学校代号  学 号

分 类 号 密 级

****

硕士学位论文

基于改进蝴蝶优化算法聚类的协同过滤推荐系统的研究与实现

学位申请人姓名 何雄

培 养 单 位 信息科学与工程学院

导师姓名及职称 戴牧红教授

学 科 专 业 软件工程

研 究 方 向 推荐系统

论文提交日期 2022年 2月10日

**The Research on Some Chemical and Physical Test Methods in the Chemical and Fiber Production Process**

**by**

**LIU Changjiang**

## B.E.(Hunan University)1995

**A thesis submitted in partial satisfaction of the**

**requirements for the degree of**

**Master of science**

**in**

**Analytical chemistry**

**in the**

**Graduate school**

**of**

**Hunan University**

**Supervisor**

**Professor Wang Xiong**

**October,2001**

湖　南　大　学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。 除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的 成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完 全意识到本声明的法律后果由本人承担。

作者签名： 日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有 关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南大学可以 将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复 制手段保存和汇编本学位论文。

本学位论文属于

１．保密□ ，在 年解密后适用本授权书。

2．不保密□。

(请在以上相应方框内打”√”)

作者签名： 日期： 年 月 日

导师签名： 日期： 年 月 日

Ⅰ

**摘 要**

石油化工行业的科学研究和产业发展对相关现代分析提出了新的要求，… …

（略）

关键词：酯；过氧化物；气相色谱；纤度；

Ⅱ

**Abstract**

The development of the scientific research and production in…………(略)

[B.E.(Hunan University)1995 3](#_Toc90403639)

**Key Words：** Ester; Peroxide; Gas Chromatography; Fiber number

Ⅲ

**目 录**

学位论文原创性声明和学位论文版权使用授权书………Ⅰ

摘要………………………………………………………Ⅱ

**Abstract**…………………………………………………Ⅲ

插图索引……………………………………………………Ⅶ

附表索引……………………………………………………Ⅷ

**第1章 绪论**…………………………………………………1

1.1 环已烷高压氧化液中酯类、过氧化物的气相色谱分 析研究………………………………………………2

1.1.1 气相色谱法最新进展…………………………2

1.1.2 气相色谱应用…………………………………7

**结论**…………………………………………………………61

**参考文献**……………………………………………………62

**附录A**（攻读学位期间所发表的学术论文目录）………63

致谢…………………………………………………………64

Ⅵ

**插图索引**

图1.1 气相色谱装置框图…………………………………2

图1.2 色谱流出曲线图(色谱图)………………………… 3

图2.1 Carbowax-20M柱110℃柱温环已烷氧化液的色谱图 … … … … … … … … … … … … … … … … … … … 21

Ⅶ

**附表索引**

表2.1 柱温与样品沸点的关系…………………………… 15

表2.2 9.26氧化液及其还原产物的色谱数据…………… 32

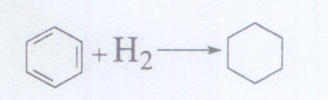
Ⅷ

**第1章　绪 论**

石化行业在我国国民经济中占有重要的地位，在质量检验和生产过程控制中，需要用到大量的化学和物理分析检验方法，这些分析检验方法的可靠性和检测分析速度对加强企业技术水平有着很重要的价值。发展与建立新的方法对促进企业的发展、丰富石化行业的理化分析方法和推动相关领域的理论和实践有着重要的意义。

鹰山石化厂化工化纤生产主要包括以下生产过程[3,4]：

苯加氢



（1.1）

**1.1 研究背景和意义**

中国互联网信息中心（CNNIC）发布的<<第46次中国互联网网络发展状况统计报告>> [1] 显示，截止2020年6月，我国网民规模达9.40亿，较2020年3月增长3625万，互联网普及率达67.0%，较2020年3月提升2.5个百分点。我国手机网民规模达9.32亿，较2020年3月增长3546万，网民使用手机上网的比例达99.2%，较2020年3月基本持平。

为了给这些互联网用户群体提供资讯服务，人们随处可见到各种各样的应用，根据上文的CNNIC报告显示，截止2020年6月底，我们国内市场上检测到的APP数量359万款，由此带来的网络数据的规模也日益膨胀。希捷科技赞助、国际数据公司（IDC）发布了《数据时代2025》[1]白皮书。白皮书显示，未来数据增长速度惊人，全球数据领域将从 2019 年的 45 ZB 增长到 2025 年的 175 ZB。数据的来源以及由数据飙升引起的应用趋势也会产生变化，这是数据未来的大势所趋。从数据中挖掘隐藏的巨大的价值，也必将促进社会的进步。

目前的社交、视频、新闻以及电子商务等互联网系统都是采用推荐系统给用户提供更好的体验。推荐系统的核心及充分考虑用户个性化的需求，即结合用户行为偏好、历史数据等，使用合适算法从海量数据中，将用户喜爱的内容推送给到用户，达到“千人千面”的目的。它能有效的提升用户忠诚度、粘性、增加流量，并提供满意度。同时也能够更深入了解用户，便于进行更多产品形态的开发和变现。

**1.2 国内外研究现状**

自从上个世纪90年代起，“推荐系统”被首先提出，最早是Tapestry首次提出协同过滤，解决电子邮件分类过滤问题。而其后协同过滤作为推荐系统的重要技术开始发展开来。其后，亚马逊等企业纷纷引入该算法，也就将推荐系统在多个领域开枝散叶。经过了经过了二十多年的沉淀与发展，推荐系统取得了瞩目的成果。

基于协同过滤的推荐。一般有基于用户的协同过滤和基于物品的协同过滤。不需要预先获取用户和新闻的特征数据，仅仅依赖于用户的历史行为数据对用户建模分析，从而给出推荐。具体是：根据用户－项目评分矩阵，寻找与目标用户（待推荐用户）有相似偏好的其他用户，即为目标用户的最近邻居，根据最近邻居进行评分预测，形成推荐。

