# 基于 SimRank++ 的课程内容相似性搜索研究*

邵炤昭[1]，张 向[2]
（1.浙江大学海宁国际校区 图书信息中心，浙江 海宁 314400；
2.中国农业银行黄冈黄州支行，湖北 黄冈 438000）

摘 要：随着在线学习平台在高等教育机构中的普及，针对在线学习平台中的内容进行相似度分析，可帮助教师更好地了解教学工作中学生关注的重点和难点。通过抓取学生在 E-learning 平台（Blackboard learn）上的课程内容访问历史记录，从学生访问课程内容的关联关系出发，基于课程内容特性定义课程内容相似度，计算课程内容相似度。针对相关学者在个性化学习中提出的算法上的不足，本文提出一种基于 simrank++的算法来分析课程内容的相似性以及学生对于知识点的关注度。通过研究结果证实，基于 simrank++的算法分析结果更加能反映学生关注的课程内容特征。基于分析结果，可以向高校负责学科资源建设，以及教学资源建设的部门提出教学提升建议。

关键词：E-learning；在线学习分析；个性化学习；SimRank++
中图分类号：G434 文献标志码：A 文章编号：1673-8454(2019)05-0044-04

## 一、在线学习平台在高等教育中的影响

随着互联网技术的快速发展，越来越多的基于互联网的信息平台被用于教育行业，特别是基于互联网的在线学习平台已经在国内外高校普及。目前主流的在线平台，例如 Blackboard、Moodle，已经成为高校课堂教育的重要补充。通过这类平台，教师可以发布与课程相关的教学资料以及作业，与学生就学习中的遇到的困惑和重点进行交流。随着大数据、数据挖掘等相关技术的普及和发展，围绕在线学习平台的关于学习分析的研究越来越多，并且已经取得一定的进展，包括：针对平台访问次

* 基金项目：2017 年浙江大学国际联合学院（海宁国际校区）一般课程“中外合作办学中信息化项目建设机制探索”（课题编号：1705）。

## 五、结论

本次研究的目的不仅是为了调查大学生对雨课堂支持下的翻转课堂教学模式的满意度，同时也是为了对以后的教学提供指导，从中找到教学中存在的真实问题，从而改进教学。根据本次调查结果，提出以下几个建议：

（1）增加教学内容的系统性。一个具体的知识系统能够促进学生更好地学习，促使新旧知识之间的联系，有利于学生对整体教学的把握，与此同时，能够让学生意识到自身存在的不足，进而查漏补缺。

（2）增强学生角色转变意识。翻转课堂更加注重学生的中心地位，着重强调学生的自主性，与传统课程是不一样的，因为学生对自身角色的转变也要有强烈的意识，不能抱怨教师讲授的时间短使自己一无所获。而如何才能提高学生角色转变的意识是需要进一步研究的。

（3）完善小组协作。在小组分工时，应该提前了解学生擅长的领域，从学生的偏好出发，自由组队，与此同时，小组成员之间应该多加沟通，在完成任务时，成员之间进行明确的分工，多组织活动，调动小组的活跃度。

（4）对弹幕提高课堂互动进行合理安排。弹幕是一个提高课堂互动的好工具，但是弹幕应该在什么时候出现、什么时候关闭是需要进一步研究的，同时，弹幕上面除了课堂外的信息应该如何控制也是应该进行研究的。

参考文献：

[1]王帅国.雨课堂:移动互联网与大数据背景下的智慧教学工具[J].现代教育技术,2017(5):26-32.

[2]王坦,吉标.“翻转课堂”模式的理性审思[J].课程·教材·教法,2016(6):55-61.

[3]肖安宝,谢俭,龚付强.雨课堂在高校思政课翻转教学中的运用[J].现代教育技术,2017(5):46-52.

[4]孙朝娟,孟洁然.翻转课堂“雨课堂”对教学效果提升的策略研究[J].汉字文化,2018(14):114-116.

