基于OLAP与数据挖掘的高考招生数据分析

何小明 张自力 肖 灿 夏大飞
(西南大学智能软件与软件工程重点实验室 重庆 400715)

摘 要 如何从海量的高考招生数据中发现有用信息,是招生主管部门迫切关心的问题,也是家长、考生以及社会各界都十分关注的问题。围绕这一问题,依据某省多年来累积的高考招生数据,建立数据仓库和多维数据集,进行OLAP分析与数据挖掘分析,得到了一些潜在的有用信息。研究分析表明,这些信息可以为招生主管部门提供决策支持,也可作为指导考生合理填报志愿的重要依据。介绍了数据仓库和多维数据集的建立过程、录取相关数据的OLAP分析及其结果的解读过程以及利用决策树算法和关联规则算法进行数据挖掘的过程。

关键词 数据仓库,OLAP,数据挖掘,决策树,关联规则
中图法分类号 TP311 **文献标识码** A

Data Analysis on National College Entrance Examination and Admission Using OLAP and Data Mining

HE Xiao-ming ZHANG Zi-li XIAO Can XIA Da-fei
(Key Laboratory of Intelligent Software and Software Engineering, Southwest University, Chongqing 400715, China)

Abstract How to find useful information from massive data of national college entrance examination and admission (NCEEA) is one of the key issues of the provinces and cities admission offices(PCAO), and it also attracts the attention of parents and examinees as well as all sectors of the society. Around this issue, a data warehouse and a data cube were built based on the admission data accumulated over the years in one province. Some potential and useful information was found through OLAP analysis and data mining analysis. Research and analysis represent that these information can provide decision-making support for the PCAO and can be important basis for guiding examinees to choose preference reasonably. This article focused on the process of building data warehouse and cube, the process of OLAP analysis on admission related data and the result interpretation, the process of data mining by using decision tree algorithm and associate rule algorithm.

Keywords Data warehouse, OLAP, Data mining, Decision tree, Associate rule

1 引言

自实行网络招生以来,全国、省市各级招办都积累了十多年的招生数据,这些数据中包含了大量有用信息。但因为统计方法与工具的欠缺,这些宝贵的数据还没有得到有效的利用。如何从海量的高考招生数据中发现有用信息,是招生主管部门迫切关心的问题,也是家长、考生以及社会各界都十分关注的问题。经验表明,从大量历史数据中找出有用信息,可以为招生主管部门和高校提供决策参考,也可以为考生填报志愿提供支持和指导。

以数据仓库、联机分析处理(OLAP)和数据挖掘为核心的商务智能技术[1]已经被应用到许多领域(如电信、银行、保险、零售等),而且正处于高速发展的过程中[2],但是在高考招生数据分析领域目前还没有成功案例可供参考。围绕上述社会各界广泛关注的问题,我们团队借助数据挖掘这一有效的海量数据分析技术,研究某省近10年来的高考招生数据,并已依据该数据建立了数据仓库。在此数据仓库的基础上进行OLAP分析与数据挖掘,可展现出该省招生历史数据中蕴含的一些有用信息。殷员分研究了数据仓库与OLAP技术在高考志愿数据分析中的应用[3];蔡海敏提出了一种应用数据仓库与OLAP技术评估高考加分政策的方案[4];曾铮研究了OLAP技术在高考志愿填报方式分析评估中的应用[5]。

本文在上述工作的基础上完善了数据仓库,建立了统一的多维数据集,加强了对于OLAP分析结果的展示,并且添加了对于数据挖掘的支持,在实际分析中应用了决策树算法、关联规则算法和朴素贝叶斯算法等数据挖掘算法。结合实际情况的分析表明,OLAP分析和数据挖掘可以得到一些有用信息,这些信息一方面可以为招生主管部门提供决策支持,另一方面也是指导考生进行志愿填报的重要依据。本文将介绍多维数据集的建立过程、OLAP分析过程及对其结果的解读,以及利用决策树算法和关联规则算法进行数据挖掘的过程。

到稿日期:2011-08-25 返修日期:2011-11-20 本文受重庆市科技攻关计划项目(CSTC,2009AC2174)资助。
何小明(1985—),男,硕士生,主要研究方向为商务智能与数据挖掘,E-mail:heming07@swu.edu.cn;张自力(1964—),教授,博士生导师,主要研究方向为多Agent系统、混合智能系统、数据挖掘;肖 灿(1986—),男,硕士生,主要研究方向为商务智能与数据挖掘;夏大飞(1983—),男,硕士生,主要研究方向为商务智能与数据挖掘。

