

基于数据仓库与联机分析技术的高考加分政策评估

蔡海敏 张自力 曾 铮 殷员分

(西南大学智能软件与软件工程重点实验室 重庆 400715)

摘 要 高考是关乎国计民生的一件大事,而高考中的加分政策,更是各省市招生考试机构和考生乃至整个社会密切关注的问题。高考加分政策是否合理,长期以来没有科学、系统的评估方法。针对这一问题,依据某省多年来高考录取工作积累的数据,通过运用数据仓库及联机分析技术,对高考加分政策的相关数据进行多维分析,得到一些有趣的结果,由此制定了一套评估加分政策合理性的方法,可供各省市招生考试机构在出台新的高考加分政策时参考。

关键词 高考加分政策,数据仓库,OLAP

中图法分类号 TP311 **文献标识码** A

Evaluation of the Scoring Policies in National Admission and Examination of Colleges and Universities Using Data Warehouse and OLAP

CAI Hai-min ZHANG Zi-li ZENG Zheng YIN Yuan-fen

(Key Laboratory of Intelligent Software and Software Engineering, Southwest University, Chongqing 400715, China)

Abstract National Admission and Examination of Colleges and Universities (NAECU) gets in closer and closer touch with the national economy and the people's livelihood. The Scoring Policies in NAECU are becoming one of key issues of the Provinces and Cities Admission Offices (PCAO), and attract the attention of every examinee as well as the whole society. Currently there is no systematic and scientific way to evaluate the Scoring Policies. To this end, data warehouse and OLAP technology were employed to analyze the data accumulated in the past years in a province. By multi-dimensional analysis of the data, some interesting results were found and presented in the paper. Based on these findings, the evaluation scheme for Scoring Policies in NAECU was then given. This scheme can provide decision making support when introducing new scoring policies.

Keywords Scoring policies in NAECU, Data warehouse, OLAP

高考是我国教育体系中的一个重要环节,对高校人才培养具有基础性作用,对基础教育具有导向性功能,对社会具有重要的影响作用;其公平与公正从根本上保证了高考制度的权威性。高考加分政策作为一项公共政策,它的每一微小的变化都关系着成千上万考生的利益^[1],是社会各界关注的焦点,因为它不仅体现着高考公平的价值导向,是社会公平在高考中的映射,而且体现着素质教育的思想^[2]。

A 省份从 2000 年开始网上招生录取工作,积累近 10 年的招生数据,这些招生数据包含大量与高考加分政策相关的数据。高考加分政策是否合理,长期以来没有科学、系统的评估方法。如何有效利用这些加分政策数据,分析高考加分政策的现状与存在的问题,以及指导高考加分政策的调整 and 改革方向,并制定一套科学、系统、有效的评估高考加分政策合理性的方案,是招生决策管理人员面临的一个难题。

本文结合历年高考加分政策数据的分析研究,利用数据仓库和 OLAP 技术,建立高考加分政策分析系统的数据仓库和多维数据模型。通过对多维分析模型的分析,展现出高

考加分政策的现状,并制定出一套有效的评估高考加分政策的合理性的方案。

1 OLAP 数据仓库总体设计

数据仓库(Data Warehouse)是一个面向主题的、集成的、相对稳定的、反映事实变化的数据集合,用于支持管理决策^[3]。OLAP 从数据仓库中提取数据并建立多维数据集,用户可以利用多维数据集多角度、多层次地观察数据仓库中的数据,从而选取有价值的信息。高考加分政策的 OLAP 数据仓库总体结构如图 1 所示。

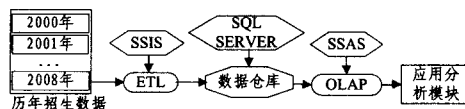


图 1 数据仓库总体结构

1) 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等。数据源层中的关系型数据库等数据通过 ETL 工具进行抽取、转换和清洗,并最终装载到数据仓库中。数据仓

到稿日期:2010-01-29 返修日期:2010-03-18 本文受重庆市科技攻关计划项目(CSTC,2009AC2174)资助。

蔡海敏(1986—),男,硕士生,主要研究方向为商务智能与数据挖掘,E-mail:jet47@126.com;张自力(1964—),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;曾 铮(1984—),女,硕士生,主要研究方向为商务智能与数据挖掘;殷员分(1984—),硕士生,主要研究方向为商务智能与数据挖掘。

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)OLAP 服务器:它介于底层的数据仓库服务器和顶层的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)前端展现工具:它包含各种报表工具、查询工具和数据分析工具,以表格或图形化等方式对分析结果进行展现。

