基于 CM-SPAM 算法的在线学习模式的研究

摘要：本文基于浙江大学海宁国际校区 blackboard 平台上课程的访问历史数据，通过 CM-SPAM 算法来分析学生在课堂中的访问课件内容的顺序行为。本研究主要针对目前主流的 Lag Sequential Analysis 算法在行为序列分析维度的不足，而提出新的分析建议。研究结果显示，采用 CM-SPAM 算法在线行为分析能够挖掘更加丰富的用户行为特征。该文期望此研究方案能够为教育机构中知识结构逻辑性较强的课程提供课程内容优化的思路以及建议。

【关键词】在线学习；在线参与模式；行为序列；序列模式挖掘；CM-SPAM

一、 引言

随着我国高等教育的快速发展，高校信息化建设已经从早期的基础建设逐渐步入教育 2.0 时代，在教育 2.0 时代，更多的强调如何利用信息化或者信息技术的手段来提高教学的质量以及教学活动的安排。近年来，在线学习（E-learning）平台在高等教育机构中日益普及，平台借助于互联网技术高速发展，摆摊了教学活动中时间和空间对于师生活动开展的限制。作为对于传统的教学模式的创新，越来越多的一线教师利用在线学习平台进行翻转课程的教学模式的改革。随着在线学习以及翻转课题的快速流行，在线学习平台记录下师生的访问课程记录，在一定程度充分表达了学生的学习行为。这些学习行为，对于研究学生在学习过程中遇到的难点以及在课程中的思维模式，提供了参考依据[1] [2]。

在数据挖掘中，特别是零售行业，对于这种基于用户购买行为进行了大量的研究并且取得了剧目的成效，特别是序列模式挖掘(sequential pattern mining)。序列模式挖掘的核心则是找出序列数据库中所有超过最小支持度阈值的序列模式。该研究方向是数据挖掘中非常重要的一个研究领域，通过分析用户的购买商品的循序，对于线下的商品促销，商品货架摆放，商品采购等提供数据依据以及决策推荐。

在教育行业中，相关学者通过利用序列模式挖掘来分析学生在访问平台中的循序行为，得出学生在学习过程中的学习行为特征，特别是在课程空间中的活动特征。针对学生学习的序列行为进行深度的挖掘，提出教学过程的优化建议。

二、 当前研究现状

目前主流的学者在分析访问顺序方面采用的是 Lag Sequential Analysis，LSA[3] [4] [5]。LAS 算法在分析学生基于在课程中的行为顺序访问习惯表现出良好的特征，并且取得一定的成效。但是在特定情况下，该算法存在一定程度的缺陷。

表一 用户访问行为序列

用户 ID	周一	周二	周三
用户 A	内容 A,内容 B，内容 C	内容 D	内容 E
用户 B	内容 A	内容 B，内容 C，内容 D	内容 E
用户 C	内容 A，内容 C	内容 D	内容 E

以表一数据为例，用户 A 在访问课程的内容时候，访问的顺序可以分为，周一访问内容 A，内容 B，内容 C，周二访问内容 D，周三访问内容 E。用户 B 在访问课程的内容时候，访问的顺序可以分为，周一访问内容 A，周二访问内容 B，内容 C，内容 D，周三访问内容 E。在采用经典 Lag Sequential Analysis（LSA）算法分析的时候，2 个用户的访问顺序都是(内容 A，内容 B，内容 C，内容 D)。以内容 B 内容 C 为例，尽管能反映出内容 B,内容 C 的顺序关系，但是没办法反映内容 B，内容 C 之间相对较强的关联关系。具体来说，内容 B 和内容 C 都是在同一天被访问的。在该样本集中，说明只要某个用户访问内容 B，当天该用户一定回访问内容 C。再以内容 C 和内容 D 为对比，可以看到访问内容 C 后，当天不一定就会访问内容 D。该场景在学习的挖掘场景中可能表现出明显不同的行为特征，比如说学生在访问课

件第二周的内容后，当天就查看第三周的课件。说明第二周和第三周课件内容相关非常高。如果学生在访问第二周课件后，基本都是第二天或者第三天后再次查看第三周的课件，尽管在这之间没有查看其它课件内容，但该行为特征不能说明第二周和第三周课件关联紧密。