2 数据仓库设计与实现

2.1 数据仓库的系统结构

数据仓库是一个面向主题的、集成的、时变的和非易失的数据集合，支持管理部门的决策过程[6]。数据仓库是数据分析和联机分析处理日趋重要的平台，并将为数据挖掘提供有效的平台。因此，数据仓库模型的设计非常重要，它影响着后来一系列的工作。本文中的数据仓库的系统结构(见图 1)分为 3 层：底层是数据仓库服务器，主要存储来源于源数据层的数据；中间层是 OLAP 服务器，为上层提供多维数据集；顶层是前端展示层，利用报表、网页、应用程序等多种方式展现分析和挖掘结果。

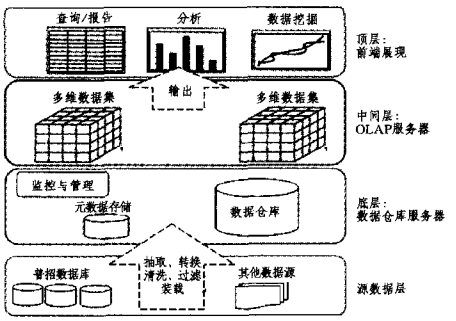

图 1 数据仓库系统结构[3]

2.2 数据仓库建模

数据仓库需要简明的、面向主题的模式，以便于联机数据分析与数据挖掘。因此，结合高考招生数据本身的特点，以及实际调研与数据分析的结果，高考招生数据仓库的主题可确定为考生成绩主题、考生志愿填报主题、考生志愿投档主题、考生加分政策主题、考生录取结果主题等。目前流行的数据仓库多维模型有星型模式、雪花型模式和事实星座型模式[6]。本文采用星型模式，它是目前最常用的数据仓库多维模型，也符合高考招生数据的特征。相对于雪花与星座模型，星型模式具有直观、查询效率高和易于数据装载的特点。

在星型模式中，数据表主要分为两种。一种是包含大批数据并且不含冗余的中心表(事实表)，另一种是小的附属表(维表)。图 2 展示了考生录取结果主题对应的设计模型。其中录取事实库是事实表，它对应着院校代号维、考生维等多张维表。

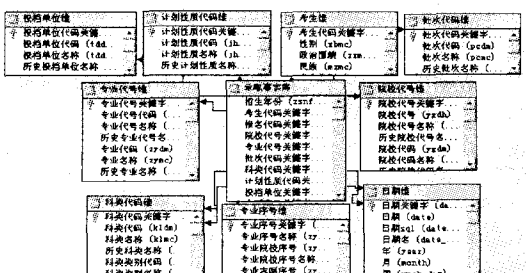

图 2 数据仓库设计模型

2.3 ETL 实现

数据存储是数据仓库的基本功能，需要利用 ETL 工具向数据仓库装载数据。ETL 亦即数据抽取(Extract)、转换(Transform)、装载(Load)[7]。良好的数据质量是分析结果可靠的基础，因此在 ETL 过程中必须保证数据的完整与准确。

高考招生是一个动态变化的过程，高考招生数据的格式每年不尽相同，引起这些数据差异的原因包括高考招生政策的调整、志愿填报方式的改变、院校名称的变化等。本文在进行 ETL 的过程中面临的最大困难就在于对维度表的处理。本文首先深入分析了历年招生数据，建立了一套编码标准；然后利用“自动处理＋人工干预＋主管部门监督”的模式逐年导入数据。在具体的实现过程中，采用的主要工具是 SSIS(Microsoft SQL Server 2008 Integration Services)。由于 ETL 过程工作量大，数据质量要求高，因此 ETL 过程的圆满完成需要大量时间和精力，约占整个项目工作量的 60%左右，这也是众多实践的普遍规律[8]。

3 多维数据集的建立

3.1 主题的确定

对数据的多维分析并不是主要针对数据仓库的，而是针对数据仓库提取的子集，如多维数据集。因此，在源数据加载到数据仓库完成以后，就要根据高考招生数据的特征来设计面向不同主题的多维数据集。根据设计数据仓库时的思路，多维数据集的主题包括考生成绩主题、考生志愿填报主题、考生志愿投档主题、考生加分政策主题、考生录取结果主题、享受加分政策的考生主题、考生专业调剂主题等。