但是在整个用户空间搜索目标用户的最近邻居非常耗时，并且存在选择邻居不合理的情况，进而导致推荐质量和推荐效率不高。于是基于聚类的协同过滤改进算法被提出。聚类技术的引入，使得在为了寻找近邻计算相似度的过程中，不用比较全局用户而只需比较同一个簇内的用户，减少了寻找相似用户时间。且聚类得到的同一个簇中对象具有较高的相似度。可以降低算法在计算相似度时的复杂性，并且增加系统的可扩展性。近些年许多学者对基于聚类的协同过滤算法提出了改进方案，Wang等人［3］设计了启发式聚类模型，根据数据集的特性为目标用户划分邻居，提高了推荐效率，仿真实验证明了这一算法的可行性。Li等人［4］提出了一种类别偏好Canopy-K-means聚类的协同过滤推荐算法，设计类别偏好比率，改善了推荐效率。

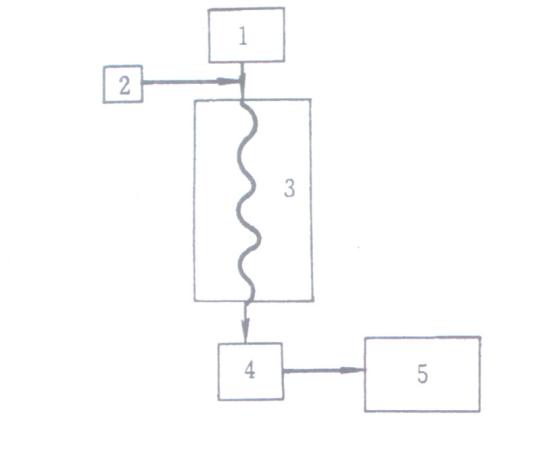


图1.1 气相色谱装置框图

1—流动相；2—样品；3—色谱柱与固定相；4—检测器；5—数据记录及处理器

**1.3 论文主要内容和组织结构**

论文分为以下五个章节：

第一章为绪论，主要介绍群智能算法、聚类算法和推荐系统的研究背景和意义，以及国内外的研究成果。

第二章主要介绍聚类算法中的 K-means 聚类和群智能算法中的粒子群优化的相关理论基础。

第三章主要介绍对粒子群算法的改进和加入改进粒子群的 K-means 聚类算法 IPK-means。

第四章主要介绍协同过滤推荐系统相关概念及其常见的协同过滤推荐算法，

协同过滤推荐算法的改进策略以及 IPK-means 算法在协同过滤推荐系统中的应用效果。

第五章主要是对全文研究工作的总结以及未来工作的展望。

**4**

**1.1.1.1 理论方面**

从色谱流出曲线（见图1.2）可知，为了获得最挂分离结果，应满足的条件是[11]

式为[13]



（1.11）.

式（1.11）还表明，在除柱温外其他操作参数都不变的情况下，给定组分的、或I等保留值与柱温有倒数的线性关系。另外，在操作参数不变的情况下，

**1.5 本文构想**

，提出工作。

**第2章　相关理论基础**

**2.1 聚类算法**

**2.1.1 聚类分析概念**

聚类分析作为数据挖掘领域中一个重要的研究方向，被广泛应用于统计学、生物学、医学、机器学习和模式识别等领域[6,27]。聚类分析的主要核心思想是物以类聚，即根据样本特征属性将类似的样本划分到相同的类群中，而差别相对比较大的样本则是分配到不同的类群中。算法是通过每个样本自身的特点属性实现这一划分的，这种划分过程的无监督的，因此我们不需要事先了解每个数据对象分别属于哪个类别。聚类可以根据样本是否只属于一个簇类，主要有硬聚类和模糊聚类。其中在数据挖掘领域中使用最为广泛的聚类方法是硬聚类，因此本文也主要针对硬聚类中的相关算法进行研究。

**2.1.2 聚类算法的要求**

受目前正处于蓬勃发展期的聚类分析技术的影响，相关人员也因此不断地研究和改进聚类分析技术中涉及到的各种聚类算法。然而，随着该技术的应用范围不断扩大，作为聚类分析核心的聚类算法也需要不断的完善以应对应用范围所带来的日益复杂的数据集。

（1）可扩展性和高维性

在当今数据呈爆炸式增长的大数据时代，算法所面对的数据量或许能达到成千上万甚至上百万。而大多数聚类算法表现出来的非常好的聚类效果可能是在那些规模相对较小的数据集上，这并不意味着它们能够在规模达到成千上万的数据集上得到同样高效的运行效果和效率。因此仅仅根据在较少的低维数据集上表现出的非常好的聚类结果，并不能评判一个聚类算法的好坏，一个良好的聚类算法同时也需要在大量的高维数据集上呈现出较好的性能。

（2）降低人工设定参数带来的影响

很多聚类算法在执行过程中都设定一些变量参数的值，这些参数的值一般都是由我们人为设定，比如 K-means 聚类中的 k 值等。然而一个稳定且良好的聚类算法在执行过程中应尽可能地降低人工设定参数的个数。首先，由于人工设定的参数值具有很大的随机性，因此用户只要通过不断尝试不同的参数值才能确定一个最佳的参数，这不仅增加了用户的使用难度，更提高了对使用者技能水平的要求；其次，人工设定的参数在算法的执行过程中很难发生改变，因此固定不变的参数过多只能保证算法在某个数据集上具有较好的效果，但是并不能说明它同样适用于其它更多的数据集。

