

基于多维度推荐的智慧在线学习系统^①

吴冉¹, 黄超¹, 卢盛祺^{1,2}

¹(复旦大学 软件学院, 上海 200000)

²(上海财经大学 信息管理与工程学院, 上海 200433)

摘要: 针对目前在线学习系统中存在的不足, 探讨如何有效地运用数据挖掘技术建立智慧的在线学习系统. 从大量的用户数据中挖掘出关联关系, 用以提供全面个性化、定制化的学习过程序列. 利用数据挖掘着重发现用户与课程之间、课程与课程之间、用户与用户间的关联, 形成一个多维度的网络. 利用多维度推荐为用户推荐有价值的课程. 实验表明多维度推荐具有良好的准确性和良好的用户体验.

关键词: 智慧在线学习系统; 数据挖掘; 多维度网络; 多维度推荐; 扩散激活

Intelligent E-Learning System Based on Multi-dimensional Recommendation

WU Ran¹, HUANG Chao¹, LU Sheng-Qi^{1,2}

¹(School of Software, Fudan University, Shanghai 200000, China)

²(School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China)

Abstract: Aiming at the shortage in the present e-learning systems, how to establish an intelligent e-learning system effectively using data mining is discussed. Associations are extracted from mass of user data in order to provide a comprehensive personalized, customized learning sequence. Associations between Users & Courses, Courses & Courses, Users & Users are mined using data mining, and a multi-dimensional network is built on that basis. Multi-dimensional recommendation is utilized to recommend useful courses to users. Experiments show that multi-dimensional recommendation has a high precision and provides a good experience to users.

Key words: intelligent e-learning system; data mining; multi-dimensional network; multi-dimensional recommendation; spreading activation

随着互联网技术的蓬勃发展, 在线学习系统(e-learning)被广泛用于公司和大学中的教育培训. 在线学习系统提供了大量电子化的教学资源, 用户可以方便地获得课程信息并进行交流共享. 然而海量的教学资源使用户需要花费越来越多的时间寻找想要的课程进行学习. 此外, 大部分用户在进行学习时, 无法确定自己需要学习什么、应该学习什么. 盲目的学习使用户的学习效果大打折扣. 用户迫切地需要能够根据用户信息提供适合的学习资源的智慧型在线学习系统^[6]. 智慧型学习系统不仅提供了大量可用的学习资源, 还能为用户提供个性化的、完善的学习体验.

在线学习系统积累了大量与用户行为相关的数据, 例如课程学习、课程收藏、交流讨论等, 隐含了大量

有价值的信息. 许多在线学习系统利用数据挖掘和推荐技术, 从这些用户数据中提取有价值的信息, 从而为用户推荐可能感兴趣的学习课程^[4]. 例如, Sunita B Asher 等人利用关联规则和聚类根据用户的课程选择信息为用户推荐可能感兴趣的课程组合^[1,2]. Paola Monachesi 等人通过挖掘在线学习系统用户的社交关系找到最相近的其他用户, 根据其他用户的喜好为用户进行推荐^[3]. 这些在线学习系统通过挖掘单个用户与所学课程间的关联规则或者以往其他类似用户的学习模式, 以此为用户推荐课程. 然而这种推荐仅仅局限于单一维度的规则, 不能很好地满足用户的需求.

本文的智慧在线学习系统从多个维度出发, 综合了用户与课程、用户与用户、课程与课程间的关联关

① 收稿时间:2015-04-03;收到修改稿时间:2015-05-28

系, 建立了一个包括课程和社交伙伴的网络. 利用这一综合网络, 获取有关用户兴趣的有效关联. 首先, 本文提取出用户和课程间的关系. 然后对用户与用户间的社交关系进行挖掘, 建立社交网络. 最后发现不同课程中的关联规则.

1 智慧在线学习系统

许多在线学习网站已经意识到智慧化的重要性. 例如, Codecademy 在线互动编程学习网站已从两个方面将“智慧”应用于网络教育体系中. Codecademy 一方面通过不同用户在网络上学习的模式来调整导航模式, 以提供个性化的课程内容. 另一方面基于以往其他类似学生的学习模式为用户推荐个性化的学习路径^[5]. 通过数据分析技术获得用户的学习行为, 并做出适当反馈可以有效提高用户学习效率和学习积极性. 以 Udemy 为代表的开放式在线培训系统则在群体智能方面表现卓越, 并且所有用户都可以在该系统上开设自己的课程^[7].