3.2 实现过程

为了方便操作与管理，本课题采取了建设一个中心多维数据集的方案，即所有的主题集中在一个多维数据集中，每个主题在多维数据集中表现为一个度量值组。每个度量值组包括若干跟本主题相关的度量值和计算值，所有的度量值和计算值共用维度。本文利用 SSAS(Microsoft SQL Server 2008 Analysis Services)建立多维数据集，采用 MLOAP[9] 方式存储。其中录取主题的多维数据集模型(部分)如图 3 所示。

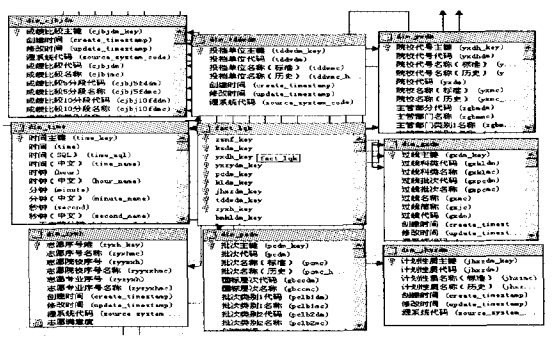

图 3 录取主题的多维数据集模型(部分截图)

鉴于 SSAS 中常用的度量值不能完全满足 OLAP 需求，本文在各个度量值组定义了若干计算值。计算值由多维数据集语言 MDX[10] 来定义，如计算录取的考生人数与报名考生人数的比值的定义为：

[录取的考生人数与报名考生人数的比值]＝[Measures].[录取的考生人数] /[Measures].[报名考生人数]

生成多维数据集之后，可以在其上进行切片、钻取等操作，从而可以多角度、多层次分析高考招生数据。

4 OLAP 分析与前端展现

4.1 OLAP 分析结果及说明

4.1.1 从考生分布看基础教育的发展

近年来，教育部比较重视基础教育，尤其是农村基础教

育[11]。研究历年来高考报名考生人数中农村考生的比例，有助于了解农村基础教育的发展状况。图 4 是历年报考考生按城镇、农村分类的情况。从图中可看出，统计的 10 年中，农村考生在报名人数中的比例一直在上升，其中 2002 年到 2004 年增长最快。

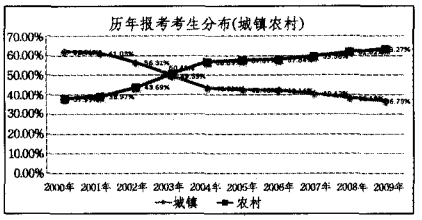

图 4　历年报考考生分布(城镇、农村)

图 5 是历年录取的考生按城镇、农村来源分类的情况。与图 4 类似，农村考生的比例呈明显的上升趋势，但在 2004 年到 2008 年前有微小起伏。

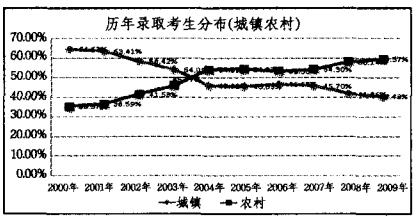

图 5　历年录取的考生分布(城镇、农村)

图 6 与图 7 列出了历年在报考和录取阶段考生人数的增长率变化情况。从报考情况可看出，2001 年－2005 年对于农村考生来说是高速增长时期，尤其是 2004 年报考人数呈现暴涨状态。从录取情况来看，2001 年－2004 年是农村考生高速增长时期。因此可以认为，2001 年－2005 年是农村基础教育好转的几年。

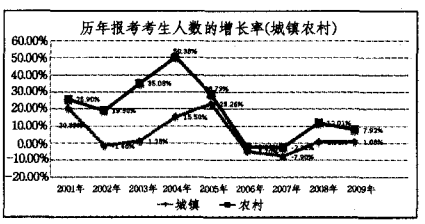

图 6　历年报考考生人数的增长率(城镇农村)

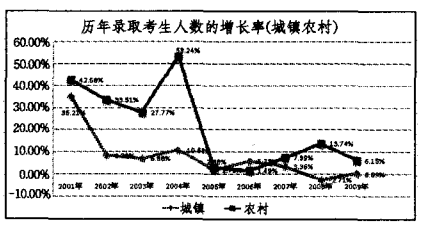

图 7　历年录取考生人数的增长率(城镇农村)

