**智能出题系统**

随着互联网的飞速发展，人们越来越多的行为在网络中发生，这直接导致互联网中人类行为相关数据呈爆 炸式增长，人类在不知不觉中已经进入了一个 “大数据”时 代。“数据驱动教学，分析变革教育”的大数据时代已经来临，利用教育数据挖掘技术和学习分析技术，构建学习教学相关模型，探索学习变量之间的相关关系，为教育教学决策提供精准、高效支持已成为一种趋势。

智能出题系统基于教育数据挖掘，综合运用数学统计、机器学习和数据挖掘的技术和方法，对海量题库大数据进行处理和分析，通过数据建模，发现学习者学习结果与学习内容、学习资源和学习行为习惯等变量的相关关系，来精准推荐相关的学习资源。

**1 教育数据挖掘的相关技术**

目前教育领域内常用的主要数据挖掘技术方法包括分类预测、聚类分析、关联分析和可视化技术等。

**分类预测**

　　分类属于预测分类（离散、无序的）标号，是一种有监督的学习过程，它依据训练数据集发现准确描述来划分类别，其实质是对已知的训练数据集表现出来的特性，获得每个类别的描述或属性来构造相应的分类器或者分类；而预测则是建立连续值函数模型，是根据分类和回归来预测未来的趋势与规律。决策树、贝叶斯（Bayes）、人工神经网络（ANN，Artificial Neural Networks）、k-近邻（kNN，k-Nearest Neighbors）、支持向量机（SVM，SupportVectorMachine）、基于关联规则的分类和集成学习（EnsembleLearning）等是比较常见的分类算法，其中决策树算法和贝叶斯分类算法中的朴素贝叶斯（Naive Bayes）算法是利用率较高的分类方法；通常运用比较多的预测方法则是回归分析法，包括线性回归、多元回归和非线性回归。

**聚类分析**

　　聚类是将大量的数据聚集到不同的群或者簇的一个过程，使得相同簇中的对象极其相似，不同簇之间的对象则存在较大的相异性。聚类分析不同于分类，不必按照已有的分类标准进行分类，而是依据大量的样本数据将其自动划分为未知类，是一种探索性的、无监督的学习过程。聚类分析过程运用的方法不同，往往会得到不同的分类结果。目前常常使用的聚类算法一般分为五类：划分聚类算法（如k-Means、k-中心点）、层次聚类算法（如BIRCH算法、ROCK算法和CURE算法）、密度聚类算法（如DBSCAN算法和DENCLUE算法）、网格聚类算法（如STING算法和CLIQUE算法）和模型聚类算法（如EM算法）。

**关联分析**

　　关联分析是从海量数据中发现对象集合或者项目集合之间存在的频繁模式、关联和相关关系，通过利用数据库事务之间相互关联、相互依赖的联系与规律，对这些事务进行预测。数据库事务之间的关联规则挖掘过程一般分为两步：一是需要发现数据库中所有的频繁项集，二是由频繁项集来挖掘数据库事务之间的关联规则。典型的关联规则挖掘算法包括了由RakeshAgrawal提出的Apriori算法和JiaweiHan等提出的频繁模式增长（FP-growth）算法。

**可视化技术**

　　可视化技术是数据挖掘中一种必不可少的且应用广泛的辅助技术，它通过借助图形、图表、图像等手段以用户易理解的方式形象地展示数据分析的最终结果，因此又被称为“绘制大数据的藏宝图”。可视化技术的独特之处表现在其持有的艺术性与交互性的特征上，不仅能够呈现多种艺术形式的效果图，而且实现了用户与数据之间的双向信息沟通。数据可视化运行的一般流程为获取、分析、过滤、挖掘、表示、修饰、交互七个步骤。在解释数据信息过程中可视化技术比较注重将抽象的、复杂的、模糊的数据转化为具体的、直观地、清晰的以视觉形式表达的信息与知识，以方便用户对可视化结果的掌握与理解。可视化技术应用的最终结果很好地揭示了隐藏在大数据背后的潜在价值。

