster center set. At the same time, it also proposes BMCMCluster algorithm based on hierarchy and density, which uses micro cluster centers to fusioning micro cluster. The first algorithm can quickly search a fine and stable initial centers to implement efficient big data cluster, but you need to specify the number of class; the last algorithm can also realize a quick and efficient big data cluster, and the number of class does not need to specify, big cluster is formed by means of micro cluster fusion, this algorithm can discover arbitrary shape cluster.

In this paper, we use the data in the fields of oil and gas exploration as our experimental data, we realize the two integrated cluster algorithm based on Hadoop platform. And we verify the two integrated cluster algorithm by clustering measures and the clustering result of visual display. The experiments show that the integrated cluster algorithm based on hierarchy and division method which called BM2Kmeans

厦门大学博硕士论文摘要库

is superior to the traditional Kmeans algorithm on effect. We also give the clustering result of visual display to the integrated cluster algorithm based on hierarchy and density method which called BMCMCluster.

Keywords: Big Data; Integrated Cluster; Cluster Feature Tree

厦门大学博硕士论文摘要库

目录

第一章 绪论 1

1.1 研究背景及意义 2

1.2 研究现状 6

1.3 本文主要研究内容 8

1.3.1 主要研究内容 8

1.3.2 论文的特色 9

1.4 论文的结构安排 9

第二章 论文相关理论 11

2.1 Kmeans 的并行化算法 11

2.1.1 Kmeans 算法的基本原理 11

2.1.2 MapReduce 编程模型 11

2.1.3 Kmeans 算法的并行化思路 13

2.2 Kmeans 并行化的改进 15

2.2.1 基于划分的改进策略 15

2.2.2 基于层次的改进策略 19

2.2.3 基于密度的改进策略 20

2.2.4 基于网格的改进策略 21

2.2.5 基于模型的改进策略 21

2.3 本章小结 22

第三章 基于 MapReduce 的 BM2Kmeans 算法模型 24

3.1 BM2Kmeans 算法原理 24

3.1.1 微簇的生成 25

3.1.2 初始聚类中心的选择 30

3.1.3 算法并行化思路 31

3.1.4 算法分析 34

3.2 BM2Kmeans 聚类算法步骤 35

厦门大学博硕士论文摘要库

3.3 BM2Kmeans 聚类算法流程 36
3.4 本章小结 37
第四章 基于 MapReduce 的 BMCMCluster 算法模型 38
4.1 BMCMCluster 算法原理 38
4.1.1 微簇融合思想 38
4.1.2 算法并行化思路 41
4.1.3 算法分析 41
4.2 BMCMCluster 算法步骤 42
4.3 BMCMCluster 算法流程 42
4.4 本章小结 44
第五章 实验设计与分析 45
5.1 实验设计思路介绍 45
5.2 实验数据集介绍 45
5.3 结果及分析 46
5.3.1 Kmeans 大数据聚类结果 47
5.3.2 Canopy-Kmeans 大数据聚类结果 53
5.3.3 BM2Kmeans 算法结果 56
5.3.4 BMCMCluster 算法结果 59
5.3.5 结果分析 64
5.4 本章小结 64
第六章 总结与展望 65
6.1 总结 65
6.2 展望 65
参考文献 67
致谢 72