4.1.2　往届生的趋势分析

往届考生对于高考招生来说一直是一个热门话题，但目前仍存在着各种各样的争议。图 8 与图 9 反映了历年来应届考生和往届考生在报考人数和录取人数中的分布情况。从图 8 可以看出，报考人数中往届考生的比例在 2004 年－2006 年有明显上升趋势，从 2006 年－2009 年一直占据着总人数的 20%左右。这种情况与目前国家与社会提出的减少往届生的目标有点南辕北辙。从图 9 来看，录取考生中往届生的比例也是在 2004 年－2006 年有明显上升过程。而且跟报考比起来，录取的考生中往届考生所占的比例都更大，这说明往届考生的录取率更高，从而证明往届考生在高考分数竞争中总体比应届考生有优势。

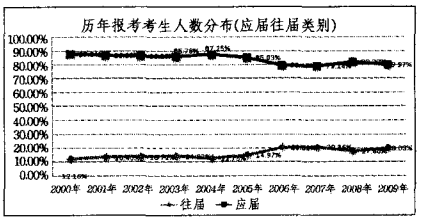

图 8　历年报考考生人数分布(应届往届类别)

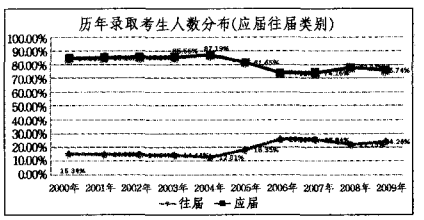

图 9　历年录取考生人数分布(应届往届类别)

4.2　OLAP 前端展现

基于 Sql Server 的 OLAP 前端展示工具包括 Excel2007、SSRS(SQL Server Reporting Services)，以及其他第三方工具。SSRS 具有展现形式丰富、开发方便、易于与 Web 平台集成等特点[12]，因此本文采用 SSRS 作为 OLAP 前端展现的工具。

从便于用户获取分析结果及部署的角度出发，本文采用了 B/S 模式来展现分析结果，利用微软公司发布的 SSRS 网页插件将报表服务器与 Web 系统集成在一起，使得用户可以通过浏览器获取多种形式的 OLAP 分析结果。为有效保证数据的安全，Web 系统中应用了比较完善的安全措施及权限管理模块。

5　数据挖掘

简单地说，数据挖掘是指从大量数据中提取或“挖掘”知识。本文的研究重点之一就是利用数据挖掘技术从大量招生历史数据中找出隐含的规律或知识。本文挖掘高考招生数据的主要步骤包括确定挖掘主题，数据选择，数据预处理，建立挖掘模型，评估挖掘模型，分析与展现挖掘结果。

根据应用场景的不同，数据挖掘有多种算法可供选择。本文采用 SSAS 平台提供的挖掘算法有决策树算法、关联规则算法、神经网络算法、聚类分析算法等 7 种[13]。下面以决策树算法和关联规则算法为例，展示数据挖掘在本文中的应用。

5.1　决策树算法

众所周知，在高考战场上是否享受加分政策对于一个考生是否被录取有比较大的影响[14]。本文中挖掘主题之一是“影响考生是否加分的因素”。根据对主题的理解，选择的输入属性包括报名单位名称、毕业类别名称、城乡名称、户口所在地域名称、户口所在一圈两翼名称、考生代码、考生类型代码、性别名称、应往届名称、中学类别名称、中学所在地区名称。输入数据源是针对此主题而建立的视图。挖掘模型中，

“是否加分”属性作为惟一的预测列，其他属性都作为输入列。

本文将挖掘结构的数据分为两部分：70%作为训练集，30%作为测试集。通过对模型的训练和处理，可得到输入属性与预测属性的依赖关系网络。如图 10 所示，左边滑块表示关联强弱，上下拉动滑块对应的关系网络图会发生变化。图中关系表明与考生是否享受加分政策最相关的 3 个输入是“报名单位名称”、“城乡名称”、“毕业类别名称”。

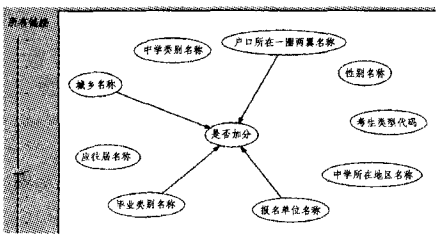

图 10　决策树依赖关系网络(截图)

