

基于关系代数运算的数据统计与分析 ——以场景教学系统数据库为例

唐娟, 潘志安

(湖北职业技术学院, 湖北 孝感 432000)

摘要: 关系代数语言是用传统的集合运算和专门的关系运算来表达查询的抽象语言, 提供了一系列的关系代数操作, 通过这种数学运行来指导数据库在关系操作上的程序实现。场景教学系统中每一个场景对应一个实际应用, 场景中设置了问题, 本文以系统数据库为例, 给出了基于关系代数运算的数据统计分析过程, 采用抽取样本法计算期望学习时长和学习时长标准差, 并设计了一个测算学习者投入度的公式, 对评价场景教学具有指导意义。

关键词: 场景教学; 数据库; 关系代数运算

中图分类号: TP391 **文献标识码:** A

Statistics and Analysis of Data Based on Relational Algebra Operations —A Case Study of the Scene Teaching System Database

TANG Juan, PAN Zhian

(Hubei Polytechnic Institute, Xiaogan 432000, China)

Abstract: The relational algebra language is an abstract language that expresses queries by traditional set operations and special relational operations. It provides a series of relational algebraic operations, which can guide the implementation of relational operations in database. Every scene in the scene teaching system corresponds to a practical application. Questions are set in the scene. Taking the database of the system as an example, this paper presents the statistical analysis process of data based on relational algebraic operation, calculates the standard deviation of expected learning time and learning time by sampling method, and designs a formula to measure learner's engagement, which has directional significance for evaluating the scene teaching.

Keywords: scene teaching; database; relational algebra operations

1 引言(Introduction)

1970年, E.F.Codd提出关系数据库理论基础, 关系数据库应用数学方法来处理数据库中的数据, 20世纪80年代后成为最重要、最流行的数据库系统。

关系代数语言是用传统的集合运算和专门的关系运算来表达查询的抽象语言, 提供了一系列的关系代数操作: 并、差、笛卡尔积、选择、投影和更名等基本操作, 交、连接和关系除等扩展操作。关系代数操作以一个或多个关系为输入, 结果是有一个新的关系, 具有一定过程性, 是学习其他数据库语言, 如SQL等的基础。关系代数运算主要运算符如图1所示^[1]。

运算符		含义	运算符		含义
集合 运算符	$\cup$	并	比较 运算符	$>$	大于
	$-$	差		$\geq$	大于等于
	$\cap$	交		$<$	小于
	$\times$	广义笛卡尔积		$\leq$	小于等于
专门的 关系 运算符	σ	选择	逻辑 运算符	$=$	等于
	π	投影		$\neq$	不等于
	$\bowtie$	连接		$\neg$	非
	$\div$	除		$\wedge$	与
				$\vee$	或

图1 关系代数运算符

Fig.1 Relational algebraic operator

场景教学系统搭设在云平台上, 每一个场景对应一个实际应用, 教师和学生可以通过网络选择其中一个场景进行交互式学习。场景教学理念为“做中学, 学中做”, 学习内容镶嵌在具体的操作流程中, 场景中设置的问题, 为一个实际操作, 完成场景内的所有操作任务, 就解决了一个实际问题^[2]。本文以场景教学系统数据库为例, 基于关系代数运算, 对数据进行统计与分析。

2 系统数据库表结构(Table structure of system database)

数据库系统是现代计算机系统的一个重要组成部分, 现代的管理信息系统几乎都是以数据库作为核心的^[3]。场景教学系统数据库的库表结构如下:

用户表(用户ID 用户名 昵称 性别 出生日期 专业学历)

课程表(课程ID 课程名 课程性质)

场景表(场景ID 场景名 授课教师 视频数 问题数 课程ID)

视频表(视频ID 名称 时长 视频内容 存储路径 场景ID)

问题表(问题ID 问题名 答案 场景ID)

答题记录表(用户ID 问题ID 答题用时正确否)

学习记录表(用户ID 场景ID 进入场景日期时间 退出场景日期时间)

3 数据统计(Data statistics)

基于云平台的场景教学, 覆盖面广、学习者需求复杂, 应该注重学习测评结果的统计分析和学习测评的质量分析^[4]。针对设置的每个问题, 给出平均正确率、平均答题用时。针对场景, 给出学习次数、平均学习时长、问题的提交率和平均正确率, 针对整门课程也能给出以上相应的参数。这些参数可以供系统设计者挑选使用, 如学习者学完一个场景, 需要生成一个学习报告, 学习报告中可以罗列出本人的学习时长及答题正确率、场景的平均正确率及平均答题用时等, 以便于学习者比对和参考。

下面列举出其中几个参数的关系代数运算过程。

(1) 每个问题平均正确率:

对于ID为001的问题平均正确率为

$$\text{count}(\sigma_{\text{问题ID}='001' \wedge \text{正确否}='正确'}(\text{答题记录表})) / \text{count}(\sigma_{\text{问题ID}='001'}(\text{答题记录表}))$$