厦门大学博硕士论文摘要库

Contents

Chapter 1 Introduction 1

1.1 Background and Significance of Research 2

1.2 Research Status 6

1.3 Main Research Content 8

1.3.1 Research Content 8

1.3.2 Innovation Points 9

1.4 Outline of the Dissertation 9

Chapter 2 Related Theories 11

2.1 Parallelization of Kmeans Algorithm 11

2.1.1 Basic Principle of Kmeans Algorithm 11

2.1.2 Programming Model of MapReduce 11

2.1.3 Parallel Idea of Kmeans Algorithm 13

2.2 Improvment of Kmeans Parallelization 15

2.2.1 Improvement Based on Classification 15

2.2.2 Improvement Based on Hierarchy 19

2.2.3 Improvement Based on Density 20

2.2.4 Improvement Based on Grid 21

2.2.5 Improvement Based on Model 21

2.3 Summary 22

Chapter 3 BM2Kmeans Algorithm Model Based on MapReduce 24

3.1 Principle of BM2Kmeans Algorithm 24

3.1.1 Formation of Micro Cluster 25

3.1.2 Selection of Initial Centers 30

3.1.3 Parallel Idea 31

3.1.4 Analysis of Algorithm 34

3.2 Procedure of BM2Kmeans Algorithm 35

厦门大学博硕士论文摘要库

3.3 Framework of BM2Kmeans Algorithm 36
3.4 Summary 37
Chapter 4 BMCMCluster Algorithm Model Based on MapReduce 38
4.1 Principle of BMCMCluster Algorithm 38
4.1.1 Idea of Micro Clusters Fusion 38
4.1.2 Parallel Idea 41
4.1.3 Analysis of Algorithm 41
4.2 Procedure BMCMCluster Algorithm 42
4.3 Framework of BMCMCluster Algorithm 42
4.4 Summary 44
Chapter 5 Design and Analysis of Experiment 45
5.1 Introduction of Experiment Design 45
5.2 Introduction of Experiment Data Set 45
5.3 Result and Analysis 46
5.3.1 Results of Big Data Clustering Algorithm of Kmeans 47
5.3.2 Results of Big Data Clustering Algorithm of Canopy-Kmeans 53
5.3.3 Results of BM2Kmeans Algorithm 56
5.3.4 Results of BMCMCluster Algorithm 59
5.3.5 Analysis of Result 64
5.4 Summary 64
Chapter 6 Conclusions and Future Work 65
6.1 Conclusions 65
6.2 Future Work 65
References 67
Acknowledgements 72

厦门大学博硕士论文摘要库

第一章 绪论

人类正面临着信息爆炸的时代，现在的信息技术发展日新月异，根据权威机构统计，从 40 年前开始，全世界每隔二十个月其产生的信息总量就会翻一倍，而且进入二十一世纪，随着网络及存储技术的发展，信息量的增长将会更快，面对如此海量的历史数据信息，要求我们必须能够处理它们，毕竟其中蕴含着知识财富，这些海量的历史信息数据要能保存一段时间，这些数据中蕴涵着潜在的有价值的知识信息，这就急切需要我们把有用的信息和知识从这些海量的数据中提炼出，从而发现一些有价值的知识和信息及数据中的隐含模式，在此背景下，数据挖掘技术应运而生，数据挖掘就是从海量的、有数据噪声的、不清晰的、有缺损的、随机产生的数据集中提炼蕴含其中的、尚未被发现的，同时极具潜在有用的信息和知识的过程。数据挖掘被称做数据库中的知识发现更合适一些，数据挖掘差不多就涵盖了知识发现的整个处理过程，知识发现囊括以下几个处理模块：数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评估、知识表示。简单的来说，从原始数据提炼出有价值的信息的过程就是一个知识发现的完整过程。下面是知识发现的一般过程示意图，如图 1.1 所示：

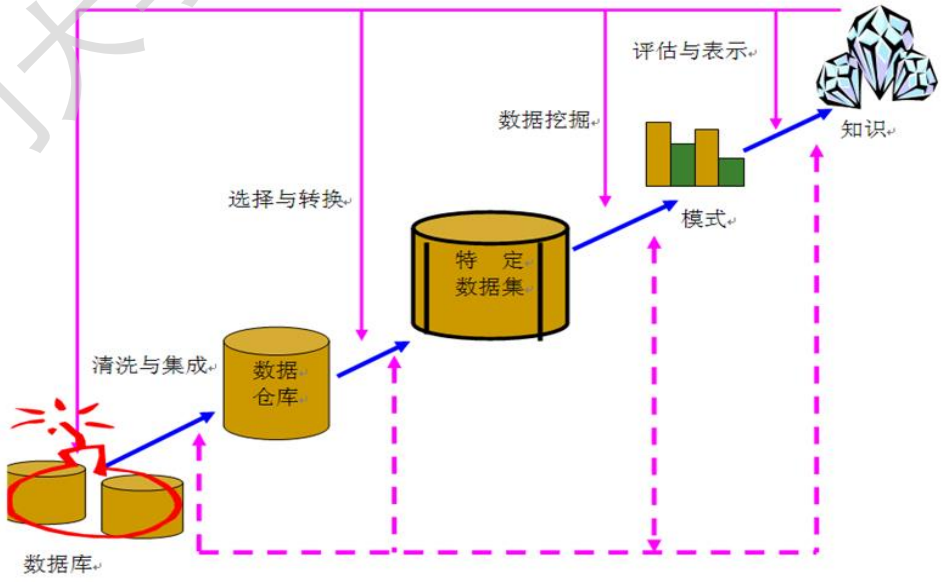

图 1.1 知识发现一般过程

数据挖掘，也是人们对数据进行进一步处理和分析的技术。数据挖掘会指定其要完成的任务，即找到需要的模式类型，通常数据挖掘任务可以分为两种类型：描述、预测。描述性的数据挖掘任务是对数据库中数据的一般性质或属性进行描述，而对于预测性的数据挖掘任务而言，就是要进行挖掘处理的数据应用特定挖掘算法来进行推断，给出预测结果。数据挖掘的功能可以用图 1.2 来进行表示，而聚类分析（Cluster Analysis）作为数据挖掘中一个非常重要的分支，也非常受人们的关注，聚类分析诞生于很多研究领域的融合，包括机器学习、概率论与数理统计、生物信息学等。