决策树的模型结果如图 11 所示，本模型生成的决策树太大，图中只反映了决策树的一部分内容。从图中可以看到，第一个树分支都跟“报名单位名称”有关，每两个分支取决于“户口所在一圈两翼名称”或“考生类型代码”。

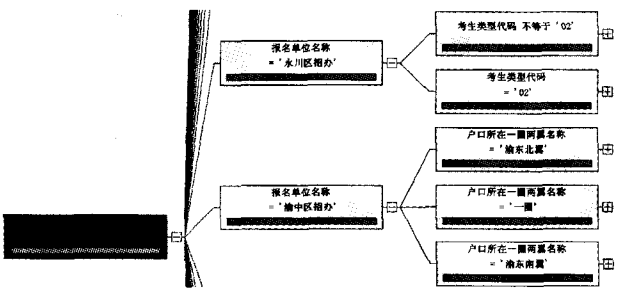

图 11　决策树模型结果(部分截图)

决策树根节点到每层叶子节点的路径代表一条规则。把决策树全部展开，可得到更详细的信息。图 12 所示为某结点的挖掘图例，其包含了该结点的规则和相关信息。从图中可看出，满足此规则的考生享受加分的概率达 88.36%。

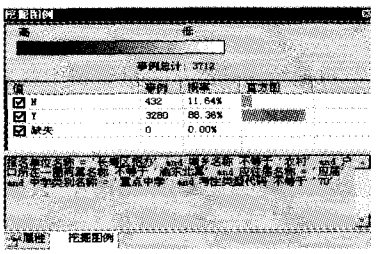

图 12　决策树挖掘图例(截图)

5.2　关联规则算法

关联规则挖掘用来发现大量数据中项集之间有趣的关联或相关联系，它是数据挖掘的一个重要研究课题，最近几年已被业界广泛研究而且实用较广，比较常见的应用是购物篮分析[15]。在本文中，研究考生填报的专业之间的关联性，可以为后续的志愿填报推荐提供数据支撑。

首先在数据仓库中建立名为“考生填报的所有专业”的视图，视图共有 3 个字段：考生代码、专业名称、专业序号，视图中的数据可理解为历年来所有考生填报的所有专业的集合。在 SSAS 的挖掘模型中，用上述视图作为数据源，“考生代码”作为关键字列，“专业名称”作为关键字列和预测列。挖掘模型采用 SSAS 提供的关系规则算法，其中算法参数采用默认值，如图 13 所示。

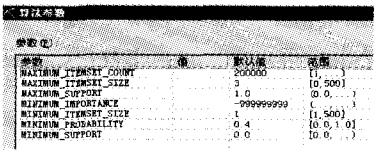

图 13　关系规则算法参数设置(截图)

同样将输入数据的 70%作为训练集，30%作为测试集。训练和处理后，得到的依赖关系网络如图 14 所示。可看出，考生填报的专业中，部分专业之间存在强弱不等的关联性。

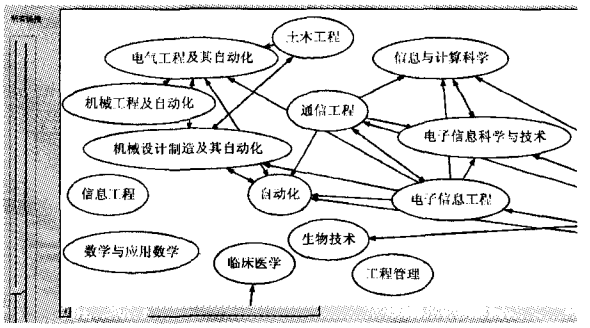

图 14　关联规则依赖关系网络(部分截图)

在查看“规则”时，将“最小概率”设为 0.60，“最低重要性”设为 0.30，得到对应的关系规则有 718 条，其中部分规则如图 15 所示。在此基础上，可进一步通过 DMX 查询或规则导出等方式，将挖掘结果用于指导志愿填报。