（编辑：王天鹏）

数来预测学生成绩走向；分析课程内学生对教学内容的访问优化和调整课程结构；基于平台中课程内容以及教师参与度的监管和教学评估。

在针对教学内容优化的研究中，李爽等人通过行为序列分析，找出课程中学习参与模式对课程最终成绩的影响。[1]陈鹏宇等人通过 Person 相关性分析学生在课程中内容的参与度和知识构建水平的关联度。[2]田阳等人分析了课程中社交行为与成绩的相互影响 [3]。目前，针对课程内容的相关性分析报告较少。在传统的电商或者社交网站中，相关性分析扮演着重要的地位，不少网站采用相关性分析来进行朋友或者商品的推荐，通过相关性算法，找出用户可能需要的产品以及可能认识的朋友，并进行推送。因此，相关性分析研究，对于教学资源的推荐以及分析学生关注的知识重点，可能存在一定的帮助。

## 二、主流个性化学习推荐服务算法介绍

个性化学习服务，即根据学生的特点、当前学习情况，向其推荐课程、学习活动、学习资料以及学习方法等，提供学习建议，动态调整学习安排，是当前在线学习行为研究的热点问题之一。目前在个性化学习中，常见的相关性算法包括：①Person 相关性分析；②基于 Aprior、FT-GROW 算法的相关性分析；③基于 K-MEAN 的聚从算法。 这些算法在一定程度上能找到不同知识点之间的关联。但是更加深度的关联分析，无法揭示之间的关联度。如图 1 所示。

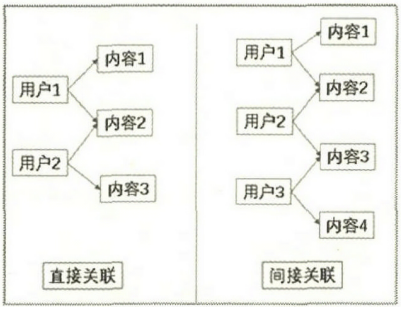


图 1　用户访问课件的关系

在传统的推荐算法中，例如关联算法、决策树算法、聚类算法。都要求物品之间存在直接的关联，如图 1 左侧所示，用户 2 和用户 1 的访问存在一定程度的交集，以课件 2 为例，通过分析课件 2，实现课件 1 对用户 2 的推荐，课件 3 对用户 1 的推荐。

假设存在另外一种情况，如图 1 右侧所示，用户 1 访问内容 1 和内容 2，用户 3 访问内容 3 以及内容 4，在常规的推荐算法中，因为内容 2 和内容 3 的存在，通常可以做到用户 1 和用户 2 的关联，用户 2 和用户 3 的关联。但是没办法做到用户 1 和用户 3 的关联。因为用户 1 和用户 3 之间不存在交集。但是从推荐的逻辑上，可以推导出内容可以推荐给用户 2，假设用户 2 阅读该内容，那么基于用户 2 和用户 3 之间存在关联，可以将内容 1 推荐给用户 3，这样的关联推荐在推荐系统中一般称为拓扑结构中节点推荐。

## 三、SimRank++算法的原理介绍

针对存在的问题，Antonellis 等人在 2002 年提出的 SimRank 算法可以用来评估课件内容的相似度[4]。SimRank 算法是一种适用于计算拓扑结构中任意 2 点关联度的算法，该算法以迭代的方式来计算目的对象的相似性，并且在很多行业都被广泛使用。例如魏琳通过SimRank 算法，对慢性胃炎的发病机理进行相似度计算，找出慢性胃炎临床症状相似度。[5]田玲等人通过 SimRank 算法找出中药方剂数据中“效-效”相似度，实现对不同药效之间的相似度归纳。[6]朱金山等人为解决城市公共自行车系统快速发展导致的潮汐问题，提出基于SimRank 的站点间关联度和相似度计算，采用最大相似度优先的原则进行聚类，为站点区域划分，公共自行车调度策略等提供理论基础。[7]王家海等人采用 SimRank 算法，设计了一套能够精准描述数控机床的故障诊断系统，并且该系统具备知识学习能力。[8]