(2) 场景平均正确率:

场景ID为01的所有问题列表S为

$$S = \pi_{\text{问题ID}}(\sigma_{\text{场景ID}='01'}(\text{问题表})) \bowtie \text{答题记录表}$$

对于S中每个问题(问题ID用i表示), 平均正确率W为

$$W = \text{count}(\sigma_{\text{问题ID}=i \wedge \text{正确否}='正确'}(S)) / \text{count}(\sigma_{\text{问题ID}=i}(S))$$

场景平均正确率为

$$(\sum_i W) / \text{count}(S)$$

(3) 每个问题平均答题用时

对于ID为001的问题, 平均答题用时为

$$\text{sum}(\pi_{\text{答题用时}}(\sigma_{\text{问题ID}='001'}(\text{答题记录表}))) / \text{count}(\sigma_{\text{问题ID}='001'}(\text{答题记录表}))$$

(4) 每个场景学习次数

对于ID为01的场景, 学习次数C为

$$C = \text{count}(\sigma_{\text{场景ID}='01'}(\text{学习记录表}))$$

(5) 场景下问题的平均提交率

场景ID为01的所有问题列表S为

$$S = \pi_{\text{问题ID}}(\sigma_{\text{场景ID}='01'}(\text{问题表})) \bowtie \text{答题记录表}$$

对于S中每个问题(问题ID用i表示), 学习了该场景的用户平均提交率M为

$$M = \text{count}(\sigma_{\text{ID}=i}(\text{答题记录表})) / \text{count}(\delta(\pi_{\text{用户ID}}(\sigma_{\text{场景ID}='01'}(\text{学习记录表}))))$$

场景ID为01的所有问题的平均提交率为

$$(\sum_i W) / \pi_{\text{问题数}}(\sigma_{\text{场景ID}='01'}(\text{场景表}))$$

4 数据分析(Data analysis)

学习分析研究正从关注大数据转向有意义数据的探索。如何抓取得到学习活动信息流中的关键行为数据并解析出学习者的个性学习行为特征是今后研究中要解决的一个问题^[5]。故基于场景教学数据库表, 需要抽取合适的样本进行分析, 以此来判断不同场景的教学效果。

4.1 期望学习时长

数学期望(mean)(或均值, 亦简称期望)是试验中每次可能结果的概率乘以其结果的总和, 是最基本的数学特征之一^[6]。它反映随机变量平均取值的大小。一个场景的期望学习时长应该是学习者正常进行学习时的学习时长平均值。针对答题记录表中某个场景所有记录测算期望学习时长, 并不合理, 因为有的用户可能只是匆匆浏览后离开场景页面, 也有的用户可能是停留在该页面长时间未进行学习活动, 这些非正常记录会干扰正常学习时长的判定, 考虑这两种情况会导致学习时长过短或过长, 可以初步定一个学习时长的上下限, 在此范围内筛选出样本表。设场景01的学习时长样本表为R1:

$$R1 = \sigma_{\text{场景ID}='01' \wedge \text{退出场景日期时间}-\text{进入场景日期时间} \leq \text{设定时长上限} \wedge \text{退出场景日期时间}-\text{进入场景日期时间} \geq \text{设定时长下限}}(\text{答题记录表})$$

为方便数据统计, 将样本R1投影为样本T1, T1中仅有属性“学习时长”。

$$T1 = \rho_{\text{学习时长}/\text{退出场景日期时间}-\text{进入场景日期时间}}(\pi_{\text{退出场景日期时间}-\text{进入场景日期时间}}(R1))$$

再根据T1计算期望学习时长。将T1中每条记录的学习时长看成是一个独立的值, 其发生的概率均等, 故可以直接求平均值作为期望学习时长, 场景01的期望学习时长E可以通过公式(1)计算。

$$E_{\text{场景01}} = \text{sum}(\pi_{\text{学习时长}}(T1)) / \text{count}(T1) \quad (1)$$

4.2 学习时长标准差

标准差表示的就是样本数据的离散程度, 通常是相对于

样本数据的平均值而定的,用 $M \pm SD$ 来表示,表示样本某个数据观察值相距平均值有多远^[7]。

μ 为均值,计算标准差理论公式:

$$\sigma(\mu) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

通过学习时长标准差,可以判断不同学习者对一个场景的学习时长是否聚集在期望学习时长附近。数据库系统中一般有相应函数计算标准差,以mysql为例,标准差函数为STD,则场景01学习时长的标准差可以通过公式(2)计算。