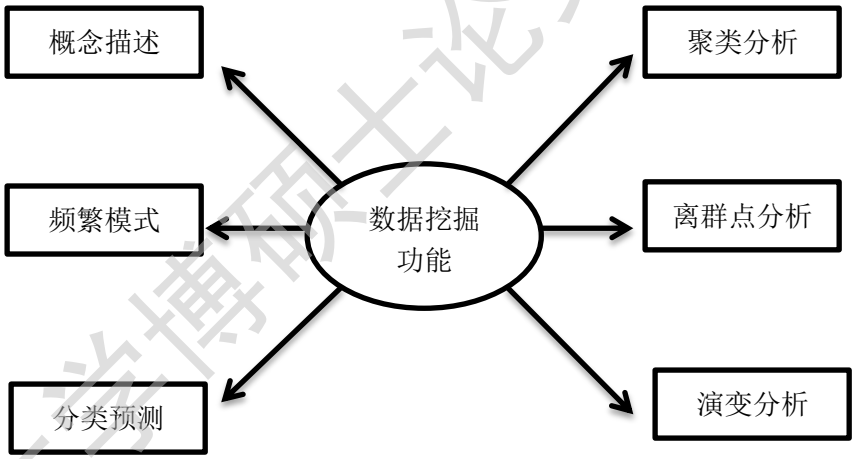

图 1.2 数据挖掘的功能

厦门大学博硕士论文摘要库

1.1 研究背景及意义

信息技术经过七十多年的发展，已经在人们日常生活、社会、国家、世界的各个方面得到广泛应用。政治、经济、社交活动中的很大一部分都与数据的产生、记录、传输和使用紧密相关，而且随着网络应用日益发达，大数据应用的影响范围逐渐扩展。据权威机构测算，全球数据总量的增速相当惊人，每隔两年会翻一

倍，从另一角度来说，近两年新产生的数据总量相当于人类有史以来所有数据量的总和。在这个大背景下，针对大数据的挖掘工作显得尤为迫切和重要。

大数据虽然源于信息技术，但其影响已经远远超出信息行业。数据已经成为一个企业的资产，不再是无用的数字串。如新兴的互联网公司，利用新技术，大规模收集数据，对消费用户进行行为预测，将积累的大数据转化为实实在在的利润。从传统行业来看，以前被认为无用的消费数据，现在已经成为克敌制胜的法宝。新的时代，要掌握先机，就必须对数据资产重新进行优化配置。

大数据的诞生有其必然性。如海上运输，开始基础设施差些，只能运输少量货物，当运输设备改善了，其运输的货物量不断增加。信息产业的发展也是如此，宽带网络建设就好比海上运输设备，而大数据则是运输设备所运输的“货物”。

从信息技术的不断进步过程，可以看出信息技术具有三个核心和基础的能力：信息存储、信息处理、信息传递。几十年来，信息技术就是围绕这三个能力在飞速发展，在此过程中，信息的处理和存储能力获得极大的提升。存储的价格从上世纪 60 年代 1 万美元 1M，已经降到现在的 1 美分 1G 的水平，其价差高达亿倍，如图 1.3 和图 1.4 所示，在几年前在线实时高清电影还遥不可及，现在已经变得很平常了。由此可见，大规模存储技术和网络带宽的飞速发展，为大数据时代提供了适合的土壤。

