Wu, Y.-T., Chang, M., Li, B., Chan, T.-W., Kong, S. C., Lin, H.-C.-K., Chu, H.-C., Jan, M., Lee, M.-H., Dong, Y., Tse, K. H., Wong, T. L., & Li, P. (Eds.). (2016). *Conference Proceedings of the 20th Global Chinese Conference on Computers in Education 2016*. Hong Kong: The Hong Kong Institute of Education.

基于 Apriori 算法的知识点序列关联分析——以《概率论与数理统计》为例

Analyzing the Correlation of Knowledge Sequences Based on the Apriori algorithm——

——take "Probability and Mathematical Statistics" as an example

冯仰存*，顾小清，王娟，钟薇
华东师范大学教育信息技术学系
* fengyangcun@hotmail.com

【摘要】 本研究基于 MOOC 平台《概率论与数理统计》课程中的数据，利用 R 语言中的 Apriori 算法分析了知识点之间的关联情况，为知识点建立了关联规则。在此规则之下，构建了更能体现学习者学习规律的学习路径，以期改善学习者的学习效果并为 MOOC 平台的进一步完善提供思路。

【关键词】 关联分析；知识点序列；慕课

Abstract: *This research which based on the data of course-"Probability and Mathematical Statistic" in MOOCs, analyzing the correlation between knowledge sequences utilizing Apriori algorithm in R, then we build the correlation rules about knowledge. On the basis of the rules, we constructed a more reasonable learning path for improving the achievement of learners and providing good idea for the further development of MOOCs.*

Keywords: correlation analysis, knowledge sequences, MOOCs

1. 前言

MOOC 的兴起，在为学习者提供了大量学习资源的同时也面临着一些问题，如难以实现个性化学习、课程完成率低、教学模式囿于传统、以及学习体验缺失等（高地，2014），其中首当其冲的是实现个性化的学习，因为 MOOC 的核心就是开放教育和个性化教学（戴朝晖，2015）。Koschmann(2001）研究指出个性化学习和自适应学习环境的研究与开发将是大数据在教育领域应用的最终指向。个性化学习主要以构建学习路径为主，基于学习者与 MOOC 平台交互过程中产生的行为数据，利用学习分析技术探寻隐藏于数据内部的学习者行为模式，为其提供个性化的学习路径，辅助其进行个性化或自适应的学习，如可汗学院、Knewton 等都取得了突出的成效。数据挖掘领域中的 Apriori 算法，是一种非常经典的关联分析算法，主要用于探究事物之间的相互关联，并建立关联规则。得益于该关联算法的支持和经典关联分析案例—沃尔玛"啤酒与纸尿裤"的启发，本文将以"中国大学 MOOC"平台的《概率论与数理统计》课程中讨论区的"知识-问题"交互数据，用 R 语言 Apriori 算法分析对 MOOC 中知识点序列做关联分析，挖掘出知识点中的频繁项集,发现有价值的数据模式。

2. 研究方法

2.1. 数据说明

通过对国内主要 MOOC 平台进行考察，并结合吴锦辉（2015）MOOC 平台对比分析的研究成果，发现中国大学 MOOC 平台的课程数据优势突出，课程视频的碎片化逻辑清晰，"形散神不散"，整体结构统一而且索引非常便捷。以《概率论与数理统计》为例，该门课程所有知识点都与视频资源建立了一对一的映射关系，学习者在进行视频学习时或是测试过程中提出的问题和疑问会自动附加到当前知识点之下，最后这些问题将自动汇总到学习者的"我的讨

论区主页”中。因此可以对其进行布尔值编码，存在问题的记为“1”，没有提出问题的记为“0”。由于该门课程还在进行之中，因此选取前四章节共 34 讲中最活跃的 100 学员数据进行编码分析，最后进行几轮数据清理之后保留数据集样本为 52，。另外每讲内容与《概率论与数理统计授课大纲》一一对应，便于分析和比对。数据编码形式及课程大纲目录部分截图如下：