（3）适用于任何类型数据集

由于过去数据类型的单一性，目前的大多数聚类算法对于非数值型元素形式的处理效果不太理想，但随着科技时代的不断发展，数据库中存储的数据形式也变得更加多样化，图片、文本等一些非数值型数据也会出现在我们需要处理的数据集中，这要求算法在处理这类数据集时也要表现出良好的性能。

**2.1.3 聚类算法**

目前为止相关研究人员提出发表的聚类算法各式各样，但是根据这些常见算法的聚类原理，我们可以把它们分为三类：划分聚类、密度聚类和层次聚类。下面是对这3类聚类方法的介绍并分析每个类型的典型算法。

（一）基于划分的聚类方法

基于划分的聚类方法是目前应用相对广泛的一种聚类方法，其基本原理是利用事先给定的方式将数据集中的数据对象划分到对应的簇类中。在得到聚簇的数目和初始化聚类中心点之后，基于划分的聚类方法会根据某个分配原则将每个数据对象依次划分到合适的簇类，然后不断重复执行更新中心点和分配数据集操作，直至算法满足结束条件。这类聚类方法的典型代表有 K-means 聚类等。

K-means 算法是一种经典的基于划分的聚类算法[28]。算法在开始时需要事先设定初始聚类中心点和聚类过程中形成的聚簇数目，然后根据距离原则将数据集中的每个数据划分到距离最近的簇类中，随后使用目标函数对聚类结果进行评价，如果不满足终止条件，则根据结果计算聚类中心重新聚类直至满足条件。算法的具体步骤如下：

（1）人为设定即将聚类的类别数目 k 的值；

（2）在原始数据集中以一定策略选择 k 个数据点作为当前每个类别的初始中心点；

（3）根据最小距离原则将数据集中的数据划分到距离最小的簇内；

（4）根据每个类中元素的属性计算平均值，作为该类的聚类中心点；

（5）根据目标函数评价当前聚类指标；

（6）如果满足终止条件，则算法结束；否则返回步骤（3）。

算法通常选用欧式距离计算数据和中心的差异，直接使用聚类评价指标函数作为目标函数，下节将介绍常见的聚类评价指标函数。

K-means 聚类算法的主要优点是原理易于理解和思想易于实现，而且算法具有快速的搜索能力，在包含球形元素的数据集上有着非常不错的聚类效果。但是算法在开始执行时要求我们人为输入类别数目 k 的值，另外聚类使用的数据集中的数据对象通常都是没有标签的，即我们事先并不知道每个数据分别属于哪一类，因此聚簇的数目事先是无法确定的，因此我们很难确定一个合适的 k 值；除此之外，聚类的初始中心点是算法开始时随机初始化得到的——即从原始数据集中随机的选取 k 个数据，而这些随机产生的中心点会对聚类结果和收敛速度产生很大的影响；另外数据集中的噪声点对于算法的聚类效果也有很大影响。

（二）基于密度的聚类方法

基于划分的聚类方法在判断数据对象之间的相似性时，通常采用的是计算并对比它们距离的方法，因此只能发现以聚类中心点为球心的球形簇类，不适用于其他非球形簇群。为了能够实现对各种形状数据集的聚类，研究人员提出了一种基于密度的聚类方法。该算法的实现原理是以包含的数据对象的数量对数据集划分聚类。这类聚类方法中比较有代表性的主要有 DBSCAN 算法等。

在 DBSCAN 聚类算法中，我们将邻域内包含数据点多且互通的数据点所形成的最大连通数据集合称作一个类别，根据这种处理方式我们将袁术数据集中的数据划分成多个聚类。算法执行之前要求人为设定两个参数——数据对象的邻域范围半径 r 和数据点的密度阈值 MinPts。

（1）r 邻域：以数据对象 xi为中心，r 为半径的球形空间区域，其中)，为数据 xi 的r邻域内的数据对象个数；

（2）核心对象：如果数据对象的r领域内所包含的对象个数不小于密度阈值 MinPts，那么 xi为核心对象，否则不是。

DBSCAN 算法的执行过程如下：

（1）输入需要聚类的数据集 S、邻域范围半径参数 r 和数据点的密度阈值MinPts。

（2）设定核心对象集合 P=∅，簇群数目 k=0，没有被访问过的对象集合 Q=S，聚类划分结果 C=∅。

（3）依次选择数据集 S 中的每个数据对象 xi，统计该对象的 r 邻域内包含的数据对象个数|xi|，若数据对象个数大于输入的参数阈值 MinPts，则xi为核心对象，即将 xi加入到核心对象集合 P 中，即 P=P∪{xi}。

（4）如果核心对象集合 P=∅，则算法结束，并输出聚类结果 C；否则转入（5）。

（5）从（3）中得到的核心对象集合 P 中取出一个核心对象 xi，建立一个当前聚类的核心对象集合𝑃′={xi}，并更新类别序号 k=k+1 和当前簇类中的数据对象集合= {𝑥𝑖}，同时修改没有被访问过的对象集合Q = Q − {𝑥𝑖}。

（6）如果当前聚类中包含的核心对象集合𝑃′=∅，则第 k 个聚类中的对象集合生成完毕，同时修改聚类划分结果 C 和核心对象集合P = P − ，返回步骤（4）；否则转入步骤（7）。

（7）从当前聚类包含的核心对象集合𝑃′中取出一个对象，找到该对象r的领域内所有对象集合，并将其中的核心对象加入到𝑃′，同时更新当前聚类的对象集合Q = Q − ，返回步骤（6）。

首先，DBSCAN 算法在非球形的数据集上的聚类效果不算差；其次，算法划分聚簇的标准为对象的密度，因此可以有效地降低个别噪声数据对整个算法聚类效果的影响；但是算法依然有着一些缺点，首先密度聚类算法不适合处理分布密度不均匀的数据集；其次算法运行过程中达到收敛的时间会比较长，并且随着数据集规模的增大，算法收敛所需要的时间会大幅度增加；最后阈值的设定也会影响聚类效果，但在实际的聚类过程中人为设定一个最佳的阈值是很难实现的。

