数据仓库与OLAP技术在高考志愿数据分析中的应用

殷员分 张自力 蔡海敏 曾 铮
(西南大学智能软件与软件工程重点实验室 重庆 400715)

摘 要 如何填报高考志愿,增加考生被自己心仪院校录取的几率,是每一位高考考生和家长密切关注的问题。以某省近9年积累的高考历史数据建立数据仓库,利用OLAP技术对这些数据进行多维分析,得到了一些广大考生可资借鉴的结果。重点介绍了考生志愿多维数据集的建立与分析的整个过程,以及涉及到的一些技术难点。

关键词 数据仓库,多维数据集,OLAP,高考志愿
中图法分类号 TP311 **文献标识码** A

Application of Data Warehouse and OLAP in Preference Data Analysis on National College Entrance Examination and Admission

YIN Yuan-fen ZHANG Zi-li CAI Hai-min ZENG Zheng
(Key Laboratory of Intelligent Software and Software Engineering, Southwest University, Chongqing 400715, China)

Abstract How to decide the preference to increase the probability of dream college matriculation is concerned by examinees and their families closely. By employing OLAP technology on the historical preference data, some interesting and referential results were mined. This paper was focused on how to build the related data warehouse as well as the related technical issues.

Keywords Data warehouse, Multidimensional datasets, OLAP, Preference in college entrance examination

志愿填报是高考过程中最重要的环节之一,也是考生与家长密切关注的问题。志愿填报是否恰当,直接关系到考生能否被心仪的高校录取。在高考志愿填报过程中,考生主要依据自己的分析以及家长和老师的经验[1]对前几年高考录取情况做单一的离散式分析。然而国家政策经常调整,高考形势不尽相同,考生、家长和老师很难凭借个人经验去了解全部的院校与专业情况,这些都给考生填报志愿带来很大的影响。

鉴于此,本文以某省招生自考办公室招生数据挖掘项目为依托,以该省自实行网络招生以来所积累的近9年海量电子招生数据(其中包含了大量有关考生基础数据、志愿填报数据、高校专业设置与院校录取结果等有用信息)为基础,首先建立数据仓库,然后根据主题建立多维数据集,最后利用OLAP技术从多角度、多侧面、多层次对考生志愿数据进行分析并用前端展现工具对其分析结果进行图形化展现,从而为考生填报志愿提供决策支持。

1 数据仓库总体结构

数据仓库(Data Warehouse)是支持管理决策过程的、面向主题的、集成的、随时间变化的持久的数据集合[2]。为了快速响应客户的请求,首先需要对关系数据库或其它外部数据源中的海量历史数据进行抽取、转换、清洗,再将加工后的数据装载到数据仓库中。

OLAP从数据仓库中抽取相关数据并建立多维数据集,使用户可以从多侧面、多角度、多层次地考察数据仓库中的数据,从而深入理解包含在数据中的信息与规律。下面以高考考生志愿数据为例,介绍数据仓库与OLAP技术在高考志愿数据分析中的应用,其数据仓库总体结构如图1所示。

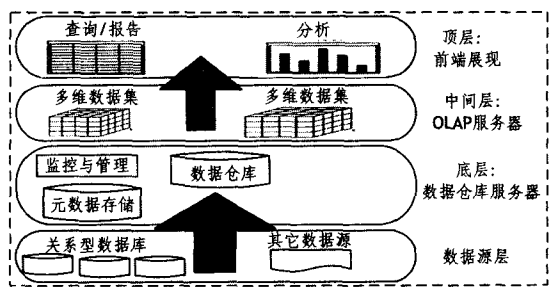

图1 数据仓库总体结构

1) 底层是数据仓库服务器,包括数据仓库、数据仓库管理和元数据管理。数据仓库负责存储分析、决策数据;而数据仓库管理则负责管理数据仓库;元数据管理负责对元数据进行管理。