表 1 知识点编码汇总表

学生 ID	第 1 讲	第 2 讲	第 3 讲	第 4 讲	第 5 讲	…
1	0	0	0	0	0	…
2	0	0	0	1	1	…
3	0	1	1	1	1	…
…	…	…	…	…	…	…

表 2 课程大纲目录一览表

概率论与数理统计授课大纲
第一章 概率论的基本概念
第 1 讲 样本空间，随机事件
第 2 讲 事件的相互关系及运算
第 3 讲 频率
第 4 讲 概率
第 5 讲 等可能概型（古典概型）
…

2.2. 关联分析

在关联分析中，由 Rakesh Agrawal 在 1994 年提出的 Apriori 算法是最为基础有效的算法。它使得支持度、信任度、提升度的计算更为高效。在 R 语言中 Apriori 算法的核心函数是 arules 包中的 apriori 函数。

2.2.1. 信任度

如表 1 所示，学生在第 5 讲出错时，在第 4 讲均出错或有疑问，于是可以说第 4 讲和第 5 讲关联性很大，很可能是第 4 讲知识没有掌握所致，可以用信任度（confidence）来衡量，若第 5 讲记为 A，第 4 讲记为 B，则的信任度等于第 4 讲和第 5 讲同时出错的次数除以第 5 讲出错的次数：

confidence（第 5 讲=>第 4 讲）=confidence（A=>B）=P{B|A}/P{A}=100%

2.2.2. 支持度

支持度（support）表示在整个集合中同时满足 A 和 B 的概率，第 5 讲=>第 4 讲的支持度等于第 4 讲和第 5 讲同时出错的次数除以集合总数，在本研究中表现为学习者的数量：

support（第 5 讲=>第 4 讲）=support（A=>B）=P（AB）=2/3=66.7%

2.2.3. 提升度

从表 1 中可以计算出第 4 讲自身的支持度 support（第 4 讲）=P{第 4 讲}=2/3=66.7%，正常情况下第 4 讲出错的概率为 66.7%，而由 confidence（第 5 讲=>第 4 讲）=100%可知，第 5 讲有问题则第 4 讲出错的可能性为 100%,可以说第 4 讲出错会导致第 5 讲出错的可能性提高（100%/66.7%）=1.5，该比值成为出错的提升度，用于判定关联规则是否可用(一般大于 1 表示关联规则可用)。

lift（第 5 讲=>第 4 讲）

=lift（A=>B）=confidence（A=>B）/support（B）=（100%/66.7%）=1.5

3. 数据分析与讨论

为了便于分析和呈现，定义第 1 讲为 V1，以下以此类推。利用 R 语言中 arules 包中的 Apriori 算法进行关联分析。根据支持度进行排序后综合信任度和提升度选择选择了信任度为 1 的 13 条关联规则，篇幅有限故仅呈现部分规则。表中 lhs 表示左侧关联，rhs 表示右侧关联。由于本研究中数据量比较小，支持度因素的影响基本可以忽略不计，故未予以呈现。如表 3 所示：

表 3 关联规则表

Rules	lhs		rhs	confidence	lift
1	{V16}	=>	{V8}	1	2.1
2	{V15}	=>	{V8}	1	2.1
3	{V6}	=>	{V8}	1	2.1
4	{V1}	=>	{V4}	1	5.25
5	{V11}	=>	{V4}	1	5.25
6	{V20}	=>	{V2}	1	5.25

从表 3 可以发现信任度和提升度都比较高，其中第 8 讲与第 16 讲、15 讲、6 讲均存在关联性，是产生关联规则较多的知识章节，其次是第 4 讲、第 12 讲和第 28 讲，与其建立的关联规则分别有两个；剩下的均为建立单个关联规则的选项。为改善学习者的学习效果，本研究根据表 3 的关联规则绘制了 1-34 讲的推荐学习路径图，见图 1。学习者在遇到学习障碍时，可以根据推荐学习路径进行有目标有规划的学习，找到问题根源所在，解决学习隐患。