**2 教育数据挖掘技术在智能题库中的应用**

2．1分类在智能题库中的应用

分类算法中应用得最多的是决策树算法。决策树是以实例为基础的归纳学习算法。决策树主要用于对离散数据进行分类，在题目筛选方面则能够用来对学生科目学习情况评估、知识点掌握程度分析、薄弱环节预测等方面提供依据。利用ID3分类算法对学生的历史综合测评数据进行分析，从而构造综合素质评分的分类器，以改进综合测评的方法，促进学生德、智、体各方面的全面发展。在在线学习中，分类可用来锁定那些学习热情不高者并找出补救措施，还可预测哪些课程将会取得不错的成绩。决策树作为分类算法中的一种，也是使用频率最多、范围最广的数据挖掘技术之一，广泛地应用于各个领域之中

2．2 聚类在智能题库中的应用

“物以类聚，人以群分”，在自然科学和社会科学中，存在着大量的分类问题。根据学生的基本信息、学习历史、学习偏好、知识结构等已有信息 ,挖掘学生特征 ,通过聚类，将具有共同学习习惯、共同学习兴趣、共同薄弱知识点等的学生划分成不同的学生群组，这样将便于题库系统采用不同的推送策略，根据学生群组特点推送学习资源，实施个性化学习与指导。同时，教师也可以根据班级学生的特点布置不同的练习作业，查缺补漏，重点强化。这样大大提高了学习与教学效率，也避免了低质或者无效学习资源的过度推荐。例如，某个同学对于某些方面的知识点已经掌握了很牢固，但是由于没有对学生群组进行聚类，普遍推送相同的学习资源，可能导致该学生用户的体验很差，重复练习已经掌握的知识点，降低学习兴趣与热情。

2．3关联分析在智能题库中的应用

关联规则是使用频率最多，范围最广的数据挖掘技术之一，也可以被广泛地应用于国内外教育教学的决策分析中。运用关联规则的方法可以判定哪些题目或者知识点适合某类学生，从而精准的进行题目的筛选及推荐，使得因材施教能够在实践中得到更进一步的应用和实施。利用数据挖掘中的关联分析，寻找学生各种学习活动之间的内在联系，从而干预和引导相关的学习行为及学习路线。通过题库中存放的有关作答结果的历史数据，利用关联规则的挖掘算法，找出学生的薄弱知识点环节，以便科学、合理地进行题目推荐。而且从大量的题库作答数据中，可挖掘出学习效果与哪些因素有关，从而为下次题目推荐提供决策支持信息 。

2.4 可视化技术在智能题库中的应用

图表是大量数据的一种简洁直观的展示方式，能方便人们在大数据分析时快速的获取有效信息，然后综合判断作出决策。根据学生用户的登录频率、资源点击时间、页面停留时间、题目作答时间、题目作答正确率、重做次数等绘制不同的曲线，能更直观的了解学生用户的行为习惯、学习特点、知识掌握程度等重要信息。结合数据挖掘的关联分析与时间序列分析等相关技术 ,就能从这些海量数据中快速挖掘出有用的信息 ,帮助分析这些数据之间的相关性、回归性等性质 ,得出一些具有价值的规则和信息 ,最终找到影响学生成绩的原因。在此基础上 ,题库对课程设置做出合理安排。也可根据学生得分情况分析出每道题的难易度、区分度、相关度等指标 ,教师也能够对试题的质量作出比较准确的评估 ,进而可以用来检查自己的教学情况及学生的掌握情况 ,并为今后的试卷测验以及作业布置提供更好的参考依据。

**用户学习行为跟踪监测**

为了更好的了解用户学习以及使用习惯、掌握用户行为，更精准、高效地推荐学习资源，引导优化学生的学习路线提供大数据支撑，需要对用户的学习行为进行跟踪监测。主要包括用户登录频率、页面停留时间、界面按钮交互次数、作答时间等。本功能模块的实现主要依靠接入TalkingData SDK。