2) 中间层是OLAP服务器:它是一种介于后端数据仓库服务器和前端展现工具之间的中间服务器,为用户提供来自数据仓库或数据集市的多维数据。

3) 顶层是前端展现工具:它是多维数据集的门面,包含

到稿日期:2009-07-25 返修日期:2009-09-30 本文受重庆市科技攻关计划项目(CSTC,2009AC2174)资助。
殷员分(1984-),男,硕士生,主要研究方向为商务智能与数据挖掘,E-mail:yuanfen200401@139.com;张自力(1964-),男,教授,硕士生导师,主要研究方向为多Agent系统、混合智能系统、数据挖掘;蔡海敏(1986-),男,硕士生,主要研究方向为商务智能与数据挖掘;曾 铮(1984-),女,硕士生,主要研究方向为商务智能与数据挖掘。

各种报表工具、查询工具和数据分析工具，以表格或图形化的手段对分析与挖掘结果进行展现。

2 OLAP 数据仓库建模

数据仓库的设计方法有两种[3]：自下而上和自上而下。自上而下的设计方法强调应用决定数据，有什么应用就获取什么数据，设计维度和多维数据集是项目的开始；自下而上的设计方法首先要依据系统文档与需求分析来创建数据仓库，然后根据不同的分析主题建立多维数据集。通过对某省市招生数据挖掘系统文档与该省招生自考办公室需求文档等进行需求分析，本文采用自下而上的设计方法，围绕考生志愿数据分析为主题来建立数据仓库。

2.1 OLAP 数据仓库建模

目前最流行的数据仓库多维数据模型有 3 种[3]：星型模式(star schema)、雪花模式(snowflake schema)和星系模式(galaxy schema)。本文采用星系模式。星系模式是多个事实表共享一个或多个维表的情况[3]。该数据模型将数据表分为两种：事实表与维度表。相对于星型模型与雪花模型，星系模式具有更高的数据准确率[4]。

数据仓库的模型设计是数据仓库设计的一个重要方面。本文以志愿报考热度与最低录取分为主题来组织数据，事实表是考生的志愿库与高校招生计划库，以招生年份、科类、计划性质、批次、考生、专业、院校、志愿序号、投档单位、成绩比较维与成绩过线维为维度表。

2.2 源数据

本数据仓库的数据来源于某省 2000－2008 年普通高考招生的历史数据，涉及到的数据主要有考生基本信息、院校基本信息、院校科类信息、院校批次信息、院校计划性质信息、专业信息、招生年份信息、历年批次分数线信息、考生志愿填报信息与院校招生计划投放信息。

2.3 数据的抽取、转换与装载

数据质量是分析结果可靠性的基础，而 ETL(数据抽取、转换、清洗和加载)是数据质量的重要保障[5]。为了确保数据仓库中的数据更加准确，本文首先对该省近 9 年的历史招生数据进行分析，形成一套标准的招生数据代码表，然后建立招生数据代码表清洗规范，最后采用 SSIS(Microsoft SQL Server 2008 Integration Services)逐年对普通高考招生录取系统中的招生数据代码表与事实表进行抽取、转换和清洗，并装载到目标数据仓库中。

2.4 技术难点及解决方法

数据仓库来源于数据库，其本身也是由数据库管理系统进行管理，但是它的结构、功能和设计与传统的数据库设计方法却不同。下面对建立高考志愿数据仓库过程中所遇到的主要问题做如下总结。

1) 分析业务数据与需求

业务数据是数据仓库的源数据，而对业务数据的了解是设计数据仓库的基础。只有通过对业务数据的分析，才可以清楚地知道原有数据库系统中已有什么，再结合用户的需求，才能明确当前数据仓库的设计能提供需求中的哪些，还缺少哪些。总之，通过对历史数据和需求的分析，可以明确用户正在使用的数据现状、他们如何使用这些数据以及将来利用这些数据干什么。

2) 事实表的粒度问题

数据越详细，粒度就越小，数据级别也就越小；数据综合度越高，粒度也就越大，级别也就越高。而事实粒度需求的不同，一方面将直接导致数据仓库模型设计的差异，另一方面将直接影响存储容量与分析效果，因此粒度的设计极其重要。由于本课题需求分析中，不仅需要对细节性数据进行分析，而且需要对综合数据进行分析，因此采用了多重数据粒度的设计方法。这也是针对业务量大、分析要求比较高的情况采用的最佳解决方法[3]。