2 高考加分政策数据仓库的建立

2.1 原始高考加分政策数据分析

从 2000 年开始,A 省份实行网络招生以来,积累了近 10 年的招生数据,这些招生数据包含了大量与高考加分政策相关的有用信息。通过对高考网上招生系统的数据库系统的研究,得到与高考加分政策相关的主要源数据表有 5 个:1)报名库表;2)加分政策代码表;3)成绩库表;4)投档单日志表;5)录取库表。

2.2 数据仓库主题的确定

数据仓库中的数据是面向主题组织的。主题是在较高层次上将企业信息系统中的数据进行综合、归类和分析利用的一个抽象概念,每一个主题基本对应一个宏观的分析领域^[4]。针对需求分析,根据得到的分析型业务需求,结合应用系统及其数据的调研与数据分析的结果,按照高考招生数据库的特点,高考加分政策分析的主题可以分为考生加分政策主题、考生成绩主题、考生志愿填报主题、考生志愿投档主题、考生志愿录取主题,如图 2 所示。

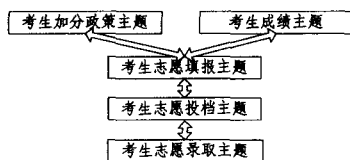


图 2 高考加分政策分析主题

2.3 设计数据仓库逻辑模型和物理模型

数据仓库的主题确定后,就需要设计逻辑模型。目前流行的数据仓库逻辑模型有 3 种:星型模式(Star Schema)、雪花模式(Snowflake Schema)和星系模式(Galaxy Schema)。本文采用星型模式,它是最常用的数据仓库设计结构的实现模式^[5]。在数据仓库的逻辑结构中,数据表可以划分为两类:一类是事实数据表(简称为“事实表”),用来存储数据仓库的实际数据,如高考加分政策数据的考生加分政策即是一个事实表;另一个是维度数据表(简称为“维度表”),用来存储数据仓库的维度数据,如考生、科类、批次、院校、时间等分析角度均为维度表。

事实表是数据仓库的核心,也是数据仓库中最大的表。事实表包含了考生享受加分政策的基本情况等信息,是对加分政策进行分析的素材。事实表的设计包括对事实的选择、量度的构造、粒度的设计和聚合的设计等。

对维度表的观察是商业智能的基本驱动力。通过维的切换,人们可以从不同的角度观察客观世界。基于不同的维度,可以看到各量度的汇总情况,也可以同时从多个不同的维度进行交叉分析。该数据仓库设计中,主要有以下 12 个维度表:1)日期维度;2)考生维度;3)科类维度;4)批次维度;5)计划性质维度;6)院校维度;7)专业维度;8)专业受限维度;9)投

档单位维度;10)志愿序号维度;11)成绩比较维度;12)成绩过线维度。图 3 是考生加分政策主题模型。

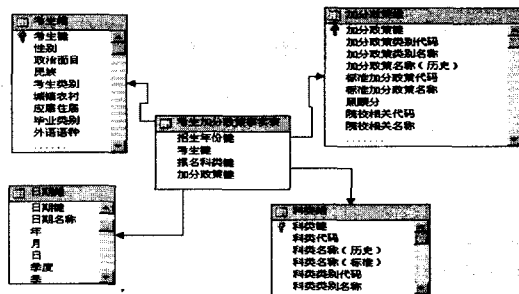


图 3 考生加分政策主题模型(截图)

在确定数据仓库逻辑模型的事实表和维度表之后,就需要确定物理模型。数据仓库的物理模型就是数据仓库逻辑模型在物理系统中的实现模式,包括了逻辑模型中各种实体表的具体化,例如表的数据结构类型、索引策略、数据存放位置和数据存储分配以及物理模型的优化操作等^[6]。

完成数据仓库的逻辑模型和物理模型的设计后,就可以创建数据仓库。数据仓库也是一种数据库,因此在逻辑结构设计完成之后可以跟普通的数据库一样创建、修改和删除。

2.4 数据抽取、转换和加载

完成数据仓库的设计后,就需要通过 ETL 工具往数据仓库中装载数据。

ETL,即数据抽取(Extract)、转换(Transform)、装载(Load)的过程,是负责完成数据从源数据向目标数据仓库转化的过程,是实施数据仓库的重要步骤,是构建数据仓库的重要一环^[7]。用户从数据源抽取所需的数据,经过数据清洗,最终按照预先定义好的数据仓库模型,将数据加载到数据仓库中。

高考网上招生系统只是在每年特定的一段时间运行,每个年度的高考网上招生系统的数据库结构基本相同,但不同年度的数据完全不一样。这就导致每年的数据编码都会发生一些变化,经过多年的变化,编码的规范问题急需解决。