**1 TalkingData**

TalkingData 成立于2011年，是国内领先的第三方数据智能服务商。借助以SmartDP为核心的数据智能应用生态为企业赋能，帮助企业逐步实现以数据为驱动力的数字化转型。TalkingData 成立以来秉承“数据改变企业决策，数据改善人类生活”的愿景，逐步成长为中国领先的数据智能服务商。以开放共赢为基础，TalkingData凭借领先的数据智能产品、服务与解决方案，致力于为客户创造价值，成为客户的“成效合作伙伴”，帮助现代企业实现数据驱动转型，加速各行业的数字化进程，利用数据产生的智能改变人类对世界以及对自身的认知，并最终实现对人类生活的改善。TalkingData不仅专注于数据智能应用的研发和实践积累，同时也在积极推动大数据行业的技术演进。

早在2011年成立初始，TalkingData就组建了数据科学团队，将机器学习等人工智能技术引入海量数据的处理、加工流程中。通过几年来的不断发展，TalkingData已在大数据、人工智能领域拥有多项国家专利。此外TalkingData还开源了大规模机器学习算法库Fregata、UI组件库iView、地理信息可视化框架inMap等项目，在海内外得到广泛支持与认可，使用者和贡献者遍布全球。目前TalkingData设立了包括硅谷边缘计算实验室、人本实验室在内的多个大数据、人工智能实验室，并与MIT媒体实验室、斯坦福人工智能实验室、加州理工航天技术实验室等国际顶尖学府、研究机构展开合作，共同加速大数据、人工智能相关技术的探索和演进，并将国际前沿技术引入高速发展的中国市场，与国内丰富的应用场景相结合，驱动新技术的落地应用与行业的飞跃发展。

**数据概览**

应用概览

应用概览展示了所选应用的累积设备、近30日可疑设备、近7日活跃设备及近30日活跃设备，以及单设备日均使用时长等统计值，从宏观上评估一款应用的表现。



时段分布

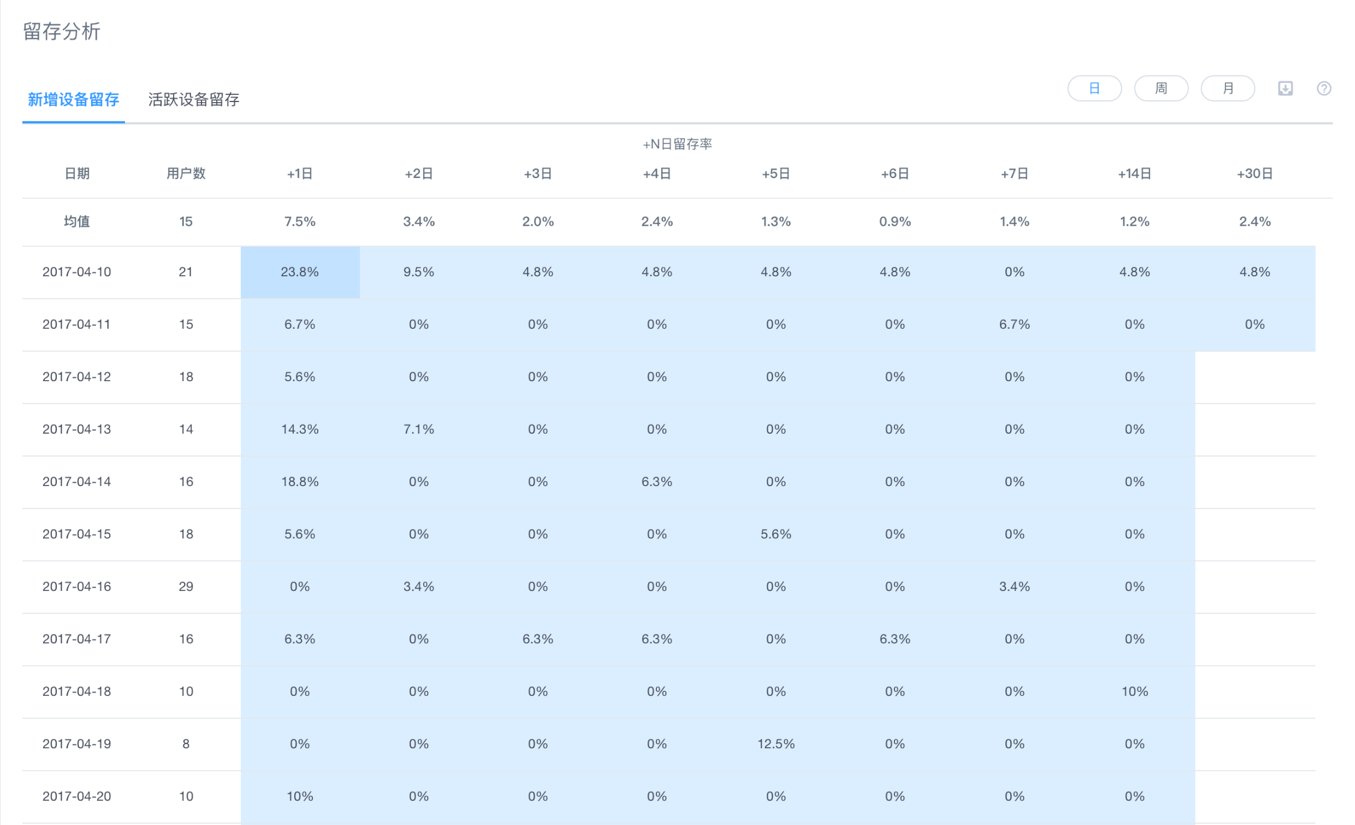
时段分布展示今日和昨日的分小时段数据量，每小时为一个区间，这条曲线可以反应用户在一天中新增、活跃的高峰期和低谷期。还可以在时段分析中通过查询一个单日来查看其他日的时段统计。