本文提出的智慧在线学习系统从多个维度提取出有价值的关联模式. 用户-课程关联抓取了用户的兴趣, 反映了用户背景与学习兴趣间的关联; 用户-用户挖掘了用户间的社交关系, 能够发现具有相似兴趣的好友; 课程-课程关系则将不同课程按照一定序列组织在一起. 智慧在线学习系统利用这些关联关系构建了一个包含用户与课程的综合型网络, 将用户个人的学习行为映射到整个网络中.

从以下几个方面提高在线学习系统的智慧性:

①挖掘用户学习行为和学习方向. 利用用户的历史信息、资源收藏信息进行数据挖掘实现智能推荐^[8]. 通过以下分析挖掘出用户的学习行为: 分析收藏教程和其他教程的相关性, 向用户推荐他们可能感兴趣的教程; 分析学习开始时间和学习结束时间计算出停留时间, 得到用户对该课程的兴趣度.

②挖掘用户的社交关系. 在线学习系统一般会在用户注册时要求用户填写个人资料, 如专业领域、教育经历等, 一般也会支持用户添加好友等功能. 在线学习系统可以根据用户的好友初步建立起用户的社交网络, 同时根据用户的背景资料将具有相同经历、背景的用户作为一种隐式的社交补充进社交网络中. 利用用户的社交关系可以推荐给用户具有相似兴趣的好友, 同时可以根据用户好友的兴趣预测用户的兴趣.

③发现课程间的联系. 用户在学习不同课程时是具有一定顺序的, 根据不同课程的内容与难度会整合成具有一定顺序的学习序列. 通过挖掘同一用户学习的课程序列, 可以建立学习课程间的关联规则. 当一个新的用户在初期学习了若干个课程后, 可以根据用户此时的学习情境从关联规则库中找到对应的关联规则, 预测该用户后续需要学习的内容并为用户推荐对应的课程序列.

2 智慧的推荐算法

本文的智慧在线学习系统从多个维度出发, 综合了用户与学习资源、用户与用户、课程与课程间的关联规则, 建立了一个综合网络. 并通过扩散激活为用户推荐与其最接近的其他用户与课程.

2.1 多维度网络构建

本文构建了一个包含用户、课程的多维网络.

2.1.1 用户-课程关联

用户与课程的关联可以用用户对课程的兴趣程度表示. 用户对某一课程表示的兴趣越强烈, 他们之间的关联越紧密. 用户对该课程的学习时间长短能够反映用户对该课程的兴趣程度. 学习时间越长, 用户兴趣越强烈; 反之则越弱. 用户对课程的兴趣度通过以下公式进行计算:

$$Intre_{u,c} = \frac{time(u,c)}{time(c)} \alpha^n \quad (1)$$

其中, $time(u,c)$ 为用户 u 对课程 c 的学习时间, $time(c)$ 为课程 c 的课时长度. α^n 为一个时间衰减因子, 表明随着时间 n 的增大用户对课程的兴趣度会相应减弱.

2.1.2 用户-用户关联

本文中, 用户-用户关联主要通过三个方面进行判断. 第一, 共同关注, 即如果两个用户具有较大比例的相同关注, 那么说明他们可能具有相同的兴趣目标, 可能构成较为紧密的关联关系.

$$Fo_{u,v} = \frac{|fo(u) \cap fo(v)|}{|fo(u) \cup fo(v)|} \quad (2)$$

其中, $Fo_{u,v}$ 反映了用户 u 和用户 v 间的共同关注关系. $fo(u)$ 为用户 u 的关注集合, $|fo(u) \cap fo(v)|$ 为用户 u 和用户 v 关注集合的交集包含的元素个数. $|fo(u) \cup fo(v)|$ 为用户 u 和 v 关注集合的并集包含的元素个数.

第二, 用户间的互动行为. 在线学习系统中用户之间可以发生互动, 包括回复消息、投票. 用户在在线

学习系统的同一版面间进行互动,可以在一定程度上反映他们在关注同一个话题.

$$Int_{u,v} = \frac{|int(u) \cap int(v)|}{|int(u) \cup int(v)|} \quad (3)$$

其中, $Int_{u,v}$ 代表用户 u 和用户 v 间的互动关系. $int(u)$ 为用户 u 的互动行为集合, $|int(u) \cap int(v)|$ 为用户 u 和用户 v 互动集合的交集包含的元素个数. $|int(u) \cup int(v)|$ 为 u 和 v 互动集合的并集包含的元素个数.