3) 维度的缓慢变化问题

高考政策与高校名称时常有所变化，数据仓库的维表结构必然也会有所改变。既要保留历史信息，又要保留新增信息是维表设计非常棘手的问题。例如，本文中需要记录院校名称的变化以及考生字段的增加情况等。本文采用的解决方法是创建额外字段来记录这些信息之间的关系，在具有缓慢变化维表中加入“创建时间”与“更新时间”字段，这样就不需要创建额外的数据行，也不需要改变维表中的键值结构。

4) 维表 ETL

ETL 是开发数据仓库项目最重要的环节之一，其困难程度占用了一个数据仓库项目的 50%～70%，而维表处理占了数据仓库 ETL 解决方案所需开发时间的最大一部分，是其最复杂的组成部分[6]。文中维表 ETL 最大的困难在于每年数据代码表不尽相同，比如院校有合并、更名情况，院校专业有变动，高考加分政策变化以及该省招生管理人员程序编写风格不同等等，导致相同代码表中同一名称在不同的年份对应的代码不同，同一代码在不同的年份表示不同的含义，也就是说代码与名称含义的对应关系在不同的年份不具有通用性。另外，代码表中极其广泛地存在用以表示同一含义所使用的名称不同的现象，因此要对数据仓库进行 ETL，首先必须对 9 年的数据代码表进行统一、规范，然后建立数据维表清洗规范，最后按照清洗规范逐年对历史招生数据进行 ETL。在 ETL 过程中，本文主要使用 SSIS 具有的模糊查找、模糊分组、数据转换、维度转换、查找、脚本组件、派生列、OLE DB 命令、条件性拆分等数据流组件依据对业务的了解进行组件组合以解决以上问题。

3 构建多维数据集

设计好数据仓库，并且将业务数据经过 ETL 以后，需要根据不同的主题建立相应的多维数据集[7]，在志愿分析数据仓库中有以下子主题：

1) 院校/专业报考热度分析：通过对考生的志愿填报行为与高校招生计划投放进行分析，得出院校或院校某专业的报考热度，从而为考生选报高校与专业提供有力参考，也可以为院校调整专业结构提供有力依据。

2) 最低录取分分析：高校所在地域是考生填报志愿时需要着重考虑的重要因素之一[8]，然而相同专业在不同的地域往往具有不同的最低录取分。本文从多个角度分析同一专业在不同地域的最低录取分，从而为考生填报志愿时根据自身情况选择合理的地域或省份提供有力参考。

本文利用 SSAS(Microsoft SQL Server 2008 Analysis Services)建立多维数据集。首先建立数据源与数据源视图，其次建立多维数据集模型，定义维度与事实度量，建立多维数

据集。志愿数据分析多维数据集的数据模型如图 2 所示。

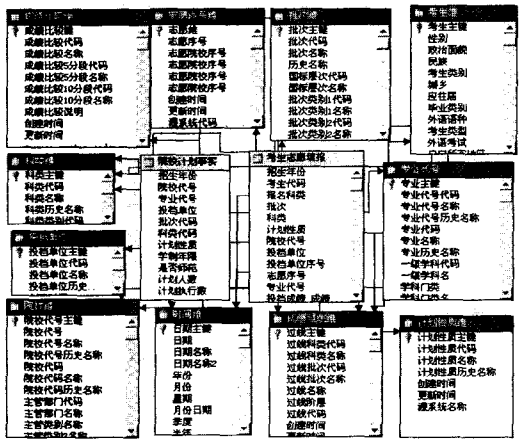

图 2　多维数据集模型(截图)