第二，相关文献提出的行为序列特征的拓扑结构生产存在一定的缺陷。在形成行为序列的拓扑结构中，基本都是基于残差表（Z 分数）中的数值，一旦行为拓扑图中存在循环结构，即导致无法找到行为序列的终止点。还是以表一的数据为例，假设内容 B，内容 C 存在较强的序列关联；内容 A，内容 C 存在较强的序列关联；如果后面发现内容 A,内容 B 存在较强的序列关联（同时内容 B,内容 A 也存在强关联，相反顺序）。按照 Z 分表，可能生成的序列候选长度是无限制的。例如，1）内容 A 内容 B 内容 A 内容 C，2）内容 A 内容 B 内容 A 内容 B 内容 C，3）内容 A 内容 B 内容 A 内容 B 内容 A 内容 C。因为 Z 很表只能得出相邻 2 个内容之间存在强关联，但是基于这样邻居关系无法找到序列长度的终止标识。 在实际中，有显著关联的序列候选长度肯定是有限的。甚至说，在序列集合中，可能存在行为序列，内容 A 内容 B 内容 A 内容 C，占总集合 90%，内容 A 内容 B 内容 A 内容 B 内容 C 占集合的 10%。该现象说明大部分人在访问内容 B，然后再次访问内容 A，最后访问内容 C。而不是访问内容 B 后直接访问内容 C。不同的序列表现，对于分析用户的访问行为特征的影响明显是不同的。

相关文献提出通过利用汇聚算法，对 Z 分表的数值进行汇聚，生成访问序列。但利用聚类方式形成的汇聚图不一定能完整反映行为序列，因为聚类分析中，单个行为序列必须被划分到某个具体的序列聚类。在这样的情况下，相对高关联的隐藏行为序列就可能无法显示。假设存在内容 A 内容 B(AB)，内容 B 内容 C(BC)，内容 C 内容 D(CD)，其关联分别为 0.9，1,0.6。其中内容 A 内容 B 的 0.9 代表访问内容 A 后，90%的可能性用户去访问内容 B，访问内容 B 的用户会 100%访问内容 C，依次类推。可以看到内容 A 到内容 D 的关联度为 0.54，即 0.9*1*0.6。假设存在，内容 E 内容 B(EB)，内容 C 内容 F(CF)的关联分别为 0.8，0.8，再考虑之前内容 B 内容 C（BC）的关联为 1，可以看到访问内容 E 后访问内容 F 的可能性为 0.8*1*0.8，但是经过汇聚算法实现的关联度分析中，由于内容 B 内容 C(BC)被划分到其他聚类中，顾无法建立内容 E 内容 F(EF)的间接关联，然而通过 LSA 算法直接得到的关联可能远低于该数值。

针对这样现状，相关学者提出了序列模式挖掘用来挖掘该场景，在挖掘的过程中，采用 AprioriAll 算法来分析学生的访问顺序的关系[6]。基于 AprioriAll 思想的行为序列算法可以很好的解决以上情况。首先 AprioriAll 算法是对 Apriori 算法的一种补充。Apriori 算法中可以快速的找到物品之间的关联，但是该算法无法体现物品之间的顺序关联。因此 AprioriAll 在 Apriori 算法的基础上，实现了行为序列的挖掘。因为是基于 Apriori 算法而衍生的行为序列挖掘。该算法更加强调访问集合的顺序关系，对隐藏序列可以直接发现。对于行为序列的长度问题，可以通过调整算法中的参数，来指定挖掘的序列的最低长度。

PrefixSpan，SPAM 等算法则是基于 AprioriAll 在计算中消耗大量的计算资源而提出的性能改进算法。研究结果显示，GSP，以及 PrefixSpan，在特定的情况下，更加容易发现高支持度的关联条目数量。

AprioriAll、GSP、PrefixSpan 等在分析行为顺序上能较好的揭露行为之间的顺序关联，相关学者提出了 SPAM，FAST 算法进一步改善计算序列模式挖掘所需的计算资源以及存储资源。近年来,相关学者提出的 CM-SPAM 算法在实验中显示出极高的计算效率以及较少的内存消耗，使得大规模的数据序列模式挖掘变为可能[7]。

三、 研究样本以及方法

研究以浙江大学国际校区 2018-2019 年 blackboard 平台中《Integrative Biomedical Sciences》课程中，67 位学生在课程中的访问日志，共计 56790 条，作为研究范围。学生的访问日志包括学生的访问课程中的课件内容，访问时间。