从结构上看，在线学习平台中的课程内容推荐是一种以课程空间知识点为节点的网络拓扑结构。知识点之间的相似数值可以用学生对于该知识点的访问频繁程度来衡量。因此，本文根据学生访问不同知识点的频率，提出一种基于网络拓扑结构的 SimRank++算法来进行个性化学习推荐。

Antonellis 等人在 2008 年针对 SimRank 算法的不足提出了 SimRank++ 算法，该算法提出了权重以及节点相关度等影响因子，进一步完善了算法的应用范围。[9]

受以上行业成果经验启发，结合在线教育平台中用户数据和用户行为，可将用户以及课程内容构建成访问关系网络。

定义 1（学生访问课程内容拓扑网络）记为 G=(S,C,E)。其中 S 为所有学生的集合，C 为所有课程内容的集合，E 为学生访问课程内容的关系。三元组(s,c,e)表示学生访问课程有向连接关系。E(c)为所有访问该内容学生的集合。

定义 2（课程内容相似度）给定 2 个课程内容(a,b)∈C，基于定义 1，内容相似度定义如下：

$$Sweight^{(a,b)}=evidence(a,b)*C\sum_{ieE(a)}\sum_{ieE(b)}W(a,i)W(b,j)$$

Sweight(i,j)

其中:

$$evidence(a,b)=\sum_{i=1}^{|E(a)|\cap|E(b)|}\frac{1}{2^i}$$

$$W(a,j)=spread(i)*normalized\_weight(a,i)$$

$$normalized\_weight(a,i)=\frac{W(a,j)}{\sum_{jeE(a)}W(a,j)}$$

$spread(i)=e^{-variance(i)}$，其中−variance(i)为变量 i 的所有关联权重的方差。

SimRank++ 算法以迭代的方式更新集合中的相似度，经过多轮计算后，结果收敛，趋向一个极值。迭代次数与相似度的精确值相关（精确到小数点后位数）。因此迭代次数可以通过计算进行调整。相关学者发现，使用 C 的参数和迭代的参数密切相关，建议在实现精确度不低于 1%的情况下，C 取值为 0.6，迭代次数为 6。[10]

SimRank++算法由于是迭代性密集计算，因此在实际操作中，可以采用多线程计算提高计算效率。具体实现方法如下：在定义 2 中，可以将所有访问 a 的集合和访问 b 的集合的组合，划分到不同线程的计算单元，线程计算单元划分依据可以是所在机器的 CPU 核数，或者其他自定义数量。然后将计算结果汇总。

算法分成 2 个阶段：首先，根据定义 2 计算课件内容的相似值矩阵 M，相似值矩阵 M 中元素数值为课程之间的相似距离值，其次为图像化表示课程内容的相关度，可以通过汇聚算法来进行聚类的划分。

在聚类算法选择中有以下因素需要考虑：首先无法预测聚类个数的范围，其次个体特征更多是内容之间的差距，随着内容数量的增加，计算聚簇的代价就越高。基于以上因素，相关学者推荐使用 hierarchy 算法作为聚簇的算法模型[7]。

## 四、实验环境以及结果分析

由于基于关联算法的文章中很少公布其数据集，和本文算法没有直接对比的样本，顾本文仅仅分析该结果的现实意义以及该结果对于教学可能的促进作用。

本次研究的数据取自浙江大学伊利诺伊大学厄巴纳香槟校区联合学院 2018-2019 年秋季 Calculs3 课程使用 blackboard 平台的数据，数据的抽取为（用户 ID、访问课程内容 ID、访问时间）。其中学生数量为 30，内容数量为 72。实现的开发环境为 Win7 平台，运行平台为 jruby1.95（因为默认的 ruby 运行环境本身不支持利用多线程提高运算效率）。