模型中,表头颜色为蓝色的是维度表,有招生年份维、考生维、院校维、批次维、科类维、计划性质维、地域维、志愿序号维、专业维、成绩比较维、成绩过线维;表头颜色为黄色的是事实表,有高校招生计划表与考生志愿填报表。多维数据集的度量值有招生计划人数、考生志愿填报数、最低录取分、报考热度。

其中,报考热度与最低录取分均值用一种访问 Analysis Services 多维数据集语言 MDX[9] 来定义:

[报考热度]=[Measures].[志愿填报计数]/[Measures].[计划人数]

多维数据集的存储方式有 3 种:ROLAP,MOLAP 和 HOLAP[10]。本文采用 MOLAP 方式存储,它将细节数值和合计数值都存放在立方体中,可将多维视图直接映射到数据立方体数组结构。这种存储方式的优点是可以对数据立方体的数据快速索引[11]。

多维数据集的操作主要有切片(slice)、切块(dice),钻取(roll-up 和 drill-down)以及旋转(pivot)[12]。通过多维数据集的多种操作,可以从多个角度分析志愿数据,图 3 可用来分析某高校各个专业在各个志愿的报考热度情况。

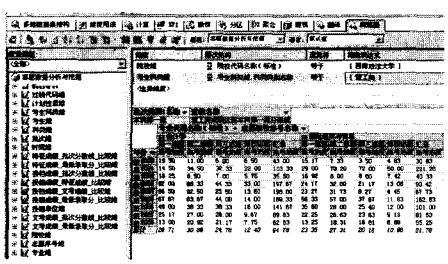

图 3　多维数据集的操作(截图)

通过对特定维度的筛选并对多维数据集进行相应的操作,得到相应的报考热度情况。图 3 选取院校、科类、批次、过线、招生年份维度等对部分维度值进行筛选并将志愿序号下钻得到某高校各个专业在该省理工类本科第一批的第一、第二、第三、第四志愿的报考热度情况。

4　志愿数据前端展现与结果分析

商业智能的前端产品负责直接面向用户,将用户的请求转发给服务器层、数据层,同时也向用户展现所需信息。本文采用 SSRS(Microsoft SQL Server 2008 Reporting Services)以图形的方式展现志愿数据分析结果。

4.1　院校报考热度分析

某院校或专业是否热门是考生填报志愿时必须考虑的重要因素之一。本文以志愿序号、科类、批次、时间、院校、专业、是否师范、考生、投档单位、成绩比较和成绩过线为维度,以报考热度为度量值,进行院校与专业报考热度分析。考生可以从以上 11 个维度查看院校或专业的报考热度,以为其志愿填报提供极大的便利。图 4 为该省理工类本科第一批考生报考某校部分专业的报考热度情况。

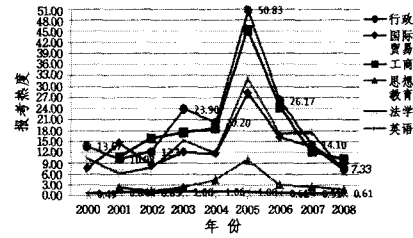

图 4　某高校部分专业第一志愿报考热度(截图)

由图 4 可知,该校的行政管理专业报考热度最高可达 50.83,工商管理专业、英语专业与国际经济与贸易专业也保持着较高的报考热度,而思想政治教育专业尤其是法学专业的报考热度相对要低很多,第一志愿往往不能完成招生计划。这说明如果高考成绩一般的考生在第二志愿甚至第三志愿报考该校的行政管理或工商管理专业,与报考该校的法学专业相比,第二志愿的作用性会更小,即面临落榜的几率会更高;反之,考生在志愿填报时,如果把该校的法学专业作为第二志愿,甚至高考成绩不太理想的考生第一志愿报考该校法学专业,则录取的几率更大。

由此不难得出,一方面高考考生需慎重地选报第一志愿。历年来的录取结果显示,在高校的录取结果中第一志愿录取率高达 85%[13],如果第一志愿报考热度极高的高校或专业,比如图 4 中的行政管理专业,则应该考虑自身的实际情况,切忌一味地报考热门专业;另一方面,考生志愿填报时应该注意志愿的梯度。选好第一志愿固然重要,但考生不可轻视第二、第三志愿的选报,只要第二、第三志愿所报考的学校与专业的档次较第一志愿有一定程度上的拉开,那么一旦第一志愿落选,第二志愿的作用就可以发挥出来,加大了录取几率。