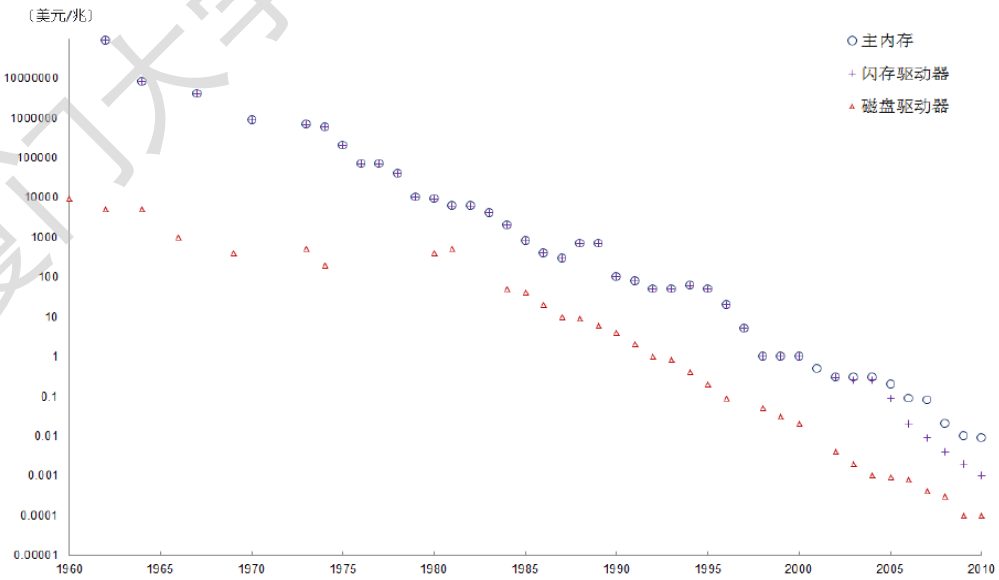

图 1.3 存储价格的下降

厦门大学博硕士论文摘要库

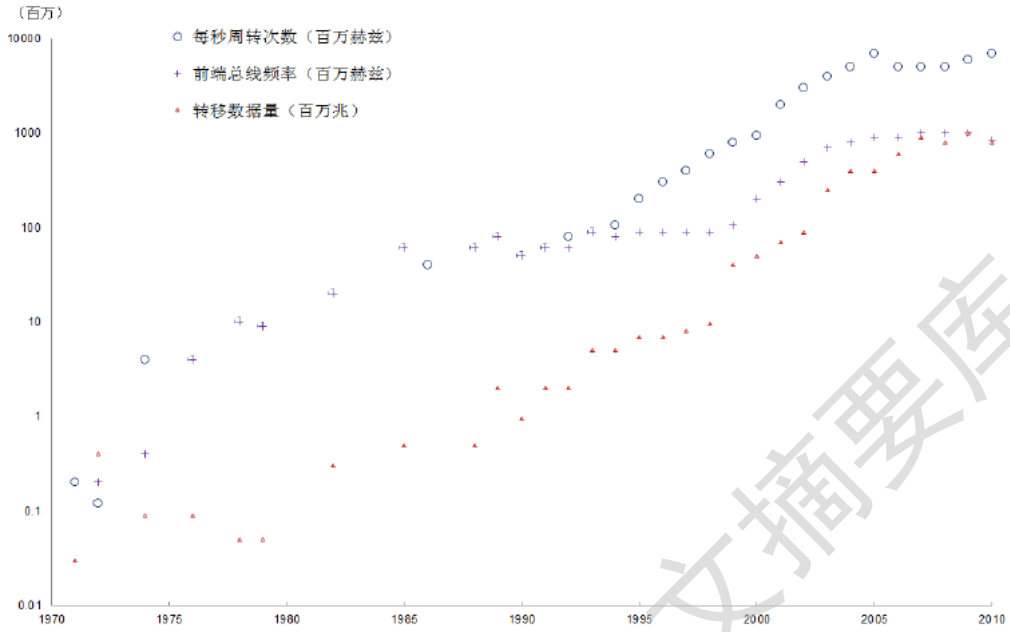

图 1.4 网络带宽的增加

互联网的出现，把每个人桌面上的计算机连接起来，改变了人们的生活，成为人们获取各类信息的首要渠道。互联网已经成为更接近消费者、最理解消费者的平台。互联网可以对用户的行为数据进行忠实的记录，随着互联网应用的急剧扩增，为大数据的产生提供了条件。云计算的出现再次改变了数据的存储和访问方式。在这之前，数据大多存储在服务器和 PC 上。云计算把所有的数据集中存储到“云端”，用户可以通过浏览器或应用程序来直接访问。同时这些云服务会积累大量的数据，实际上这些大数据已经成为企业的核心资产，近年来，国内外兴起了建设云计算基地的热潮，也为大数据的诞生提供了必备的物质条件。同时随着物联网技术的发展，传感器无时无刻不在产生大量的数据，当数据被持续的收集，就会成为大数据的来源之一。社交网络也是互联网发展史上的一个重要里程碑，它完美地将现实生活的中人际关系映射到网络空间，并借助互联网的特性而得到进一步发展。人们可以在虚拟的网络空间中分享各自相关的心情和事件，并相互传染和传播。在社交网络中，有一个重要的研究内容就是如何利用网友间的关系链数据来为研究消费者行为提供预测。深入了解社交网络，就会明白大型社交网络平台已经构成了以“个人”为枢纽的不同方面的数据集合。社交网络把网友在不同网络留下的“足迹”链接起来，形成完整的行为轨迹和“偏好”链。