概率	重要性	规则
0.903	0.591	信息与计算科学，电子信息科学与技术 -> 计算机科学与技术
0.896	0.597	信息与计算科学，信息管理与信息系统 -> 计算机科学与技术
0.891	0.641	信息与计算科学，电子信息工程 -> 计算机科学与技术
0.880	0.603	信息与计算科学，通信工程 -> 计算机科学与技术
0.880	0.573	信息与计算科学，自动化 -> 计算机科学与技术
0.878	0.579	电子信息科学与技术，信息管理与信息系统 -> 计算机科学与技术
0.877	0.567	信息与计算科学，机械设计制造及其自动化 -> 计算机科学与技术
0.873	0.585	信息与计算科学，工商管理 -> 计算机科学与技术
0.869	0.412	行政管理，市场营销 -> 工商管理
0.865	0.410	行政管理，经济学 -> 工商管理
0.863	0.563	信息与计算科学，土木工程 -> 计算机科学与技术
0.863	0.409	行政管理，金融学 -> 工商管理
0.862	0.651	信息与计算科学，通信工程 -> 电子信息工程
0.861	0.649	电子信息科学与技术，通信工程 -> 电子信息工程
0.855	0.405	行政管理，会计学 -> 工商管理

图 15　关联规则算法挖掘结果(截图)

5.3　数据挖掘前端展示

挖掘模型处理完成后，本文同样利用了相关的 Web 插件来展示挖掘结果。为了更丰富地展现挖掘结果，还利用了微软公司的 ClickOnce 部署技术，用户可以通过点击网页打开基于 Windows 的应用程序来查看挖掘结果。利用 ClickOnce 部署的应用程序，具备了安装便利和用户界面丰富的双重优点，提高了整个系统的交互性和用户友好性[16]。

由于篇幅有限，不能展示更多的挖掘过程和结果。针对不同的主题，采用聚类、神经网络等更多挖掘算法，会得到更多有用的知识。

结束语　本文以某省近 10 年的高考招生数据为源数据，在已有的数据仓库基础之上建立了多维数据集，进行了 OLAP 分析和数据挖掘，得到了一些有用信息，证明了商务智能技术应用于高考招生数据分析的可行性。下一步的工作除了继续完善数据仓库和深入数据挖掘外，更重要的是研究将 OLAP 与数据挖掘的结果应用于指导考生填报志愿，形成一

(下转第 187 页)

表 2　美国 10 个城市的距离

城市	0	1	2	3	4	5	6	7	8	9
0 芝加哥	0	960	1050	500	410	860	460	290	560	700
1 达拉斯	960	0	780	490	940	510	640	630	410	370
2 丹佛	1050	780	0	600	840	610	540	860	760	510
3 塔萨斯	500	490	600	0	450	350	200	260	170	180
4 明尼阿波利斯	410	940	840	450	0	800	360	550	590	640
5 俄克拉荷马	860	510	610	350	800	0	460	500	280	80
6 奥马哈	460	640	540	200	360	460	0	450	370	300
7 圣路易	290	630	860	260	550	500	450	0	210	450
8 斯普林菲尔德	560	410	760	170	590	280	370	210	0	250
9 卫奇塔	700	370	510	180	640	80	300	450	250	0

通过最近邻法的求解，最优路径为(5,1,2,6,4,0,7,8,3,9)，距离和为 3530；通过本文设计的 GA 算法求解，最优路径为(0,7,3,8,1,9,5,2,6,4)，距离和为 3500。比较发现，本文设计的 GA 算法更好，且 3500 确实是最优路径。

下面的表格记录的是种群规模为 50 时，传统 GA 方法和本文设计的 GA 方法的结果比较，程序运行了 10 次。

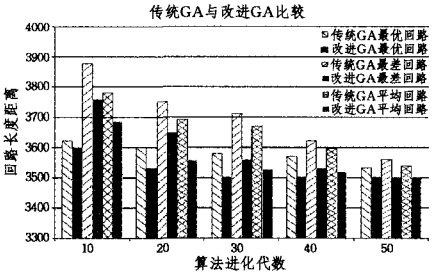

图 3　传统 GA 与自适应贪婪 GA 的比较

通过上面的实验数据，很容易发现本文设计的 GA 方法解决 TSP 是十分有效的。当进化代数变大时，本文 GA 算法的最优回路和最差回路长度都在变小，这说明算法是在朝着一个更优的方向前进，这也是 GA 方法所希望看到的。当算法进化到 50 代时，本文设计的 GA 方法已经找到了最优解，也就是回路长度为 3500 的那组解。