编码过程碰到的主要问题包括:

- 1)历年的命名不规范,即同一内容在不在年份用不同的表达方式,如“湖南理工”和“湖南理工大学”;
- 2)更名问题,例如“陕西工学院”改名为“陕西理工学院”;
- 3)合并问题,例如,“西南师范大学”和“西南农业大学”合并成“西南大学”。既要完成代码的规范,又要保留历史变化信息,就引进了渐变维度^[8]的方法。具体的实现方法如下:

创建额外字段来记录这些信息之间的关系,在渐变维表中加入“创建时间”与“更新时间”字段,再加上“原始名称”和“标准名称”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,即可解决上述问题。

ETL 贯穿于整个解决方案的全过程,管理数据的质量,并且对整个系统进行数据处理与调度。ETL 涉及到大量的业务逻辑和异构环境,在一般的数据库转换项目中 ETL 部分成为了牵扯精力最多的部分,故而 ETL 规则设计和实施约占在整个项目中工作量的 60%~80%,这是从众多实践中得到的普遍共识^[9]。

2.5 建立 OLAP 多维数据集

对数据的多维分析并不是主要针对数据仓库的,而是针

对从数据仓库提取的子集^[10],如多维数据集。将源数据加载到设计好的数据仓库后,就要根据高考加分政策主题来设计多维数据集,在高考加分政策分析系统中,建立的分析主题主要包括:

1)享受加分政策的考生分析。通过对享受加分政策的考生人数的分析,能够得到享受加分政策的考生人数占所有报名考生人数的比率,即反映出加分政策的考生享受率。该享受率说明了加分政策在总体上的分布现状,如图4所示。

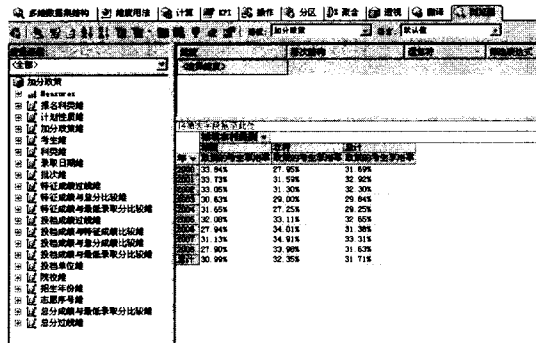


图4 享受加分政策的考生分析主题(截图)

2)利用加分政策上线的考生分析。此类考生指的是考生的总分成绩达不到某批次的分数线,但通过考生所享受的所有加分政策的加分后使考生的志愿成绩达到该批次的分数线,从而使得考生有资格参加对应批次的投档,考生因此多了一个投档批次的机会。此分析以批次为角度,反映出加分政策对“成绩临近批次分数线的考生”的影响程度,从这一侧面来评估加分政策的合理性。

3)利用加分政策录取的考生分析。此类考生指的是考生的总分成绩达不到其填报志愿院校的最低录取分数线,通过考生所享受的所有加分政策的加分后使考生的投档成绩达到该最低录取分数线,从而使得该考生被院校录取。所有加分政策不管在上述两个分析中的表现如何“抢眼”,只有在本主题分析中的考生才是高考加分政策的真正有效受益者。利用加分政策录取的考生人数的高低才是高考加分政策制定人员最为关心的,这可以用来检验加分政策的调整 and 改革的实践效果,是评估加分政策是否合理的最有效标准,可以为高考加分政策的调整 and 改革提供决策支持。

3 高考加分政策的实例分析和评估方案

本节首先给出评估高考加分政策合理性的评估方案。然后,根据实例分析的结果来说明评估方案的有效性。

3.1 高考加分政策合理性的评估方案

评估方案的第一组指数是享受加分政策的考生人数和加分政策享受率。加分政策享受率是指享受加分政策的考生人数占报名考生人数的比率。享受加分政策的考生人数、加分政策享受率可以反映出整体的加分政策、部分的加分政策的覆盖范围,可为决策者为高考加分政策的整体规划提供指导作用。

评估方案的第二组指数是利用加分政策录取的考生人数和加分政策录取利用率。加分政策录取利用率是利用加分政策录取的考生人数占享受加分政策的考生人数的比率。利用加分政策录取考生人数和加分政策录取利用率是反映加分政策的真正“效用性”,是对加分政策的合理性评估指数,这组指

数也是最有解释意义的评估标准。

下面将通过对两个实例结果的分析来探讨评估方案。在实例中,分析结果采用 SQL SERVER 2008 的 SSRS(Microsoft SQL Server Report Services)进行前端展示,并利用表格和图形等形式将分析的结果直观地呈现给最终用户,使用户更容易理解。