（三）基于层次的聚类方法

基于层次的聚类方法实现原理是利用数据对象间的属性相似成都来划分数据集的聚类结构，聚类方法的过程就是合并与分解每层结构的过程。根据这种聚类思想，相关人员将层次聚类分为两种类型：凝聚聚类和分裂聚类。自底向上的凝聚层次聚类的原理是首先假设原始数据集中的每个数据对象都属于不同的类别，即每个数据对象自成一类，然后通过计算每个类别之间的相似度大小以判断哪些类群可以合并为一个更大的新类群，直至数据集中所有数据合并成一个类群或者算法达到某个终止条件。自上而下的分裂层次聚类的思想则与之相反，首先将整个原始数据集中的所有数据对象看作为一个类别，然后计算每个类别内的相似性大小并判断这些类别是否需要分裂为两个较小的类别，直至每个元素都是一个单独的类别或者达到终止条件。

作为一种经典的自底向上的凝聚层次聚类算法，AGNES 算法的执行过程如下：

下：

（1）输入数据集最终簇类数目 k 作为终止条件。

（2）将 S 中的每个对象初始化为一个单独的簇类，得到初始簇类集合，其中。

（3）根据簇类之间的距离计算公式找出距离最近的簇类 Ci和 Cj。

（4）将簇类 Ci和 Cj 合并为一个较大的簇类，即生成新的簇类 Cm=Ci∪Cj，同时将 Ci和 Cj 从C中删除，并将 Cm 加入到C中。 判断当前簇类数目|C|是否等于 k，若不是，则返回（3）；否则结束算法，输出聚类结果。

AGENS 聚类算法的效果与参数 k 的取值相关，通过设置不同的参数 k 值，算法可以得到不同粒度的多层次聚类结构，因此算法适用于任意形状的数据集；在聚类过程中算法合并聚类之后就无法撤销，这样尽管可以减少计算代价，但不可避免会出现合并后的簇类效果不如之前的问题；最后，随着聚类数据集规模的不断增大，聚类算法的运行时间也会大幅度增加，因此不具备较好的扩展性。

**2.1.4 聚类评价指标**

聚类评价指标是用来评价聚类效果的好坏，也是聚类算法的好坏的重要依据[29]。而聚类算法的目的是把数据对象划分为一个个簇类，同一簇内元素之间的属性相似度越高，不同簇间元素之间的相似度越小，则聚类效果越好。基于这一思想，研究人员引入了多种聚类评价指标的计算方式。另外由于目标函数也是算法

在聚类过程中的聚类效果好坏的体现，因此我们可以将目标函数设定为聚类评价

指标。在介绍聚类评价指标之前，我们需要先了解数据对象的相似性是如何计算的。

（一）数据元素的相似度计算

聚类算法大多数是根据不同数据元素的相似度来对数据集进行划分聚类，因

此相似度的计算是聚类过程中至关重要的一步，不同的相似性计算方法也会对聚

类结果产生一些影响。目前常用相似性的计算方法主要有基于距离的欧几里得距离公式，即：

dis) = ;当距离值 dis 越小，说明两个数据对象越相似。

（二）聚类评价指标

介绍了数据对象的相似性计算之后，接下来我们讨论的是如何判断一个聚类结果的好坏，也就是聚类评价指标的计算方式。一个好的聚类算法产生的簇类特点为类内相似度高，簇间相似度低[30]；一般评估聚类质量的指标分为内部评价指标和外部评价指标。外部质量评价标准只适用于有标签类别的数据集，在已知每个数据真实类别的前提下进行评价；而内部评价指标适用于任何数据集，它是基于数据集的数据属性特征评价聚类效果，因此后者的适用性相对更加广泛。

常用的内部评价指标轮廓指数（SI）、邓恩指数（DVI）和戴维森堡丁指数（DBI）。

（1）SI 主要测量每个簇自身距离和其它最近簇的距离比值：

（2-1）

SI(i) = （2-2）

其中  （2-3）

（2-4）

dis分别表示簇类和 任意两个不同簇类数据的平均距离和簇类中任意两个数据的平均距离。SI 指数越高说明同一簇类内数据对象的距离越小，不同簇类间数据对象的距离越大，聚类效果也就越好，但是该指标的计算复杂度较高。

（2）DBI 主要测量任意两个簇类的类内平均距离之和与对应质心间距的比值最大值：

（2-5）

其中 （2-6）

表示簇类的簇心。DBI 指数越小说明同一簇类内数据对象的距离越小，不同簇类间数据对象的距离越大，也就是说聚类效果越好。但是环状分布的簇类得到的该指标数值上会相对较大。

（3）DVI 主要测量任意两个簇之间的类间最短距离和任意簇的类内最大距离的比值：

（2-7）

DVI 值越大说明同一簇类内数据对象的距离越小，不同簇类间数据对象的距离越大，也就说明聚类效果越好。但是具有离散类的簇类结果计算得到的该指标数值会相对较大。

**2.2 群智能算法**

群体智能是研究人员通过观察和模拟群居生物的群体行为提出的一些用于解决传统问题和实际应用问题的智能计算或优化方法[6,31]。人类在大自然中的生物种群之间体现出的智能行为下得到了启迪，以此提出了一种新的人工智能方式，即为群体智能；也就是群体中的多个简单个体之间可以通过相互合作做出一系列更加复杂的群体行为表现。这个由很多主体组成的群体被称之为智能群体的前提是群体必须能够体现出学习性、自主性和自适应性等复杂的行为特征。遗传算法（GA）、粒子群优化（PSO）算法和蚁群（ACO）算法就是目前学术界常用的群体智能算法。