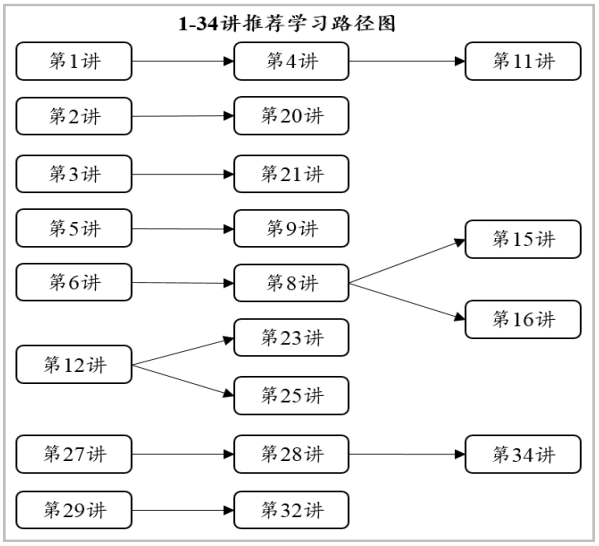

图 1 知识点关联规则路径图

为了进一步搞清楚各知识点的出错情况，本研究又对于各讲的问题数目进行了统计汇总，见图 2 和图 3。通过对比发现，第 8 讲出错比较多，说明其确实处于比较重要的地位，而第 6 讲虽然出错较多（位居第二位），但是其与其他知识的关联性较低，仅与第 8 讲存在关联规则，故可以推测该知识点比较独立或是数据量不够大没有体现出第 6 讲的关联价值所在。

图 2 每讲问题统计图

另外，通过表 1 还可看观察到关联规则几乎产生于前几讲，这与 MOOC 中学习者结业率低有着密切的联系。随着学习的不断深入，学习者不断跳出，相应的互动数据就会不断减少，如图 3，这也是 MOOC 面临的亟待解决的问题，同时也可能造成分析的误差。

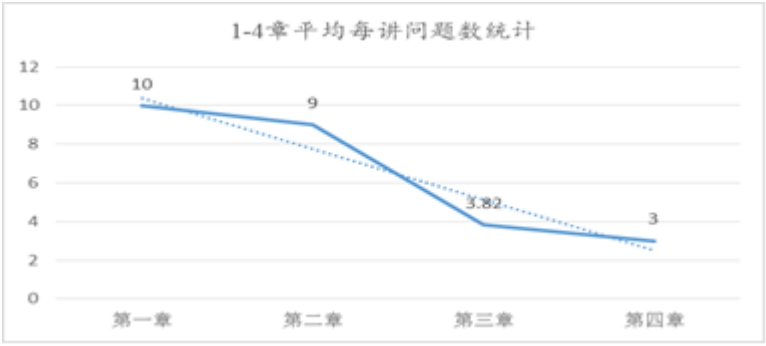

图 3 章每讲平均问题统计图

4. 总结与展望

通过关联分析和统计分析可以发现，学习者的学习是循序渐进的，知识之间有一定的相互承接关系。本研究存一些在不足之处，首先是数据的样本量偏小，分析结果会有一定误差；其次，由于参与人数不断减少数据质量不是非常理想；最后，由于课程还在进行，数据结构没有十分完善。下一步的工作将会对整门课进行关联分析，并将其与学习者的测试成绩进行相关分析和回归分析，切实为提升学习者的学习体验和学习效率而努力奋斗。

致谢

本论文得到华东师范大学研究生出国（境）短期研修专项基金资助。

参考文献

吴锦辉（2015）。我国主要慕课(MOOC)平台对比分析。**高校图书馆工作**，**35**（1），11-14。

李桥和阳春华（2010）。关联规则 Apriori 算法在教学评价中的应用。**计算机与数字工程**，**38**（6），49-51。

李明（2015）。**R 语言与网站分析**。北京:机械工业出版社。

高地（2014）。MOOC 热的冷思考——国际上对 MOOCs 课程教学六大问题的审思。**远程教育杂志**，**32**（2），39-47。