通过与传统 GA 方法的比较可以发现，本文设计的 GA 方法在执行 10 个城市的 TSP 时的表现要明显优于传统 GA 方法。当进化代数达到 50 时，本文的 GA 方法最优回路和最差回路已经趋同，而传统 GA 方法还未完全收敛，这说明本文设计的 GA 方法的收敛速度要快于传统 GA 方法，即当进化代数一定时本文的 GA 方法能更大可能地找到最优解；本文的 GA 方法还能更好地利用局部信息，并且能够很好地继承父代的优秀基因片段。

结束语　通过实验可以看出，本文提出的改进 GA 方法在求解 TSP 上相对于传统 GA 方法有着很大的优势。遗传算法是一种随机算法，但算法的每一步操作并不能被精确地控制。遗传算法能够计算的基础是适应度的确定，应该还能找到一些更好的确定适应度的方法。

参 考 文 献

[1] 陈国良. 遗传算法及其应用[D]. 北京：人民邮电出版社，1996

[2] 王桂平，王衍，任嘉辰. 图论算法理论、实现及应用[D]. 北京：北京大学出版社，2011

[3] 陈国良. 遗传算法及其应用[D]. 北京：人民邮电出版社，1996

[4] 玄光男，程润伟. 遗传算法与工程优化[D]. 北京：清华大学出版社，2004

[5] 王文杰，叶世伟. 人工智能原理及应用[M]. 北京：人民邮电出版社，2004

[6] 王小平，曹立明. 遗传算法：理论、应用与软件实现[M]. 西安：西安交通大学出版社，2002

[7] Chan K C C, Lee V, Leung H. Generating Fuzzy Rules for Target Tracking Using a Steady-State Genetic Algorithm[J]. IEEE Transactions on Evolutionary Computation, 1997, 1(3): 189-200

[8] Holland J H. Adaption in Natural and Artificial Systems[M]. The University of Michigan Press, Ann Arbor, 1975

(上接第 178 页)

个对考生有较高指导价值的高考志愿在线模拟填报系统。

参 考 文 献

[1] 张巧. 商务智能发展现状与趋势分析[J]. 中国证券期货，2009，12(02)：14-17

[2] 胡翠华，陈登科. 商务智能在我国的发展现状、问题及其对策[J]. 科技管理研究，2007(10)：50-52

[3] 殷员分，张自力，蔡海敏，等. 数据仓库与 OLAP 技术在高考志愿数据分析中的应用[J]. 计算机科学，2010(5)：162-164

[4] 蔡海敏，张自力，曾铮，等. 基于数据仓库与联机分析技术的高考加分政策评估[J]. 计算机科学，2010，37(6)：223-225

[5] 曾铮，张自力，殷员分，等. OLAP 技术在高考志愿填报方式分析评估中的应用[J]. 西南师范大学学报：自然科学版，2010，35(3)：239-242

[6] Han Jia-wei, Kamber M. 数据挖掘：概念与技术(第 2 版)[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007

[7] 吴远红. ETL 执行过程的优化研究[J]. 计算机科学，2007，34(1)：81-83

[8] Ralph K, Joe C. The Data WareHouse Toolkit: Practical Techniques for Extracting, Cleaning[M]. Wiley, 2004: 29-48

[9] Melome E. SQL Server 2005 Analysis Services 标准指南：中文版[M]. 武桂香，等译. 北京：电子工业出版社，2008：56-61

[10] George S, Sivakumar H. MDX 解决方案(第 2 版)[M]. 李仁见，董霖，等译. 北京：清华大学出版社，2008：10-18

[11] 吴亮奎. 寻找失落的基础：基础教育"基础"的反思[J]. 天津师范大学学报：基础教育版，2010(3)：1-4

[12] Larson B. Microsoft SQL Server 2005 商业智能实现[M]. 赵志恒，武海峰，译. 北京：清华大学出版社，2008：468-469

[13] MacLennan J, Crivat B, Tang Zhao-hui. Data mining with Microsoft SQL server 2008[M]. Indianapolis, IN: Wiley Pub., 2009

[14] 万永生. 高考加分政策的现状及思考[J]. 教学与管理，2009，36(10)：75-77

[15] 赵岩. 数据挖掘中的关联规则技术研究[D]. 西安：西安电子科技大学，2008

[16] Noyes B. Smart client deployment with ClickOnce: deploying Windows Forms applications with ClickOnce[M]. Upper Saddle River, NJ: Addison-Wesley, 2007