**2.2.1 遗传算法**

遗传算法（GA）是 20 世纪 60 年代由美国学者 J.H.Holland 提出的一种模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型[32]。它通过模拟大自然中生物群体优胜劣汰的演变过程不断优化问题解，并逐渐逼近最优解。在问题解集搜索过程中，算法具有较强的鲁棒性和适应性。

在介绍遗传算法的原理之前，我们先引入以下遗传算子。

（1）选择算子：选择算子是算法模拟自然界中生物群体受自然界影响所表现出的优胜劣汰的过程，即种群中越优秀的个体在自然界的筛选下就越容易传递给下一代。算法根据个体适应度值判断种群中的个体优劣，然后依据优胜劣汰原则采用选择算子对种群进行择优筛选。目前常用的选择操作有轮盘选择机制、适应度比例方法和随机抽样等。

（2）交叉算子：大自然中的大多数生物个体都是由父辈的雄性和雌性个体

交配生成的，交叉算子就是模仿大自然中两个生物个体交配生成新个体的步骤。

从种群中选取两个目标个体，利用交叉算子所制定的个体交叉方式对所选个体的

基因进行交叉操作，生成两个具有新基因的下一代。这种交叉操作能够有效地保

证种群的多样性。目前常用的交叉操作有单点交叉和多点交叉等。

（3）变异算子：大自然中生物群体可能会因为适应新的生存环境而发生小概率的基因突变，这一过程是由算法中的变异算子所模拟的，即每个单独个体染色体上的基因都有一定概率发生改变。根据自然界的优胜劣汰，良性的变异会保留下来，恶性的变异则会被淘汰，因此变异操作能够加速种群向最优解收敛，同时增加了生物群体在优化过程中的个体多样性，有效地避免种群进入早熟收敛。

当前常用的变异算子包括实值变异、二进制变异等。

基本遗传算法的思想过程如下：

（1）设置种群规模大小 N、交叉率 pc、和变异率 pm；

（2）在解空间内随机生成 N 个个体作为初始种群；

（3）利用算法给定的适应度函数计算种群中每个个体的适应度值；

（4）利用选择算子制定的选择方式对种群进行选择操作；

（5）将（4）中得到的种群根据交叉算子设定的交叉方式和交叉率对种群中个体进行两两交叉；

（6）将（5）中得到的种群根据变异算子设定的变异方式和变异率对每个个体进行变异操作。

（7）如果满足终止条件，则算法结束，输出适应度最优的个体；否则返回步骤（3）。

遗传算法最大的优点是算法的原理以及种群的搜索过程简单，且容易与其它算法结合；但是算法可能会对问题的解集进行编码和解码，因此在实际编程实现时会相对比较复杂；其次算法中涉及到的参数较多且严重影响最终的解；而且算法过程中种群内部的个体之间没有交互作用，最后算法的好坏一定程度上也依赖初始种群的选择。

**2.2.1 蝴蝶算法**

蝴蝶算法（Butterfly Optimization Algorithm，BOA）是一种元启发式仿生群智能优化算法，它模拟蝴蝶的觅食行为。

思想的条件假设如下:

1) 所有的蝴蝶都应该散发出某种香味, 使蝴蝶能够互相吸引。

2) 每只蝴蝶都会随机移动, 或朝着散发出更多香味的最佳蝴蝶移动。

3) 蝴蝶的刺激强度受目标函数值的影响或决定。

当蝴蝶能感觉到其他任何蝴蝶的香味时并朝它移动, 在该算法中, 该阶段称为全局搜索. 在另一种情况下, 当蝴蝶不能感觉周围的香味时, 然后它会随机移动这个阶段称为局部搜索. 利用转换概率控制全局和局部搜索过程, 其迭代公式为

式 (1)中的f是香味感知量, c是感觉形式,I是刺激强度,a是香味,通常 a和c的取值范围为[0, 1]之间;式(2)中的表示第i只蝴蝶在第t+1代的位置, 的随机数, 表示全局最优解, 表示第i只蝴蝶的香味感知量; 式(3)中的, , 分别表示第i,j,k只蝴蝶在第t代的位置;式(4)中的为常数, 表示第t代的值,Ngen 为最大迭代次数。

该优化算法，能够在局部搜索和全局搜索之间得到好的平衡，可以解决很多优化问题，

**第3章　改进蝴蝶优化聚类算法**

**3.1蝴蝶优化算法的改进**

**3.1.1 指数收敛因子**

为了增强BOA算法的探索能力和提供收敛精度，引入线性收敛因子(Linear convergence factor)α在蝴蝶森爱的全局位置更新处，为其设计的迭代前期α较大可以增强全局探索能力且递减速度快，而在迭代后期α收敛到较小值且递减速度变缓慢，以实现前期加速收敛，后期保证收敛精度在较高水平。Α随着迭代次数由2减小到0。文献收敛因子和黄金指引机制的蝴蝶优化算法，设计的公式是：

式中：t是当前迭代次数，是最大迭代次数。

一种改进的蝴蝶优化算法中实验指出，α递增算法性能也越来越差，当α算法随指数性能最优时。本文引入一种新的基于指数变化的收敛因子(Exponential connective factor)更新方式。



**3.1.2 融入差分进化策略和精英策略**

差分进化是一类基于群体的自适应全局优化算法，与遗传算法类似，不过不采用二进制编码，而是实数编码。它的进化也包括变异，交叉和选择。

变异操作：在每次迭代中基于指数收敛因子的蝴蝶算法融入差分进化策略（differential evolution）和精英(elite)策略。为了加强种群个体交互信息，做出了优化。