3.2 加分政策享受率分析

通过构造的享受加分政策的考生分析主题,分析从2000年到2008年的文理科考生的加分政策享受率情况,如图5所示。借助变化图可以看出,2000年到2008年的加分政策享受率的值总体维持在0.30到0.34之间,这说明了A省份的招生部门在制定加分政策时,在加分政策的覆盖范围有着整体的规划。再把考生分成城镇考生和农村考生两类,就可以发现两者的加分政策享受率的曲线图变成完全相反的两个方向。城镇考生的加分政策享受率呈现下降态势,而农村考生的加分政策享受率却显现出增长态势。这正好反映出该省招生部门制定的加分政策逐步向农村考生倾斜,并且取得了良好的效果。

从本实例的分析可以看出,加分政策享受率对高考加分政策的整体规划有很好的指导作用。

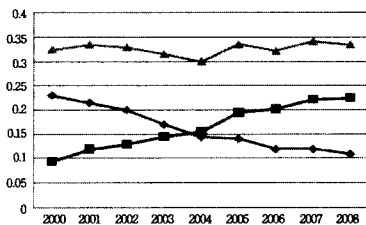


图5 加分政策享受率(截图)

3.3 加分政策录取利用率分析

通过构造的考生加分政策主题和考生志愿录取主题,分析文理科考生从2000年到2008年利用“集居少数民族”加分政策被本科批次的院校录取的人数及加分政策录取利用率情况,如图6和图7所示。从图6中的曲线图可以看出,享受加分政策的考生人数的增长速度是非常快的,特别是从2003年到2007年。那么这么快的增长速度有没有拉动利用加分政策录取考生人数的快速增长呢?从图中利用加分政策录取考生人数的曲线图的变化可以看出:从2000年到2004年,其值并没有随着加分政策享受人数的高增长而有相应的高增长,其增长量是非常缓慢的;但从2005年到2007年,每年的增长量都在400个左右,增长非常明显。

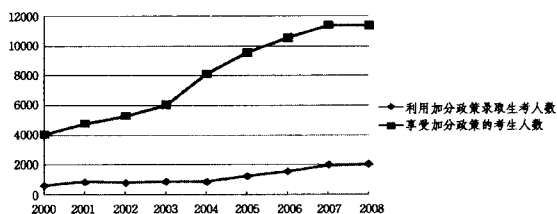


图6 “集居少数民族”利用加分政策录取考生人数(截图)

我们分析另一张图表。图7表示的是文理科考生本科批次的“集居少数民族”加分政策的录取利用率情况。借助变化图可以看出,从2002年到2006年,加分政策利用率都不理想,其值基本都在0.15以下;但从2007年之后,加分政策利

出的方法比起其他算法都有所提高,MP 平均提高 21.4%(从 2.4%~44.6%),MR 平均提高 7.5%(从 5.9%~10.6%),F-measure 平均提高 15.7%(从 4.8%~31%)。整体说来,SME 算法具有较好的综合性能。

结束语 本文提出一种新的本体映射发现方法 SME。作为一种实例压缩学习方法,SME 从训练数据集中获得能最大程度覆盖训练数据的数据依赖球的集合(压缩集),通过压缩集中数据依赖球在测试集上的交操作来判断测试数据的类别(是否存在本体映射),其泛化误差边界依赖于压缩集的规模^[11]。怎样降低泛化误差,提高映射发现的准确性,是今后进一步的研究方向。

参考文献

- [1] Doan A, Madhavan J, Domingos P, et al. Learning to map between ontologies on the semantic Web[C]//Proc. of 11th World Wide Web Conference. 2002;662-673
- [2] Lacher M S, Groh G. Facilitating the exchange explicit knowledge through ontology mappings[C]//Proc. of 14th Intl. Florida Artificial Intelligence Research Society Conf. 2001;305-309
- [3] Ehrig M, Sure Y. Ontology mapping - an integrated approach [M]. Bussler C, Davies J, Fensel D, et al., eds. Lecture Notes in Computer Science. Vol. 30653, Springer-Verlag, 2004;76-91
- [4] Melnik S, Molina-Garcia H, Ralm E. Similarity flooding: a versatile graph matching algorithm[C]//Proc. of the 18th International Conference on Data Engineering (ICDE 2002). San Jose, California; IEEE Press, 2002;117-128
- [5] Madhavan B J P, Rahm E. Generic schema matching with cupid [C]//Proc. of VLDB. 2001;49-58