为对比在使用 SimRank++算法前后课件内容的关联度，基于篇幅所限，本文截取部分数据来比较在使用 hierarchy 算法时候的汇聚效果。样本数据如表 1 所示。

表 1 部分学生访问课件数量

| 课件 / 学生 | 课件 1 | 课件 2 | 课件 3 | 课件 4 | 课件 5 | 课件 6 |
|---|---|---|---|---|---|---|
| 学生 1 | 0 | 1 | 17 | 20 | 2 | 1 |
| 学生 2 | 1 | 0 | 22 | 31 | 5 | 2 |
| 学生 3 | 0 | 0 | 10 | 21 | 1 | 2 |
| 学生 4 | 0 | 0 | 12 | 15 | 1 | 2 |
| 学生 5 | 0 | 0 | 15 | 9 | 1 | 1 |

在不使用 Simrank++算法情况下，课程内容汇聚效果如图 2 所示。

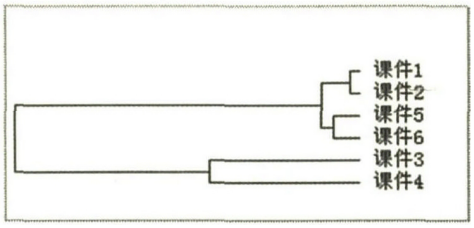


图 2 不使用 simrank++算法课件的汇聚效果

在默认情况下，Hierarchy 汇聚算法采用访问的次数作为汇聚的依据，因此，可以看到课件 3、课件 4 的相似度最高，其他课件相似度相对较低。

在使用 Simrank++算法后，得到课件相似矩阵数据如表 2 所示。其中 1 代表最相关，0 代表不相关，例如课件 1 和课件 2 最不相关，课件 1 和课件 5 最相关。

表 2 课件相似度矩阵

| 课件 / 课件 | 课件 1 | 课件 2 | 课件 3 | 课件 4 | 课件 5 | 课件 6 |
|---|---|---|---|---|---|---|
| 课件 1 | 1 | 0.17 | 0.28 | 0.28 | 0.32 | 0.27 |
| 课件 2 | 0.17 | 1 | 0.27 | 0.27 | 0.27 | 0.25 |
| 课件 3 | 0.28 | 0.27 | 1 | 0.52 | 0.53 | 0.51 |
| 课件 4 | 0.28 | 0.27 | 0.52 | 1 | 0.53 | 0.52 |
| 课件 5 | 0.32 | 0.27 | 0.53 | 0.53 | 1 | 0.52 |
| 课件 6 | 0.27 | 0.25 | 0.51 | 0.52 | 0.52 | 1 |

进行汇聚的效果如图 3 所示。

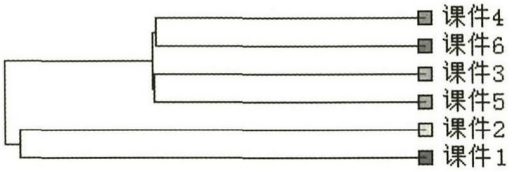


图 3 基于相似度的矩阵的课件汇聚效果

因为 SimRank++算法是一种基于拓扑结构的推荐算法，大量学生访问课件 3 和课件 4 后，也在一定程度

上访问了课件 5 和课件 6。该算法认为课件 3、课件 4、课件 5、课件 6 相似度较高。但是学生在访问课件 3、课件 4 后，较少访问课件 1 和课件 2。通过汇聚图，可以看到这种明显的区别。

1.数据结果分析

通过对整个班级中课件访问次数的相似度计算，得出课程内容的汇聚效果，数据体现出以下特征。

(1)发现一：第一周的内容基本和其他教学周的相关程度比较低，因为第一周的课件内容基本都是课程的入门介绍，教师的联系方式等，和后面的相关教学内容关联度不大。

(2)发现二：在学期中期，课程有期中考试，且考试成绩被计入课程总成绩。数据显示学期中期的课程内容(教学周第 10 周至教学周第 13 周)和期中考试关系度紧密。实际上，考试的重点也是集中在这 3 周的学习内容。