为了验证改进的指数收敛因子的有效性，实验首先进行了比较经典的蝴蝶优化算法，线性收敛因子蝴蝶算法和指数收敛因子蝴蝶优化算法在求解Sphere（单峰函数，公式）和Rastrigin函数（多峰函数）最小值问题过程中的优化结果。种群数量50，最大迭代次数500。动态转换概率0.8，C取0.01，α取0.1。结果如下，结果为算法独立运行30次取平均值。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 测试函数 | BOA算法 | LBOA算法 | EBOA算法 | EDEIBOA算法 |
| Sphere函数 | 0.026949 | 0.039431 | 1.255e-09 | 9.69674e-18 |
| Alpine函数 | 1.679137 | 0.839685 | 0.0001138 | 4.803655e-05 |

算法的改进算法的时间复杂度分析算法的种群规模设置为 N，最大迭代次数为 maxiter，搜索空间的维度为 dim，增加 差分进化策略，相当于增加一个内层循环( 由改进算法的流程图可以直接看出) ，则增加的运算量为 O( maxiter × dim × N) ，相当于增加了一个内层循环，根据改进算法的寻优效率来说，增加的计算量是可以接受的。

可以看到我们最终的最优的算法。相对于最初的基本蝴蝶算法。对于Sphere函数精度提升了16个数量级，对于Alpine函数精度也提升了5个数量级。

验证了EDEIBOA算法的有效性之后，我们在进一步验证与之结合的K-means 算法在不同数据集中的聚类结果表现。

|  |  |  |
| --- | --- | --- |
| 数据集 | 数据个数 | 属性个数 |
| Iris | 150 | 4 |
| wine | 178 | 13 |
| Absenteeism\_at\_work | 740 | 21 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | DBI指数 | K-means | BOAK | EK |
| Iris | 最优值 | 0.4042 | 0.3747 | 0.3747 |
| 最差值 | 1.1313 | 0.4648 | 0.3827 |
| 平均值 | 0.6671 | 0.3855 | 0.3782 |
| Wine | 最优值 | 0.5380 | 0.2959 | 0.3842 |
| 最差值 | 0.7181 | 0.5300 | 0.2724 |
| 平均值 | 0.6692 | 0.4014 | 0.3252 |
| Absenteeism\_at\_work | 最优值 | 1.0237 | 1.0602 | 0.8907 |
| 最差值 | 1.3118 | 1.2151 | 1.2043 |
| 平均值 | 1.1755 | 1.1786 | 1.0796 |

3.4.2 实验结果分析

为了比较本实验中方法的性能，在推荐指标的选择上，选择用 RMSE（均方根误差）

作为评价指标。RMSE 指标广泛应用于评估推荐系统中，具体来说，RMSE 提供了有关用

户选择的真实内容和预测的可能性内容之间的差异。RMSE 值越小表明推荐系统性能越好。计算公式如(3-7)：

其中𝑝,𝑖是用户𝑢对内容𝑖的预测评分，𝑟,𝑖表示实际评分，而𝑁表示对此内容的评分总分。在实验中我们将本文提出的基于降维和聚类的协同过滤算法与基于 K 均值算法的协同过滤算法和基于 K 近邻的协同过滤算法进行比较，邻居范围设定为 10-50。实验结果

如表 3-10 所示。从表中可以看出本文提出的算法在三个数据集上的表现都不错，都要比另外两个基于传统的聚类算法的效果要好。

kmeans

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 聚类数量 | 3 | 5 | 9 | 10 | 11 | 15 |  | 20 |
| MAE | 0.81 | 0.82 | 0.81 | 0.782 | 0.81 | 0.82 |  | 0.826 |

与传统的协同过滤算法UCF，基于K-means的协同过滤算法，

**第3章　基于蝴蝶优化聚类的协同过滤电影推荐系统**

系统的开发采用Windows 10系统，数据库设计采用Mysql数据库，后端语言采用Python，此外，此外后端框架采用Flask，前端界面采用Boostrap和Jquery框架来实现。对于系统整体，采用MVC设计模式，该模式在系统开发中最为常用，让系统易于维护。

功能需求分析

功能需求分析就是要明确该系统所含有功能。系统的使用对象不同，提供的功能也会有所不同。根据系统的使用对象，对系统进行功能需求分析。

(1) 游客

我们将未登录的用户称之为游客。当游客使用该系统时，系统内并没有该用户的历史信息，所以系统对该类用户只会提供两种推荐列表，即热门电影列表和最新电影列表。此外，该用户还可以进行分类电影的查询以及相关电影信息的查看，但不能进行电影的观看以及对电影的评论、评分等操作。

(2) 普通用户

我们将已注册的用户称之为普通用户。普通用户可以使用已注册的账号、密码登录电影推荐系统，登陆成功后，系统会对该类用户提供三种推荐列表，即多样化电影列表、热门电影列表和最新电影列表。此外，用户还可以进行个人信息的查看以及修改，电影信息的查看以及相关电影的查询，进行电影的观看，并且可以对电影发表个人评论，用户还可以根据个人对电影的喜爱程度对电影进行打分，分数为 1-5 之间的整数，分数越高，表明该用户对此电影越喜欢。图 4.2 是普通用户的用例图。

(3) 系统管理员

系统管理员是系统内不可缺少的角色。管理员登录系统，进入后台管理界面，其主要任务是电影信息的管理。对于新上映的电影，管理员需要收集相关影讯信息，主要是电影剧情介绍、导演、主要演员、上映日期、电影类型等信息，并将电影添加到数据库；对于数据库内已有的电影，管理员可以执行查看、修改、删除等操作。此外，管理员还拥有系统说明管理、系统前台轮播图管理等操作。图 4.3 是系统管理员的用例图。