依据学生的访问内容的时间间隔来进行访问顺序集合的划分，依据划分的集合。对集合中的内容进行序列模式挖掘。

依据以上方法，进行如下定义

定义一：事务数据库(transaction database)，以 blackboard 平台用户在课程中的访问历史记录为例来说明，即由用户访问课程内容记录组成的数据库。StuID 、Transaction _Time、content 分别代表学生的 ID、访问时间和访问内容的集合。

定义二：在特定时间内的访问记录的集合，记为项集。

定义三：最小频繁序列：给定最小支持度阈值（support 值），如果序列 a 在序列数据库中的支持数不低于该阈值，则称序列 a 为频繁序列。

由于 blackboard 平台本身不记录学生在平台中的登陆登出的行为，顾通过基于 k-mean 的 cluster 算法来估算出学生在平台中单词登陆的最大时长。通过分析后，70%左右的访问时间间隔集中在 3 个小时以内，加上 blackboard 平台本身的登陆回话的间隔时间设置为 1 个小时，顾研究将 1 个小时后续发生的行为视为第二次登陆的行为。基于此规则，本文采用 ruby 语言编写算法，对学生每次访问平台的数据进行频率的转换以及项集的创建，即将学生连续行为（低于 1 个小时）创建一个访问集合。例如，学生在 13 点 1 分钟，13 点 20 分钟，14 点 30 分别访问课程的内容 A,内容 B,内容 C。这访问的集合为{内容 A，内容 B，}，{内容 C}。

研究首先采用 ruby 语言编写算法从原始数据集中提取学习行为数据并生成研究所需格式的行为序列文件；之后，采用基于 java 平台的数据挖掘工具 SPMF 对课程中学生的访问进行进行行为序列的分析；

四、 研究结果以及分析

相关文献从三个角度分析行为序列和课程成绩的关系，首先是行为序列与成绩的相关分析。基于 Lag Sequential Analysis 算法通过 Pearson 相关分析计算行为序列和成绩。以序列课件 1 课件 2 为例，首先计算出所有学生访问课件 1 课件 2 的次数，然后计算该次数和学生成绩的 person 相关系数值。该算法在计算低纬度的行为序列和成绩关系时候，对于计算性能要求不高。但是在计算高纬度的行为序列时候，会消耗大量的计算资源。

假设有课程 1，课程 2，课件 3。通过 Lag Sequential Analysis 算法，3 个课件的排列组合为{（课件 1，课件 2），（课件 1，课件 3），（课件 2，课件 3），（课件 2，课件 1），（课件 3，课件 2），（课件 3，课件 1）}。然后统计具体某个序列的学生的访问次数和成绩的关联系数。

但是在基于序列模式挖掘算法中，除以上组合外，还存在组合{（课件 1，课件 2，课件 3），（课件 1，课件 3，课件 2），（课件 2，课件 1，课件 3），（课件 2，课件 3，课件 1）等。随着访问序列长度上升，可以看到不同的行为访问序列和成绩相关系数的计算空间呈指数性上升。

为解决这样的问题，研究将学生分为 2 组，即成绩较高的学生（课程排名前 20%）为一组，成绩较低的学生（课程排名后 20%）为一组。通过 CM-SPAM 算法，找出不同小组中学生的访问行为频率较高的访问集合。因为篇幅所限，仅仅截取数据，如图 1 所示。