第三, 用户学习的课程的相似度. 如果两个用户学习过的课程非常相似, 那么就认为这两个用户具有相似的学习目标, 那么他们间的关联关系也会较紧密.

$$Sim_{u,v} = \frac{|sim(u) \cap sim(v)|}{|sim(u) \cup sim(v)|} \quad (4)$$

其中, $Sim_{u,v}$ 代表用户 u 和用户 v 间的互动关系. $sim(u)$ 为用户 u 的互动行为集合, $|sim(u) \cap sim(v)|$ 为用户 u 和用户 v 互动集合的交集包含的元素个数. $|sim(u) \cup sim(v)|$ 为 u 和 v 互动集合的并集包含的元素个数.

为了量化两个用户间关联关系的紧密程度, 本文定义了紧密度:

$$Close_{u,v} = \delta_1 Fo_{u,v} + \delta_2 Int_{u,v} + (1 - \delta_1 - \delta_2) Sim_{u,v} \quad (5)$$

其中, δ_1 、 δ_2 代表了不同关系对紧密度的贡献.

2.1.3 课程-课程关联

为了量化课程之间的关联, 本文利用关联规则挖掘发现用户在学习课程时的规则. 由于不同课程具有不同的难度, 因此用户对课程的学习通常是具有一定序列的. 本文假设如果许多用户在学习了课程 A 之后又学习了课程 B, 那么就构成了一个从 A 到 B 的关联. 本文利用 Apriori 挖掘课程间的关联规则, 设置最小支持度和置信度, 用来过滤掉支持度和置信度较低的规则. 同时将课程间的关联关系用以下公式进行量化:

$$Asso_{c,d} = \frac{P(cd)}{P(c)} \quad (6)$$

其中, $P(cd)$ 为课程 c 和课程 d 同时出现的概率, $P(c)$ 为课程 c 出现的概率. 可以看出, 课程间的关联关系 $Asso_{c,d}$ 本质是规则 $c \Rightarrow d$ 的置信度.

将以上三种关系进行整合, 得到如图 1 所示的多维度网络. 图中笑脸表示用户, 方框表示课程. 笑脸间的实线代表了用户间的社交关联, 笑脸与方框间的虚线代表用户与课程的关联, 方框间的带箭头实线表示课程间的有序关联.

2.2 多维度推荐

对于一个用户 u , 本文利用多维网络从用户-用户关联、课程-课程关联、用户-课程关联为用户推荐可能感兴趣的课程.

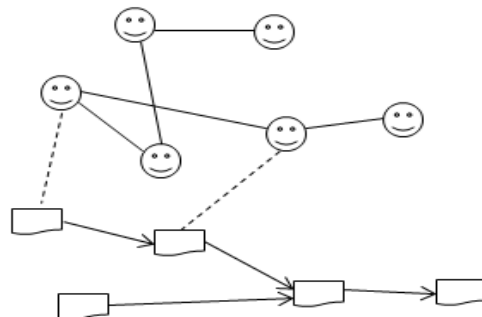


图 1 多维度网络

首先, 找到与 u 最为紧密的 k 个用户, 综合 k 个用户的兴趣与用户 u 的兴趣找到综合兴趣度最高的 m 门课程. 综合兴趣度的计算方法如下:

$$SumIntre_{u,c} = \frac{\varphi}{k} \sum_k Close_{u,uk} \times Intre_{uk,c} + (1 - \varphi) Intre_{u,c} \quad (7)$$

综合兴趣度即这 $k+1$ 个用户(包括用户 u)的兴趣加权和.

然后, m 门课程以综合兴趣度为初始值, 沿着课程-课程网络进行扩散激活, 以激活值最高的 t 个节点推荐给用户 u .

扩散激活(Spreading Activation)模型是基于内容的过滤方法的扩展, 可以通过网络连接寻找与初始激活节点高度关联的节点^[9]. 扩散激活是一个迭代的过程, 由网络中的若干初始激活节点开始, 沿着节点周围的连接向四周扩散, 激活其他节点. 当这些节点被激活后, 也开始激活其他相连的节点. 这一过程不断重复, 直到满足停止条件. 在扩散激活模型中, 激活值、扩散阈值与最大扩散距离是扩散过程中非常重要的三个参数. 激活值反映了激活节点的激活能量, 扩散距离是扩散模型扩散过程经过的连接数量. 扩散阈值与扩散最大距离是扩散过程是否停止的阈值, 如果激活节点的激活值低于扩散阈值, 或者扩散距离已经达到了最大扩散距离, 那么激活节点的扩散激活过程就会停止.