游客用例图如图 4.1。

本文的系统使用 Mysql 数据库保存信息，将收集到的用户信息和推荐算法生成的推荐结果都保存在数据库中。数据库中包括四张表来保存信息：User 表保存客户基本信息；Movie 表中保存影视内容基本信息；Ratings 表中保存影视内容属性及其评分等信息；Recommend 表中保存预测结果。

1. 用户信息表User表

用户信息表包括用户 id、用户姓名、用户密码、邮箱、手机号、简介、头像和用户评论外键，电影收藏外键

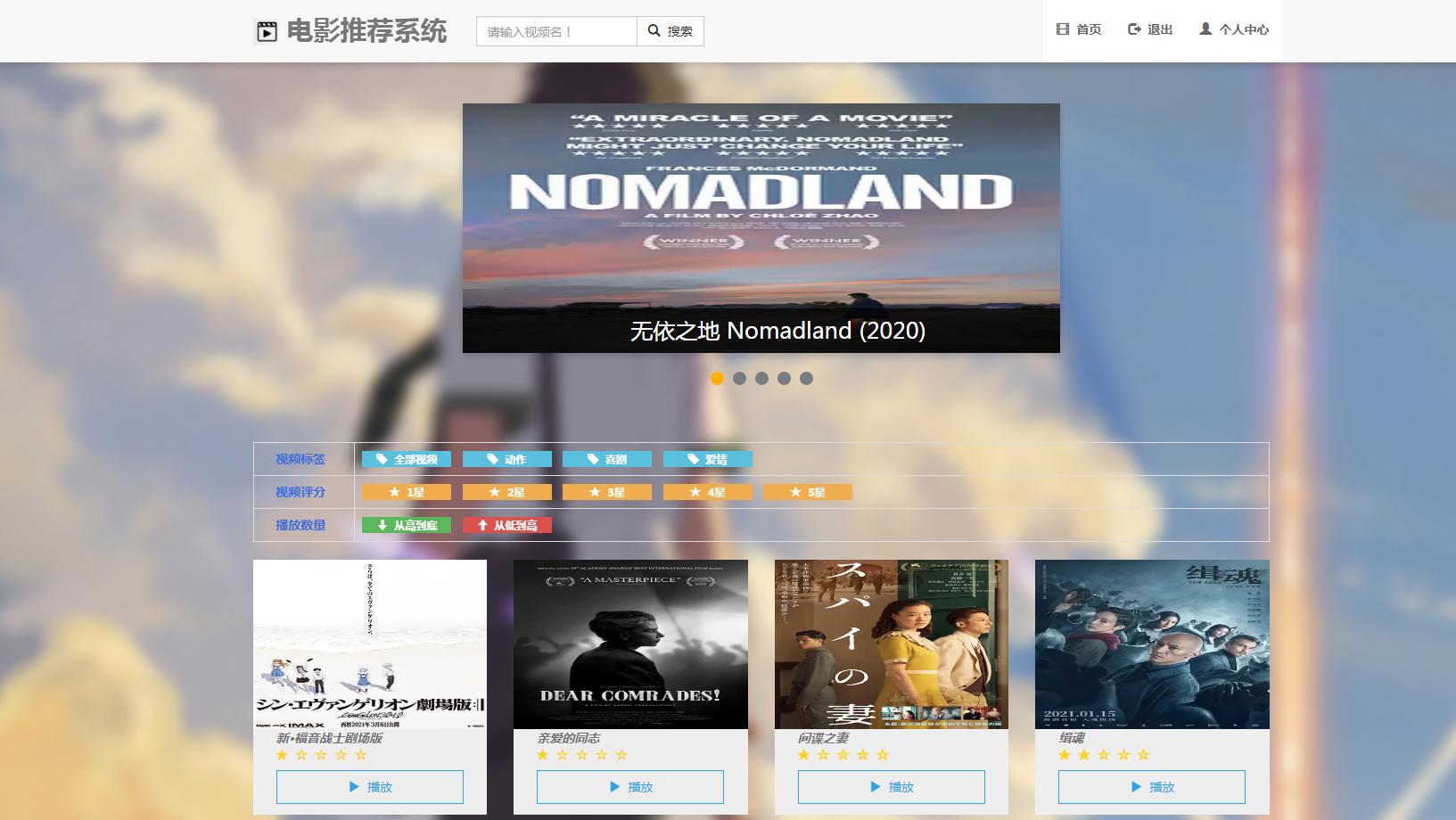
，如表 4-1 所示。

|  |  |  |
| --- | --- | --- |
| Field | Type | Desc |
| id | Bigint10) | 用户id,主键，自增 |
| name | Varchar(100) | 用户名称 |
|  |  |  |
|  |  |  |