4.2　最低录取分分析

院校所在地域也是考生填报志愿时必须考虑的问题,如何在理想地域与自身高考成绩之间找到平衡点,一直是考生很头疼的问题。本文所建立的志愿数据分析多维数据集可以从志愿序号、科类、批次、时间、专业、是否师范、考生、投档单位、成绩比较与成绩过线 10 个维度来分析同一专业在不同地域的最低录取分。图 5 可用来分析该省文史类本科第一批考生第一志愿报考中国语言文学类专业在各个地域的最低录取分情况。

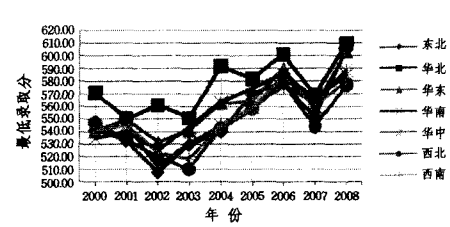

图 5　最低录取分分析(截图)

(下转第 177 页)

即 $P1>P2>P3$，敌机最可能选择“相持”策略。

如果态势因素在各计划中的权重系数未知，则由多个领域专家给出各因素的权重系数，进行加权平均，得到各态势因素在各计划中的权重系数，如表 3 所列。

表 3　专家给定态势因素在各计划中权重系数

		距离	能见度	释放电磁干扰	释放导弹
专家 1	相持(P11)	0.32	0.31	0.28	0.09
	规避(P12)	0.48	0.19	0.21	0.12
	攻击(P13)	0.28	0.12	0.31	0.29
专家 2	相持(P21)	0.30	0.33	0.26	0.11
	规避(P22)	0.45	0.22	0.19	0.24
	攻击(P23)	0.29	0.10	0.31	0.30
专家 3	相持(P31)	0.26	0.3	0.35	0.11
	规避(P32)	0.47	0.28	0.20	0.05
	攻击(P33)	0.30	0.06	0.32	0.32

各专家权重为 0.3，0.3，0.4，则加权结果如表 4 所列。

表 4　加权平均后态势因素在各计划中的权重系数

	距离	能见度	释放电磁干扰	释放导弹
相持(P1)	0.29	0.312	0.302	0.104
规避(P2)	0.467	0.226	0.20	0.128
攻击(P3)	0.291	0.09	0.314	0.305

其他计算步骤同实例，最终得到得分矩阵为

$$\begin{pmatrix}0.397-0.381\\0.366-0.441\\0.307-0.489\end{pmatrix}=\begin{pmatrix}0.016\\-0.075\\-0.182\end{pmatrix}$$

即 $P1>P2>P3$，敌机采取“相持”策略。

结束语　本文在态势评估领域将直觉模糊理论与计划识别相结合，建立了基于直觉模糊多属性决策的计划识别模型，克服了计划识别中信息在传播中“易”丢失的缺点。实验表明，此模型计算量小，能直观地反应出采取某计划的倾向度。

参 考 文 献

[1] Hall D L, Llinas J. An Introduction to Multisensor Data Fusion [J]. Proceedings of IEEE, 1997, 85(1): 6-3

[2] ENDSLEYMR. Toward a Theory of Situation Awareness in Dynamic Systems[J]. Human Factors, 1995, 37(1): 3264

[3] Das S, Lawless D. Trustworthy Situation Assessment via Belief Networks[C] // Proceedings of the Fifth International Conference on Information Fusion. 2002, 1: 543-549

[4] Rao B S, Durrant-Whyte H A. Decentralized Bayesian Alogorithm for Identification of Tracked Targets[J]. IEEE Transaction on Systems, Man and Cybernetics, 1993, 23(6): 1683-1698