①初始化. 选择得到的 m 门课程作为初始的激活节点, 将初始节点的激活值设置为综合兴趣度, 剩余节点激活值设置为 0; 设置扩散阈值与扩散最大距离 sd .

②扩散激活过程. 由初始激活节点开始, 按照图

中的连接激活其他节点. 当一个激活节点通过一个连接扩散激活另一个节点时, 激活值与连接权重的乘积将作为被激活节点的激活值. 比如, 一个激活值为 4 的节点通过一个关联度为 0.3 的连接激活了另一个节点, 那么被激活节点的激活值就为 4×0.3 . 如果一个节点被多个激活节点激活, 那么它的激活值就是不同激活值之和.

③终止条件. 每次完成一次激活, 检查激活节点的激活值是否低于扩散阈值或者扩散距离是否超过了最大距离, 如果是的话激活节点就停止激活其他节点. GT 扩散激活后得到的扩展话题集为 $CS = \{C_1, C_2, \dots, C_j\}$.

④节点筛选. 设定一个阈值 ϕ_s , 去除 CS 中激活值低于该阈值的课程, 得到最后的课程集合作为要推荐给用户的课程集合.

3 实验

实验部分对本文提出的方法进行了验证, 并与其他方法进行比较.

3.1 实验数据

实验采用的数据来源与在线学习系统网易云课堂 (<http://study.163.com/>). 网易云课堂是一个开放式在线教育系统, 用户可以在该系统中学习课程. 实验使用的数据集包括 2319 条课程资源信息(包括课程名称、教师、机构、一级分类、二级分类、预备知识、价格、课时)、4536 条用户学习课程信息(用户、课程、学习时间)以及 322 条用户信息(用户 ID、教育经历、职业经历、擅长技能、关注/粉丝). 首先对数据进行预处理:

①数据选择. 本文选取源数据中 60% 的数据作为训练集训练模型, 剩余 40% 的数据作为测试集.

②异常处理. 由于人为操作失误或是系统缺陷, 系统数据库中的错误数据很难避免, 将不完整的记录删除. 同时, 将异常数据删除. 比如, 在数据处理过程中发现用户学习课程信息中有很多异常值, 如用户学习某一课程的时间过短(远远低于课程的时长), 说明许多用户没有完成该课程的学习. 因此, 根据课程的课时信息设置一个阈值, 将学习时长低于该阈值的记录标记为异常记录并删除.

③数据集成. 对于包含不同来源的数据源数据, 需要将不同数据集集成到一个数据集中. 在本实验中, 源数据包含两个部分: 网站数据库的表数据(包括课程

信息表、课程分类表、机构信息表、用户账号信息表、用户档案信息表、课程注册表)以及系统日志(用户日志). 将这些数据集集成到一个临时表中, 以进行后续实验.

④数据转换. 把处理后的数据转换为 WEKA 支持的 ARFF 格式的文本数据文件, 以便应用于特定的数据挖掘算法模型.

数据预处理结束后的实验数据如表 1 所示.

表 1 实验数据

数据集	课程资源信息	用户学习课程信息	用户信息
训练集	1294	2596	207
测试集	983	1686	95

3.2 评价指标

本文使用准确率(Precision)和查全率(Recall)这两个指标对实验结果进行评估. 准确率和查全率是推荐系统中最常用的两个指标^[10,11], 其定义如下:

$$Precision = \frac{tp}{tp + fn} \quad (8)$$

$$Recall = \frac{tp}{tp + fp} \quad (9)$$

在实验中, 推荐的结果被标注为正(positive)或负(negative). 正代表推荐的结果符合用户兴趣, 负说明推荐的结果不符合用户兴趣. tp 代表真正, fp 为假正, fn 为假负. 本文设置了实验中的最小支持度和置信度分别设置为 0.25 和 0.5.

3.3 性能评估

为了对本文提出的学习系统进行评估, 本文将提出的多维度推荐(简称为 MDR)与传统的基于 Apriori 的推荐(简称为 ABR)进行了比较.