$$\sigma_{\text{场景01}} = \text{STD}(\pi_{\text{学习时长}}(T1)) \quad (2)$$

现有两个场景01和02,分别取样本T1和T2,两个样本的学习时长发生散点图如图2所示。

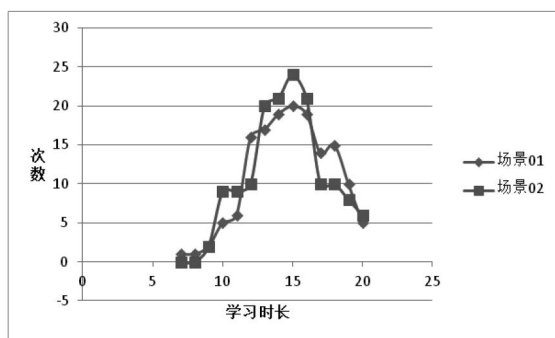


图2 学习时长发生散点图

Fig.2 Scatter plot of learning duration

用公式1和公式2计算期望时长和标准差,得到数值:

$$E_{\text{场景01}} = 15.007 \quad \sigma_{\text{场景01}} = 2.846$$

$$E_{\text{场景02}} = 14.66 \quad \sigma_{\text{场景02}} = 2.647$$

标准差越小,表明数据越聚集;标准差越大,表明数据越离散。从图2中可以看到,场景01和场景02比较,场景02学习时长的聚集度比场景01略高,经过计算标准差值较小。

4.3 学习者投入度

将文中给出的统计和分析参数进行组合和加工,可以从不同的维度分析场景教学效果。如为了比较不同场景的教学内容是否能吸引学习者,可以测算不同场景的学习者投入度。学习者投入度至少应从两个参数进行分析:

一是场景的学习时长标准差。学习时长标准差越小,则学习时长聚集度高,这表明学习该场景时学习者的学习状态越一致和越稳定。

二是场景问题提交率,其计算方法前文已述,场景教学理念是课中边学边做,问题设置在课中,问题是否提交能很好地判断学习者是否在参与学习,问题提交率高表明学习者的参与度高。

场景的学习时长标准差和问题提交率两参数呈跷板式,一个要求越低越好,另一个要求越高越好,跷板斜度越大,

说明该场景的教学吸引力越强,学习者的投入度越高。

假设一个场景学习时长方差为 $\sigma_{\text{场景ID}}$, $\pm 1/\sigma_{\text{场景ID}}$ 的学习者为学习投入者,同时场景中问题提交者也在这部分人群中, $M_{\text{场景ID}}$ 为问题提交率,则可以设计一个测算学习者投入度att的公式(3)。

$$\text{att} = 2/\sigma_{\text{场景ID}} \times M_{\text{场景ID}} \quad (3)$$

存在以下两组数据:

$$\text{场景01: } \{ \sigma_{\text{场景01}} = 2.846 \quad M_{\text{场景01}} = 75\% \}$$

$$\text{场景02: } \{ \sigma_{\text{场景02}} = 2.647 \quad M_{\text{场景02}} = 79\% \}$$

$$\text{通过公式3计算得: } \text{att}_{\text{场景01}} = 52.71\% \quad \text{att}_{\text{场景02}} = 59.69\%$$

计算结果表明,场景02的学习者投入度比场景01略高。

教学系统中可以设定一个标准,如某场景学习者投入度低于40%,则应着重分析该场景教学的过程和内容,做出相应的诊改。

5 结论(Conclusion)

本文以场景教学系统数据库为例,基于关系代数运算给出了数据统计和分析过程,通过这种数学运行来指导数据库在关系操作上的程序实现,熟悉关系代数的操作就很容易写出复杂的SQL语句。本研究采用抽取样本法计算期望学习时长和学习时长标准差,设计了一个测算场景学习者投入度的公式,对评价场景教学具有指导意义。当然,分析场景教学效果应从多维度进行,如答题效果、访问量、学习者评价等等,这些维度本文没有探讨,为后续研究和其他教学评测人员提供研究思路。

参考文献(References)

- [1] 何玉洁.数据库原理与应用(第三版)[M].北京:机械工业出版社,2017(6):31-33.
- [2] 唐娟.面向软件技术专业的云平台场景教学应用研究[J].软件工程,2018,21(05):60-62.
- [3] 苏俊.数据库基础教程[M].北京:中国人民大学出版社,2014:1.
- [4] 何晓新,陶水龙,孙志娟.论远程开放教育环境下工科课程的学习测评模型[J].广播电视大学学报(哲学社会科学版),2008(04):117-121.
- [5] 牟智佳,武法提.MOOC学习结果预测指标探索与学习群体特征分析[J].现代远程教育研究,2017(03):58-65.
- [6] 陈希孺.概率论与数理统计[M].中国科学技术大学出版社,1992:25.
- [7] 刘忠广.标准差、标准误差和估计标准误差辨析[J].消费导刊,2018(11):203-204.

作者简介:

唐娟(1976-),女,硕士,实验师,讲师.研究领域:软件工程,实践教学.

潘志安(1967-),男,硕士,教授.研究领域:计算机网络.