留存分析

留存可以用作评价用户粘性，是非常重要的一个指标。包括新增留存与活跃留存两个评估角度。

新增留存是以分析粒度，日，周，月，为时间区间，以时间区间内的首个时间区间的新增用户数为基数，取各个时间区间的留存新增用户数，从而得出各个时间区间的留存率，用户留存率体现了应用的质量和保留用户的能力。

活跃留存是以分析粒度，日，周，月为时间区间，以时间区间内的首个时间区间的活跃用户数为基数，取各个时间区间的留存活跃用户数，从而得出各个时间区间的留存率，通过这个比值可以了解用户的整体活跃度，但随着时间周期的加长，用户活跃率总是在逐渐下降的，所以经过一个长生命周期（3个月或半年）的沉淀，用户的活跃率还能稳定保持到5%-10%，则是一个非常好的用户活跃的表现，当然也不能完全套用，需要结合应用特点具体分析。



用户趋势

App Analytics平台允许自选时间段来查看任意时期的数据趋势，不光可以选定一个较长的日期段来看一定时期的长期走势，还可以单独选定某一日来看一天的具体数据，也可根据"版本"，"渠道"筛选查看。数据趋势部分详细展示了选定时段内应用用户和用户使用类指标的每日表现：新增设备，日活，周活，月活，启动次数，单次启动时长。



场景分析

通过提供整体场景分布和新增首日访问场景分布两个分析，对比用户常用的场景和用户在新增当日体验的场景，便于找到更合理的学习场景。



**用户行为分析**

页面访问

在正确调用页面统计接口后，系统可以统计相应页面的页面受访次数（PV）、页面受访人数（UV）、平均停留时长，及当前页面的离开率。通过记录页面的上一个页面来源，可以统计用户在访问某个页面后，下一步将访问哪些页面及占比。



事件分析

事件数据是指为了获取某些更精准的数据，需要定义一些事件继而进行的统计，我们可以看到各个事件在一段时间完成的情况，这样更能精确的了解用户使用情况。比如我们定义练习按钮点击事件，可以了解用户练习的频率；定义学科选择事件，可以了解用户是否有偏科现象，也可以统计出来最受欢迎的学科。通过事件趋势我们可以了解到所选事件在一段时间内的趋势走向，以及可以了解到活跃用户对这些事件的完成情况，同时我们可以看出应用启动时这些事件的完成情况。