厦门大学博硕士论文摘要库

图 1.5 反映社交网络 Facebook 上人们活跃程度的世界地图

图 1.5 为 Facebook 上人们相互联系的数据通过建模、渲染得到的一幅图片，越是明亮的地方，人们相互交流越是活跃。此外智能终端的普及，如智能手机、iPad 等先进移动设备也为大数据带来了丰富的数据。

国际数据公司通过四个特征来对大数据进行定义：海量的数据规模（Volume）、快速的数据流转与动态的数据体系（Velocity）、多样的数据类型（Variety）、巨大的数据价值（Value）。麦肯锡公司在《大数据：创新、竞争和生产力的下一个前沿领域》中对于大数据的定义如下：大数据是指数据规模已经超过一般的数据库能够获得、储存、管理、计算分析能力极限的数据集。从另一个角度来说，并不是说大数据集的量越大，就是一定是大数据，这取决于实际情况。在亚马逊从事大数据研究的科学家 John Rauser 对于大数据的定义如下：大数据是超越一台计算机计算能力极限的数据量。大数据是一个很泛的概念，我们通常面对的“大数据”，就是指在你所拥有的单机环境中，难以处理的数据集。数据的获得、修正、储存、处理、展示、应用、共享等都可看成大数据产业的生成活动，其业务模式包括网络数据和信息服务、企业和政府职能决策、企业流程改造和变革等，应用领域更是涉及广泛，包括智慧城市、金融、信息服务、科学研究、制造业等，几乎涵盖国民经济的所有部门。伴随着当今全世界数据的高速增长，数据已经成

厦门大学博硕士论文摘要库

为生产因素中不可或缺的重要因素，大数据更是成为企业发展和竞争力的关键，对大数据的利用将支撑新一波生产力的增长和消费浪潮。

数据已经与自然资源、人力资源一样成为我们重要的战略资源，掌控数据资源的能力是国家数字主权的体现。同时大数据的研究及应用已经成为推动现有产业升级与新产业崛起的重要力量。既然大数据有如此重要的作用，所以对于大数据的分析利用就显得尤为重要，比起数据量大更难以应付的是数据的多样性、实时性、不确定性。我们更加关注从海量数据中获得价值的能力，本文将针对大数据分析方法中聚类方法进行研究。本文将以油气勘探领域的大数据为研究背景，下面简要介绍下油气勘探领域的一些基础知识。

地震数据中通常包含着大量的地质信息[1-3]，而地震属性通常能更好地从不同角度将这些信息表达出来。目前实际生产中已经利用地震属性进行储层特征描述和储层预测，并且取得了不错的效果，同时也促进了含油气储层预测技术的发展，正是因为这些成果，也促使得针对地震属性的研究越来越热门，到目前为止，可提取的地震属性[4-6]已多达上百种。正是由于现在地震属性越来越多，通过人工的方式很难断定何种属性对储层更敏感，更贴近于反映储层的分布情况。实际上作为研究对象的地质目标与地震属性与也并非是一对一对应的关系，很多地震属性是对岩性与油气、地质构造、地层信息等因素的总的展现，因此这就需要我们将大量的地震属性综合起来使用，而关于地震多属性的聚类分析已经成为一个很有前景和重要的研究方向。本文正是针对油气勘探领域的大数据聚类需求，研究能够有效处理大数据聚类的算法。

1.2 研究现状

近几十年来，聚类算法得到了很大的发展，目前已知的方法有很多种，常用的主要聚类方法分为以下五类：划分聚类方法、层次聚类方法、密度聚类方法、网格聚类方法、模型聚类方法。大数据聚类作为一个关注度越来越高的研究领域，已经有不少的关于大数据的聚类算法被相继提出来。

Kmeans 算法是一种比较经典的基于划分的聚类方法，通过迭代它能够快速对数据集进行有效聚类，目前针对它的主要改进集中于初始聚簇中心点集的选择[7-21]。现在针对大数据，分布式 Kmeans 算法[16, 17, 22]也被设计出来了，但由于

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”.

Fulltexts are available in the following ways:

1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.

2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库