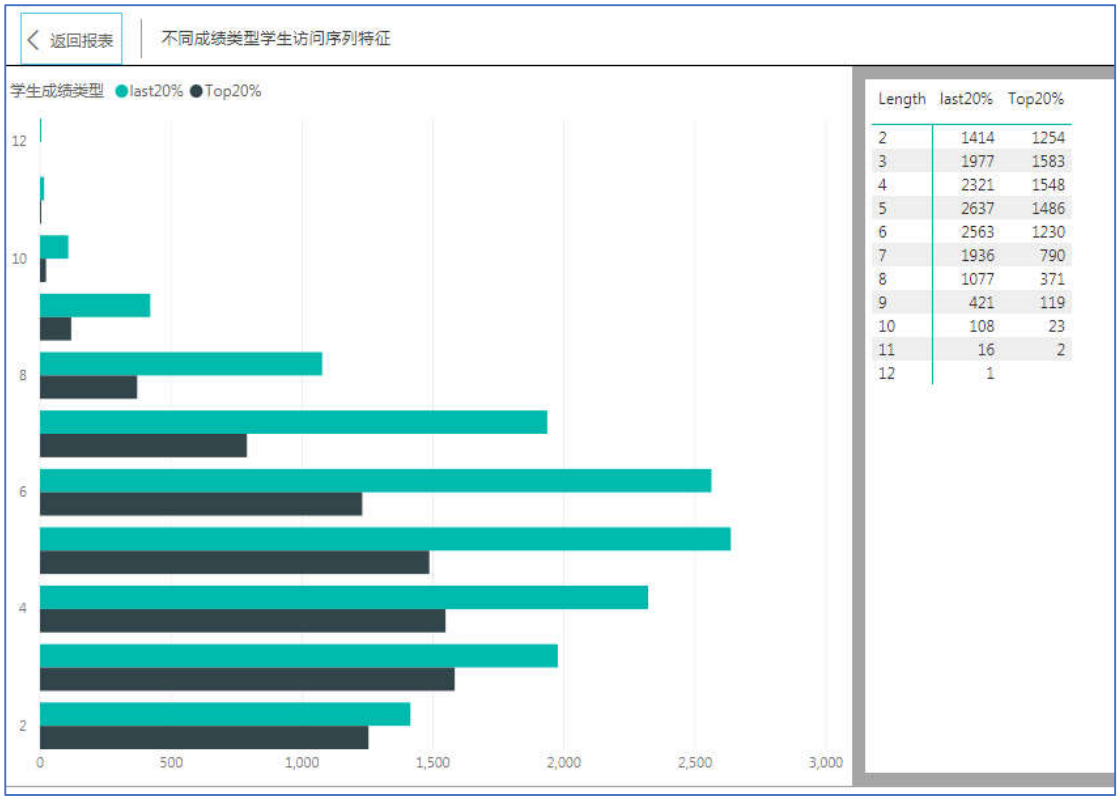

图 1 不同成绩类型学生行为访问特征

图 1 中,左侧 X 轴代表访问行为特征的发生的次数；Y 轴代表行为访问行为特征的长度。黑色代表该课程中,课程成绩排名前 20%学生访问。蓝色代表排名最后 20%学生的访问。图右侧为具体数值。例如，长度为 4 的访问序列，在排名前 20%的学生的行为特征中，发生 1548 次，在排名后 20%的学生特征中，发展了 2321 次。其中，成绩排名较后的 20%学生访问序列长度集中在 3 到 7 之间，而排名前 20%学生的访问长度集中在 2 到 5 之间。数据显示，在该课程中，学生成绩较差的学生花费较多的时间在平台中进行相关资料的寻找，反而成绩较好的学生在平台中花费的时间相对较少。

为进一步分析不同成绩类型的学生在平台中的访问特征，对访问的频率较高的序列内容进行统计。统计的结果如图 2 所示。

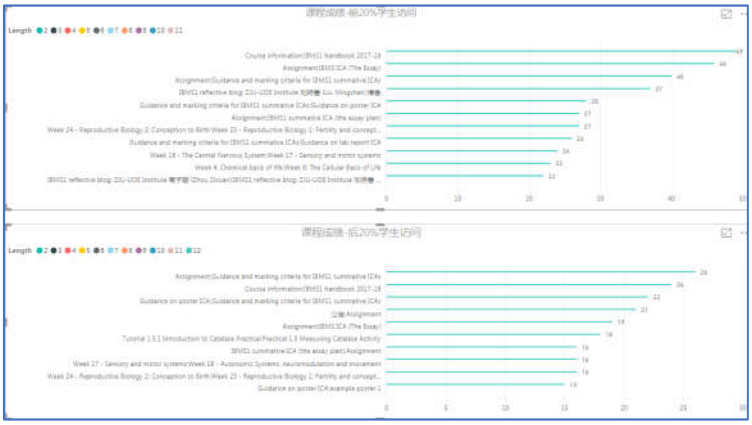

图 2 不同访问序列特征的频率

结果显示出值得思考的数据，在之前的访问统计中，成绩较差一组学生在平台的活跃度较高，按照该数据，很容易认为，该类型学生群体中，单个序列中重复的次数应该较也较高。

但是结果显示，该类型的学生在一些频繁出现的访问序列特征的总次数上，明显低于成绩优异的学生。以课件 23 课件 24 为例，成绩较好的一组学生访问该课件内容的次数明显多于成绩较差一组学生，其他的课件访问顺序也体现出同样的特征。通过访谈发现，成绩较差的学生在课程的学习中，普遍没有太好的梳理课程的思路。故在复习的时候，频繁的在不同的内容知识点之间进行跳转和查看。虽然访问次数明显多于成绩较好的学生，但是在访问上行为上没有规律。但是学习成绩较好的学生，非常好的掌握了课程的主要框架。尽管在访问的总体数量上较少，但是在平台中的访问特征比较集中。