(3)发现三：如果相关教学周有课后作业，那么这些教学周的相似度极高，可能说明学生积极访问该批课程内容的主要原因是在完成相关课后作业时，需要查看该教学周的讲义。

(4)发现四：教学周最后几周的课件内容不被学生广泛访问，进过调研，该课程内容主要是进阶阅读，不作为期末考试的重点。

(5)发现五：如果教学周没有课后作业或者习题，该课件内容不会被学生广泛访问。

2.实验结果对于教学活动提升的建议

(1)高校图书馆学科资源建设

高校的教学资源建设一直是教学质量提升的重要保障举措。在过去的教学资源建设过程中，学科建设和教学过程存在一定程度的信息不对称，即购买的学术资源不是学生或者教师所关注的。导致采购的图书或者电子数字资源利用率不高。通过发现的问题，可以在图书资源采购、数据库采购或者优秀教学课程录制和引进的时候，重点考虑学科建设资源能够覆盖学生关注的重点或者难点。通过分析课件中学生访问图书馆资源链接的次数，可以对相关图书馆资源购买优化。例如在购买图书资源的过程中，更多考虑该出版社或者该作者的著作。通过信息化手段，特别是数据挖掘等工具，提高数字资源的使用效率，提升教学质量。

(2)教学单位课时分配

其次，可以建议相关教学管理单位提高相关课程内容的讨论课时，或者利用其他手段，对课程中的难点予以更多解答。

(3)教师课程准备

在高校教学活动的开展过程中，相关课程的任课教师可能发生变动，对于新的任课教师来说，可以通过研究历史数据，找到课程中学生关注的重点或者难点。通过在课堂中重点讲解，提高学生的学习成效。

(4)个性化学习知识推荐

最后，可以对该课程中学习成绩较差的学生，进行课程内容的推荐，通过该方法，让学生快速抓住课程的核心或者重点，进行有针对性的预习和复习。

五、结束语

本文提出基于 SimRank++算法来推断出课程内容相似性或者关联度，结果揭示了教学中学生关注的重点以及难点。相关教学机构可以利用该数据对教学的过程或者学科资源建设进行相应的优化。未来的研究工作是扩展课程内容的关联度边界。

参考文献：

[1]李爽，钟瑶，喻忱.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，2017(3): 88-95.

[2]陈鹏宇，冯晓英，孙洪涛.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，2015(8):59-63，84.

[3]田阳，冯锐，韩庆年.在线学习社交行为对学习效果影响的实证研究[J].电化教育研究，2017(3):48-54.

[4]SimRank:a measure of structural-context similarity. Jeh G,Widom J. Proc of the8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2002.

[5]魏琳.基于 SimRank 的慢性胃炎相似关系挖掘的研究与分析[J].福建电脑，2014(9):93-96.

[6]田玲，曾涛，陈蓉.基于 SimRank 的中药"效-效"相似关系挖掘[J].计算机工程，2008(12):242-244.

[7]朱金山，刘良旭，周超兰.基于 SimRank 的公共自行车站点聚类算法[J].计算机工程，2018(4):12-16.

[8]王家海，徐旭辉，沈佳豪等.基于粗糙集结合 SimRank 算法的数控机床故障诊断研究[J].组合机床与自动化加工技术，2018(2):84-86.

[9]Simrank++: Query rewriting through link analysisof the click graph. Antonellis I,Molina H G,Chang C C. Proceedings of the VLDB Endowment,2008.

[10]Dmitry Lizorkin,Pavel Velikhov,Maxim Grinev, Denis Turdakov. Accuracy estimate and optimization techniques for SimRank computation[J]. The VLDB Journal, 2010,19(1).

(编辑：王晓明)