****







**42**

**图 4.30 多样化电影推荐图**

**4.6 系统测试**

多样化电影推荐系统开发完成后，需要对其进行测试。功能测试是系统测试的重要环

节之一。利用功能测试，可验证系统功能是否符合用例图的设计，其功能是否达到用户要

求。文中我们选用的功能测试方法包括页面链接检查、相关性检查、search 检查等，对用

户主要功能和管理员主要功能进行了测试。

表 4.7 是管理员主要功能的测试结果。

表 4.7 管理员功能测试结果

测试功能 测试次数 成功次数 失败次数 未运行次数 是否通过

登录验证 30 30 0 0 是

电影搜索 30 30 0 0 是

电影添加 40 40 0 0 是

电影删除 20 20 0 0 是

电影修改 10 10 0 0 是

类别添加 30 30 0 0 是

类别删除 20 20 0 0 是

轮播图添加 20 20 0 0 是

轮播图删除 20 20 0 0 是 第 4 章 基于 DivRec\_LSH 方法的电影推荐系统的设计与实现

43

表 4.8 是对用户主要功能的测试结果。

表 4.8 用户功能测试结果

测试功能 测试次数 成功次数 失败次数 未运行次数 是否通过

注册验证 30 30 0 0 是

登录验证 30 30 0 0 是

依据类别 查

询电影

20 20 0 0 是

依据片名 查

询电影

20 20 0 0 是

个人信息 修

改

10 10 0 0 是

电影评论 20 20 0 0 是

电影评分 30 30 0 0 是

多样化推 荐

列表查看

20 20 0 0 是

**第 5 章 总结与展望**

随着互联网的发展，网络中的信息呈爆炸式增长，这虽然可以满足用户的信息需求，但也带来了“信息过载”的问题。错综复杂的信息让用户眼花缭乱，以至于用户无法快速发现自己所需要的信息。对于这一问题，推荐技术是最为有效的解决方案。通过不同的推荐方法，推荐技术可以自动筛除掉一些垃圾信息，向用户推荐其感兴趣的信息，不仅提高了用户信息搜索的效率，而且丰富了用户体验。

协同过滤方法，凭借着“易解释、领域无关性”等特点成为了最为常用的一种推荐方法。然而，对于传统的协同过滤方法来说，在进行相似项或相似用户查找时常常需要大量的计算，这就会使得时间消耗增多，推荐效率较差，系统扩展性不高。与此同时，常用的协同推荐方法中一般只考虑推荐的准确度，并没有考虑推荐列表中项与项之间的相似性问题，由此带来推荐结果冗余化，使得用户满意度降低，从而影响用户对推荐系统的使用感。针对上述问题，本文用以下方法进行了解决：

(1) 为提高传统协同过滤方法低效率推荐以及冗余推荐问题，本文利用局部敏感哈希技术实现协同过滤，对数据库中的项进行项目索引值的计算，该过程离线即可完成。当系统需要查找与用户历史项目相似的项时，只需利用用户历史项目的索引值与数据库中其它项的索引值进行匹配即可，经过此过程，相似搜索的时间复杂度即可由 O(m\*n)降至 O(1)，系统可快速找到候选项，从而推荐效率也得到提高。在为用户生成推荐列表时，使用评分多样性对列表进行二次优化，除去列表内相似影片，在保证推荐准确度的情况下，尽可能为用户推荐多样的电影，使用户满意度有所提高。

(2) 为了验证上述两点的有效性，本文选用 MovieLens 数据集，通过设计对比实验，根据常用的推荐系统指标，对实验结果进行分析，证明了该方法是可行的。 (3) 基于 B/S 架构模式，结合上述研究成果，设计并实现了多样化电影推荐系统，通过对系统进行可行性分析、功能性需求与非功能性需求分析，以及对系统的总体的设计，使用 Java 编言对系统的功能模块进行实现。 虽然本文提出的方法可以在保证推荐效率和推荐准确度的同时提高推荐多样性，且设计的多样化电影推荐系统能够独立运作。但是由于本人能力问题，改进的算法与实现的系统都有相应的不足之处，在未来的研究中需进一步弥补。总体来说，可以概括如下：

(1) 本文利用局部敏感哈希技术实现协同过滤，而局部敏感哈希技术是一种近似方法，它是基于概率的，推荐性能受哈希函数个数的影响。所以在实际应用中，需要将哈希函数个数定义为一参数，对其参数调节，以确定最优的哈希函数个数，实现最优的推荐结果。在之后的研究工作中我们将尝试划分数据量的范围来限定哈希函数的个数，以减少参数调节的过程。

(2) 对于多样化方法，我们考虑的维度较为单一，并没有考虑电影的多个维度，例如导演、演员等，在之后的工作中，我们可以进一步考虑电影的其它维度，并考虑使用加权函数将电影的多个维度进行结合，以衡量电影间的相似性。此外，我们并没有考虑每个用户对推荐列表多样化的要求，比如有些用户可能并不太希望列表多样，这些都是我们以后工作中可以进一步完善的地方。 由于时间与能力有限，本文所设计实现的电影推荐系统，有些模块功能可能并不够完善，页面可能不够美观。除此之外，当用户量过大时，该系统可能面临服务器负载过大问题。在之后的研究中，需继续对系统进行完善。

**参 考 文 献**

-5-

[1] 赵晓，庞劲风，杨春光.浅析我国锦纶6帘子布生产现状.合成纤维工业，2001，24（1）：43-46

[2]  王德诚.我国轮胎帘子布供求现状与发展前景.中国化纤信息，1998，（9）：14-16

[3] 中国科学院大连化学物理研究所.气相色谱法. 北京：科学出版社，1972，44-46

[4] Crippen R C. Indentification of Oragnic Compounds With the Aid of Gas Chromatography. New York: Mc Graw Hill, 1993,36-39

1. Pecsok R L, Shieds L D, Cairns T, et al. Modern Methods of Chemical Analysis, New York: John Wiley & Sons, 1976,67-70

附录A 攻读学位期间所发表的学术论文目录

-6-

-10-

-8-

致 谢

时光飞逝，白驹过隙，转眼三年研究生生活即将告一段落。但这段学习的时光，我也在不断成长，我感受到这是一种财富。

感谢戴牡红老师， 在学业方面的指导，戴老师对我学习期间，给了我充分的自由，让我可以依照自己的兴趣，学习并做一些自己的研究，我的论文离不开他指导。也感谢学校可以给我们宽松的环境，能够快乐学习，感谢其他老师的付出，感谢同学的关爱。

也感谢一直默默支持我的家人。他们是我一直学习的坚强后盾，让我没有后顾之忧。他们的无私的爱让我感到力量满满。

最后，感谢所有评阅专家于百忙之中评阅此文。