- [6] Noy N F, Musen M A. PROMPT: Algorithm and tool for automated ontology merging and alignment[C]//Proceedings of the 2000 National Conference on Artificial Intelligence. Austin, Texas, 2000;450-455
- [7] Hu Wei, Cheng Gong, Zheng Dongdong, et al. The results of falcon-ao in the oaei2006 compaign[C]//Proc. of the ISWC 2006 Workshop on Ontology Matching. Athens, GA, USA, November 2006
- [8] Tang Jie, Li Juanzi, Liang Bangyong, et al. Using Bayesian decision for ontology mapping[J]. Web Semantics: Science, Service and Agents on the World-Wide Web, 2006,4(12)
- [9] 权光日. 集合覆盖问题的启发函数算法[J]. 软件学报, 1998,9(2)
- [10] Chvátal V. A greedy heuristic for the set covering problem[J]. Mathematics of Operations Research, 1979,4;233-235
- [11] Marchand M, Shawe-Taylor J. The Set Covering Machine[J]. Journal of Machine Learning Research, 2002,3;723-746
- [12] Valiant L G. A theory of the learnable [J]. Communications of the Association of Computing Machinery, 1984,27(11);1134-1142
- [13] Haussler D. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework [J]. Artificial Intelligence, 1998,36;177-221
- [14] Haussler D. Convolution kernels on discrete structures [R]. USCS-CRL-99-10. Computer Science Department, University of California in Santa Cruz, July 1999
- [15] <http://oaei.ontologymatching.org/2007/>
- [16] Rodriguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies[J]. IEEE Trans on Knowledge and Data Engineering, 2003,15(2);442-456

(上接第 225 页)

用率的值一下跃至 0.18 左右,显然这时的加分政策利用率值是让人非常满意的。让我们来解读这组数据。“集居少数民族”的加分政策享受人数的高增长没有带动加分政策录取利用人数的高增长,体现在图 7 中的 2002 年到 2006 年的加分政策录取利用率一直维持在低位。显然,招生部门发现了这一问题,故而在 2007 年开始调整了该项加分政策的加分规则:由原先的“考生报考重点本科院校享受 10 分的加分、报考一般本科院校享受 20 分的加分”调整为“考生报考所有院校都享受 20 分的加分”。很显然,该调整方案取得了很好的效果,加分政策的录取利用率重新回到高位值。

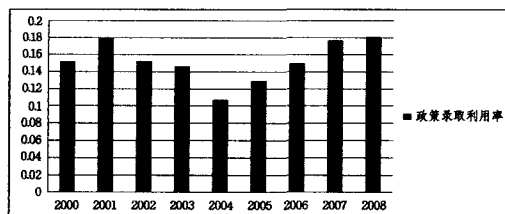


图 7 “集居少数民族”加分政策录取利用率(截图)

通过以上的分析,利用加分政策录取利用率来评估一项加分政策是否合理的有效性是显而易见的。

3.4 评估方案小结

加分政策享受率是用来评估高考加分政策的整体规划合理性的,而加分政策录取利用率则是用来评估单个加分政策的合理性的有效手段。

结束语 本文的研究意义在于将数据仓库与 OLAP 技术应用于高考招生工作的加分政策。通过建立数据仓库和多

维数据集,为高考加分政策分析提供准确的信息,更为后续的数据挖掘等分析提供准确、更具有针对性的数据。本文同时提供一套评估高考加分政策的合理性的方案,可为决策人员制定和调整高考加分政策提供依据。下一步的工作重点除完善数据仓库和 OLAP 模型,还将引入数据挖掘技术来分析高考加分政策的潜在规律等。

参考文献

- [1] 向莉娟,孟立军.教育公平视阈下我国高考加分政策探析[J].内蒙古师范大学学报:教育科学版,2008,21(06):1-3
- [2] 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
- [3] Inmon W H. 数据仓库[M].北京:机械工业出版社,2003;20-24
- [4] 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;77-80
- [5] Vincent R. Building a Data Warehouse: With Examples in SQL Server[M]. New York: Springer, 2008;71-72
- [6] 陈文伟,黄金才.数据仓库与数据挖掘[M].北京:人民邮电出版社,2004;54-55
- [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, Margy R. The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling[M]. Wiley, 2002;89-95
- [9] Ralph K, Joe C. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning[M]. Wiley, 2004;29-48
- [10] 陈京民. 数据仓库与数据挖掘技术[M].北京:电子工业出版社,2007;93-100
- [11] 李红良. 智能决策支持系统的发展现状及应用展望[J]. 重庆工学院学报:自然科学版,2009,23(10):140-144