从图 2 可以看出, 随着推荐数量的增加, 准确率呈下降趋势, 而 MDR 的准确率一直高于 ABR. 这是由于随着推荐的课程数量增多, 引入了较多用户兴趣值较低的课程导致的. 由于某些情况下准确率和召回率是互相矛盾的. 为了综合考虑这两个指标, 本文使用 F1-measure:

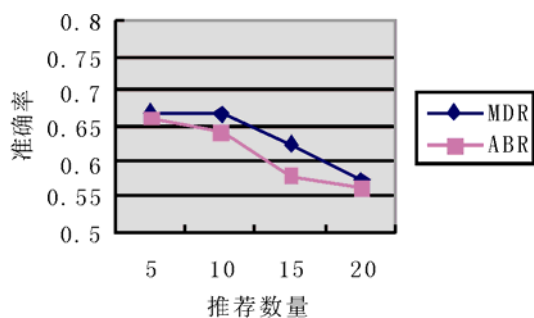


图2 准确率比较

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

MDR与ABR的F1-measure的比较如图3所示。

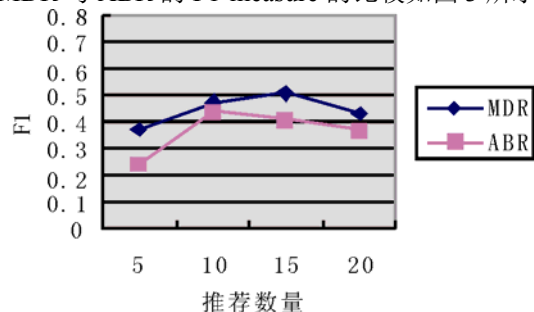


图3 F1-measure 比较

从图3可以看出，随着推荐数量的增加，F1整体呈上升趋势。而MDR的F1一直高于ABR。

4 结论

利用多维网络能够从用户行为数据中提取出有价值的关联关系，使在线学习系统的推荐更具有主动性。相比于传统的Apriori关联规则挖掘，本方法抓住了课程与用户、用户与用户、课程与课程的关联关系，不仅保证了推荐的准确性和良好的用户体验，而且显著地提高了智慧在线学习系统的效率。

如何挖掘知识网络改善推荐效果是本文接下来的改进方向。同时，在进行实验时仍然存在一些缺陷，如阈值的设定、模型检验部分，需要大量的调试对比和定量分析才能得出理想的结果。

参考文献

- 1 Aher SB, Lobo L. Data preparation strategy in e-learning system using association rule algorithm. *International Journal of Computer Applications*, 2012, 41(3): 35–40.
- 2 Aher SB, Lobo L. Prediction of course selection in e-learning system using combined approach of unsupervised learning algorithm and association rule. In: Vinu VD, Yogesh C, eds. *Mobile Communication and Power Engineering*. Berlin: Springer Berlin Heidelberg, 2013: 149–154.
- 3 Monachesi P, Markus T, Posea V, Trausan-Matu S, Osenova P, Simov K. Supporting knowledge discovery in an e-learning environment having social components. In: Elleithy K, Sobh T, Iskander M, Kapila V, A. Karim M, Mahmood A, eds. *Technological Developments in Networking, Education and Automation*. Berlin: Springer Netherlands, 2010: 157–162.
- 4 Aher SB. EM&AA: An Algorithm for Predicting the Course Selection by Student in e-Learning Using Data Mining Techniques. *Journal of the Institution of Engineers*, 2014, 95(1): 43–54.
- 5 Wu M, Zhao Y, Wang C. Knowledge-based approach to assembly sequence planning for wind-driven generator. *Mathematical Problems in Engineering*, 2013(2013): 1–7.
- 6 Murdock JL, Williams AM. Creating an online learning community: is it possible? *Innovative Higher Education*, 2011(5): 305–315.
- 7 El Ahrache SI, Badir H, Tabaa Y, Medouri A. Massive open online courses: a new dawn for higher education? *International Journal on Computer Science and Engineering*, 2013, 5(5): 323–327.
- 8 Z Jianmin, X Aichun, L Haixiang, C Yongliang. An automatic engineering configuration system for substation automation based on SVG/XML/CIM. *Automation of Electric Power Systems*, 2004, 28(14): 53–56.
- 9 Liang TP, Yang YF, Chen DN, Ku YC. A semantic-expansion approach to personalized knowledge recommendation. *Decision Support Systems*, 2007, 45(3): 401–412.
- 10 Salehi M, Kamalabadi IN, Ghouschi MBG. Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering. *Education and Information Technologies*, 2014, 19(4): 713–735.
- 11 Vaithyanathan V, Rajeswari K, Phalnikar R, Tonge S. Improved Apriori algorithm based on selection criterion. 2012 IEEE International Conference on Computational Intelligence & Computing Research (ICIC). IEEE, 2012: 1–4.