[5] 雷英杰，王宝树. 基于直觉模糊决策的战场态势评估方法[J]. 电子学报，2006，34(12)：2175-2179

[6] 李伟生，王三民，王宝树. 基于计划识别的态势估计方法研究[J]. 电子与信息学报，2006，28(3)：532-536

[7] 雷英杰，王宝树，路艳丽. 基于直觉模糊逻辑的近似推理方法[J]. 控制与决策，2006，3：306-310

[8] Zadeh L A. Fuzzy Sets [J]. Inf orm ation and Control, 1965, 8(3): 338-353

[9] Atanassov K. Intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems, 1986, 20: 87-96

[10] Xu Z S. Intuitionistic fuzzy aggregation operators [J]. IEEE Transactions on Fuzzy Systems, 2007, 15: 1179-1187

[11] 王晓帆，王宝树. 基于贝叶斯网络和直觉模糊推理的态势估计方法[J]. 系统工程与电子技术，2009，31(11)：2742-2746

(上接第 164 页)

由图 5 可知，该省文史类本科第一批考生第一志愿报考中国语言文学类专业在各地域保持着较稳定的趋势。其中，华北地区的最低录取分一直高于其他地域，而西北地区几乎一直处于最低位置，二者在 2002 年分数之差高达 53 分；另外，2006 与 2008 年各地域的最低录取分都有所升高，而 2007 年各地域的最低录取分都有所下降。

因此，倘若考生在填报志愿时第二志愿为华北高校，则落榜的风险较大。反之，如果考生的第一志愿为华北地区高校，而第二志愿为西北地区高校或者直接避开录取分数较高的华北、华东地区而选择西北地区同类高校，可以扩大自己的选择范围，降低风险，提高录取几率。总之，考生填报志愿时需要正确处理地域问题。

结束语　本文利用某省招生自考办公室丰富的数据资源，将普通高考招生录取系统的海量数据经过抽取、转换、清洗预处理后装载到 OLAP 数据仓库中，并建立志愿数据分析多维数据集，最后对这些数据进行多维分析并将结果以图形的方式进行展现。本文的分析结果对高考考生填报志愿具有指导意义。下一步的工作重点是借助数据挖掘技术，利用已建立的数据仓库或多维数据集，进一步揭示出隐含在志愿数据中的规律和信息，从而为考生的志愿填报进行合理导向。

参 考 文 献

[1] 李德铭. 高考志愿填报问题及其对策[J]. 甘肃教育，2007，3(05)：7-8

[2] Inmon W H. 数据仓库(原书第三版)[M]. 王志海，林友芳，等译. 北京：机械工业出版社，2003

[3] 朱德利. SQL Server 2005 数据挖掘与商业智能完全解决方案[M]. 北京：电子工业出版社，2007：189-191

[4] Vincent R. Building a Data Warehouse: With Examples in SQL Server [M]. New York: Springer, 2008: 6-7

[5] Brian K, Erik V. SQL Server 2005 Integration Services 专家教程[M]. 冯飞，译. 北京：清华大学出版社，2008：20-29

[6] 张宁，贾自艳，史忠植. 数据仓库中 ETL 技术的研究[J]. 计算机工程与应用，2002，70(24)：213-216

[7] Brian L. Microsoft SQL Server 2005 商业智能实现[M]. 赵志恒，武海锋，等译. 北京：清华大学出版社，2008：135-139

[8] 高永金. 走进象牙塔——与广大考生谈如何填报好高考志愿[J]. 中学生数理化：高中版，2005(05)：34-35

[9] 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

[10] Melome E, et al. 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

[11] 柳进，胡政，唐降龙. OLAP 数据仓库在电网调度决策中的研究与应用[J]. 计算机工程与设计，2005，26(2)296-311

[12] Han Jiawei, Micheline K. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2001：39-41

[13] 陈祖力. 填报志愿有技巧[J]. 广东教育，2005，10(3)：67-68

[14] 李红. 智能决策支持系统的发展现状及应用展望[J]. 重庆工学院：自然科学版，2009，23(10)：140-144