在分析访问序列长度在 3 至 10 之间的集合中，学生的行为访问序列分为 2 种特征，一种是时间间隔较短的课件行为访问序列，比如说课件 6.1，课件 6.2，课件 7.1，即课件发布时间间隔比较短，直观上课件内容的相关度比较高。这序列行为主要集中在成绩较好的一组学生中。另外一种是时间间隔较长的课件行为访问序列，比如说课件 6.1，课件 8.2，课件 10.3。这些访问行为序列主要集中在成绩较差的一组学生成绩中，因为课程 6.1 讲述的实验内容是一个复杂的生物循环系统，然而课件 8.2 的内容，以及课件 10.3 的内容，都是需要在完全理解课件 6.1 的内容上才能理解。 从学生的主观收集到的反馈中，学生表示课件 6.1 的内容在开始学习的时候处于一种似懂非懂的状态。但是在后面学习其他和该知识点密切相关的课程内容时，发现对于之前课件的细节理解不完全，导致后面的内容无法理解，因此需要反复的回顾之前的课件内容。

五、 讨论与总结

- 促进对课程结构以及学生的认知过程的理解

在教育行业中，资历较深的优秀教师的显著特征之一就是能够很清晰的找出学生在学习中普遍认为较难的知识点，并且针对知识点进行针对性的课件设计或者案例解析。然而，对于刚进入教师行业的年轻教师来说，由于缺乏相关的经验，在课程的教授中无法找到重点，导致学生在学习中，被特定知识点给难倒。进而导致后期的课程学习进展缓慢和吃力。

- 设计有效的课程讲义

针对这样的现状，可以通过对于课程中的学生行为特征进行分析，优秀的学生表现出来一个很明显的特征就是掌握课程的结构和逻辑，通过分析优秀学生在课程空间中的行为特征，针对该特征，或者知识点进行针对性的研究，并且提出相关的讲义或者案例来帮助学生快速的理解知识点，对于教学质量的提升有着显著的促进作用。

- 提高课程干预

在教学的过程中，针对成绩较差的学生，可以观察其学习轨迹，如果存在明显的访问行为特征的无规律，可以给予相关的课程学习建议，特别是课程学习的逻辑结构，帮助其快速的掌握课程的主要知识点。

参考文献:

[1]贺超凯,吴蒙(2016). edX 平台教育大数据的学习行为分析与预测[J]. 中国远程教育, (6): 54-59.

[2]彭文辉,杨宗凯,黄克斌(2006). 网络学习行为分析及其模型研究[J]. 中国电化教育, (10): 31-35.

[3]李爽,钟瑶,喻忱等(2017). 基于行为序列分析对在线学习参与模式的探索[J]. 中国电化教育, (3): 88-95

[4]陈鹏宇,冯晓英等.在线学习环境中学习行为对知识建构的影响[J].中国电化教育,015,(8):59-63

[5]杨现民,王怀波,李冀红 (2016). 滞后序列分析法在学习行为分析中的应用[J]. 中国电化教育,(2): 17-23,32

[6] Romero, C., Ventura, S., Zafra, A., & Paul, B. et al.(2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems[J],Computers & Education,53:828-840.
[7] Fournier-Viger, P., Gomariz, A., Campos, M., Thomas, R. (2014). Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information. Proc. 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2014), Part 1, Springer, LNAI, 8443. pp. 40-52

Exploring the Online Learning Participation Behavior Pattern Based on CM-SPAM algorithm

Based on the access history data of the course on Blackboard platform of Haining International campus of Zhejiang University, this paper analyzes the sequential behavior of students ' access to course content through Cm-spam algorithm. This study mainly aims at the shortcomings of the current mainstream lag sequential analysis algorithm in behavior sequence analysis dimension, and puts forward new analysis suggestions. The results show that using Cm-spam algorithm can explorer more user behavior characteristics. This paper expects that this research can provide the idea and suggestion of curriculum content arragement optimization in educational institutions.

Keyword: