

面向知识推荐服务的选课决策

张维国

(南京旅游职业学院 南京 211100)

摘 要 面对 Internet 的快速发展以及 Web 上的海量信息资源,用户如何快速并准确地定位到需要的信息成为了一个亟待解决的问题,面向知识推荐服务的选课决策由此产生。面向知识推荐服务的选课是基于个性化推荐研究的核心问题,其以关联规则 Apriori 算法的理论为基础,运用协同过滤推荐算法,对 Apriori 算法进行改进,结合学生的专业、兴趣爱好、学习成绩,构建了选课推荐系统模型和基于该模型的个性化推荐算法分析,通过对学生成绩数据库中的数据进行挖掘,指导学生选择更适合自身情况的课程,为学生高效的学习、个性的发展提供帮助。

关键词 知识推荐,数据挖掘,关联规则,推荐服务,推荐引擎

中图法分类号 TP399 **文献标识码** A

Decision Making of Course Selection Oriented by Knowledge Recommendation Service

ZHANG Wei-guo

(Nanjing Institute of Tourism & Hospitality, Nanjing 211100, China)

Abstract Facing the rapid development of the Internet and the massive information resources on the Web, it is urgent to enable users to quickly find the information they want, hence the course selection oriented to knowledge recommendation service is generated. Course selection oriented to knowledge recommendation service is the core issue in the research of personalized recommendation. Based on the theory of the Apriori algorithm of association rules, this method makes use of the traditional collaborative filtering recommendation algorithm to improve Apriori algorithm. Combined with students' majors, hobbies and academic records, constructs the model of course recommendation system as well as the personalized recommendation algorithm analysis based on this model. Through data mining in the students' academic record database, it guides students to choose more suitable courses and helps them to learn efficiently and develop with personal characteristics.

Keywords Knowledge recommendation, Data mining, Association rules, Recommendation service, Recommendation engine

随着信息技术的飞速发展,运用其生产和收集数据的能力有很大的提高。“人们被数据淹没,人们却饥饿于知识”也是高校面临的一个新问题。知识推荐服务强调服务工作的个性化、专业化,服务过程强调知识增值、服务增值。学生在网上选课时,需要对课程的难易度、专业的关联度、兴趣度等进行比较与分析,从而做出选择,因此课程选择决策过程是网上选课的决定性环节。在高校学生网上选课的个性化服务方面,数据挖掘技术有着十分广泛的应用前景,能有效避免学生选课的盲目性、从众性,指导学生进行科学合理的选课。

1 面向知识的数据挖掘架构

面向知识的数据挖掘架构(见图 1)通过挖掘引擎对数据库、数据仓库或其他数据源中的数据进行抽取,从大量数据中识别出有价值的数据进行推荐,可视化用户界面为系统和用户之间提供了交流平台,挖掘系统在对用户提交的挖掘要求及任务进行搜索时通过知识库提供相关知识,同时挖掘的结果通过可视化界面向用户展示或做出解释。面向知识推荐服

务的选课决策针对不同学生的特点提供了不同的选课策略,改变了传统的选课方式,更加注重数据的快速产生、信息采集渠道的多样性,提高了系统推荐的实时性及结果的准确性。

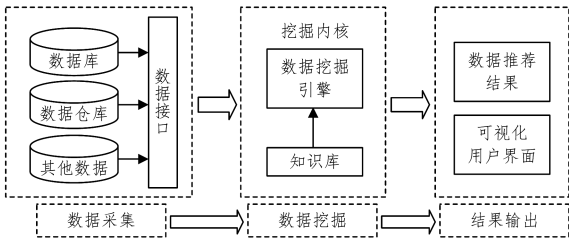


图 1 数据挖掘架构图

2 关联规则算法

关联反映的是一个事件和其他事件之间依赖或关联的知识。数据挖掘中一类重要的算法则是关联规则算法^[1-4],关联规则挖掘的相关算法包括 Apriori 算法、FP-树频集算法、基于

本文受 2015 年度江苏省高等教育教改立项研究课题(2015JSJG375),全国旅游职业教育教学指导委员会 2015 年科研项目立项课题(LZW201505),南京旅游职业学院科研课题(2016YKT18)资助。

张维国(1978—),男,高级实验师,主要研究方向为数据挖掘、数据库技术、计算机网络,E-mail:Jamesnj@163.com。

划分的算法。其中最有力影响力的要属 Apriori 算法,该算法依赖于频繁项集特性的先验知识(Prior Knowledge),因此被称为 Apriori。Apriori 算法使用逐层搜索的迭代方法寻找频繁项集,其优点是算法结构简明,易于理解,理论的推导不复杂,性能很好,容易实现。

2.1 Apriori 算法的缺点

通过 Aprior 算法可以发现关联规则是一个两步的过程:1)利用迭代检索出事务数据库中所有的频繁项集,即支持度不能比用户设定阈值低的项集;2)用频繁项集构建满足用户最小信任度的规则。挖掘或识别出的所有频繁项集占整个计算量的大部分,也是本算法的核心,其缺点有以下几点:

- 1)对数据库进行扫描的次数过多;
- 2)Apriori 算法会产生大量的中间项集;
- 3)采用唯一支持度;
- 4)算法的适应面窄。

2.2 改进 Apriori 算法的描述

- 1)通过扫描事务数据库构建出矩阵。
- 2)由最小支持度 $minsup$ 生成频繁 1 的项集 L_1 ,采用矩阵压缩技术删除相应的列和行,对矩阵进行压缩。
- 3)由频繁 1-项集进行自连接 $L_1 \circ L_1$,生成 2 项候选项集 C_2 。新压缩矩阵中, C_2 的每个候选项集支持度采用向量内积法进行计数,满足最小支持度 $minsup$ 的项集,组成频繁 2-项集 L_2 。
- 4)根据减枝步的改进内容,对 L_2 中单个项目 I_i 进行计数,删除计数小于 2 的项目对应的项集,产生 L_2' ,并删除该项目对应的列,运用矩阵压缩的技术再次实行压缩矩阵。
- 5)由 L_2' 自连接 $L_2' \circ L_2'$,并进行减枝,生成 3 项候选项集 C_3 ,在新压缩矩阵中, C_3 的每个候选项集支持度采用向量内积法来计数,满足最小支持度 $minsup$ 的项集,组成了频繁 3 项集 L_3 。
- 6)在 k 维频繁项集搜索中,依据步骤 4)和步骤 5)两步的思想同理查找,直到 L_k 中的元素个数小于 $(k+1)$,且不能产生 $k+1$ 维候选项集 C_{k+1} 。

3 关联推荐引擎的设计

3.1 关联推荐引擎的大纲

面向知识推荐服务的选课数据库中可以用来关联推荐分析的属性如表 1 所列。

表 1 用作关联推荐的属性	
字段	属性
姓名	用户的属性
性别	用户的属性
年龄	用户的属性
出生日期	用户的属性
院系	用户的属性
学时	课程的属性
学分	课程的属性
是否必修课	课程的属性
任课教师	课程的属性
课程性质	课程的属性
选课次数排名	其他的属性
课程间的关联度	其他的属性

设计关联推荐首先需要发现可用作关联推荐分析的历史数据属性集,这类属性集必须同时满足 2 个要求:必须是系统中存在的字段或数据内容;必须对用户选课存在某

种直接或间接的关系。

找到数据集后,再加入课程权值的概念,用于人为对推荐课程进行影响,增加对推荐课程的可控性。课程关联推荐算法如图 2 所示。

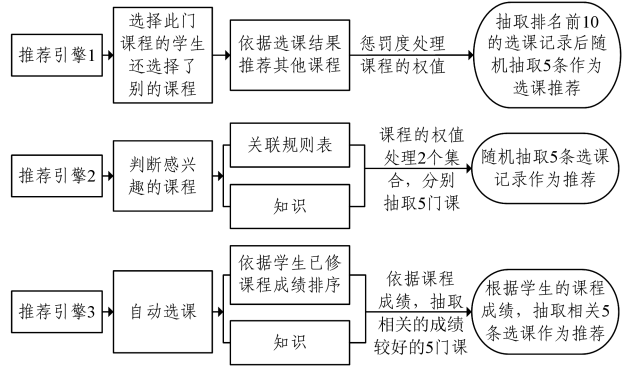


图 2 课程关联推荐算法

面向知识推荐服务的选课关联推荐设计分为 3 个:推荐引擎 1 “选择此门课程的学生还选择了别的课程”。通过计算选了此门课程的学生所选课程次数 * 课程的惩罚度(根据全体学生选课次数的平方根的倒数得到),然后选取前 10 条记录,随机抽取其中的 5 条作为推荐。

推荐引擎 2 “判断感兴趣的课程”。统计设计 8 条可用的关联规则,从 8 条关联规则中分别抽取课程后,随机抽取 5 门课程作为推荐。

推荐引擎 3 “自助选课”。对学生所修课程的成绩进行由高分到低分排序,并抽取与课程成绩较好的课程相关的 5 门课程作为推荐。

3.2 推荐引擎 1 的设计

在学生选课成功之后,根据所选的课程计算选这门课程的学生中,其他课程被选次数的排行。这样的排行是与关联挖掘中 Apriori 算法按支持度和置信度排行是一样的,学生选课次数越多,支持度就越高,置信度也越高,但 Apriori 算法的缺陷在推荐算法 1 中也会存在^[5-9]。

3.2.1 Apriori“支持-置信”的缺陷分析

Apriori“支持-置信”存在很多缺陷,一个缺陷是设定支持度阈值;若最小支持度阈值定得较低,则许多无意义规则充斥着挖掘过程,挖掘的效率及可用性会降低;若最小支持度阈值定得较高,则速度得到提高,有意义的规则会不易发现。另一个缺陷体现在置信度上:在课程选课次数超出支持度阈值的情况下,系统倾向推荐课程选择次数多的。如在某家网上超市的商品推荐栏中发现了超低价促销的饮料时,不管顾客购买食品类、服装类、电器类,均被推荐饮料;当促销方便面时,推荐栏中经常会连续出现各类品牌的方便面,这些推荐造成了关联推荐的冗余,未能给顾客很好的体验,顾客也会将该行为误解为变相做广告。通过引入惩罚度对被选次数多的课程进行减值计算,可以减少上述的冗余关联推荐,大幅度提高关联推荐的价值。

3.2.2 推荐引擎 1 算法

面向知识推荐服务的选课决策系统针对 Apriori 关联算法的缺陷,在推荐引擎 1 中引用惩罚度概念来重新计算课程选课次数,重新排列。学生选课后,系统后弹出窗口推荐课程,根据式(1)的结果进行排序,排出 10 门课程并随机抽取 5 门课程作为推荐。惩罚度由学生的课程选课次数的平方根的倒数得到。数据库中存储着每门课程的课程名称,全体学生选

课次数、惩罚度的倒数、课程推荐的权值。

(与刚刚所选课程关联的课程的选择次数)×惩罚度×课程的权值

(1)

推荐引擎 1 算法的流程如图 3 所示,引入惩罚度后学生选课的推荐更为合理,大部分的推荐结果等同于热门课程推荐(即选课最多的课程,这部分为推荐引擎 2 的子集),且不管选什么课程,推荐的结果都几乎没有变化,这不利于推荐引擎为用户找到感兴趣的课程。

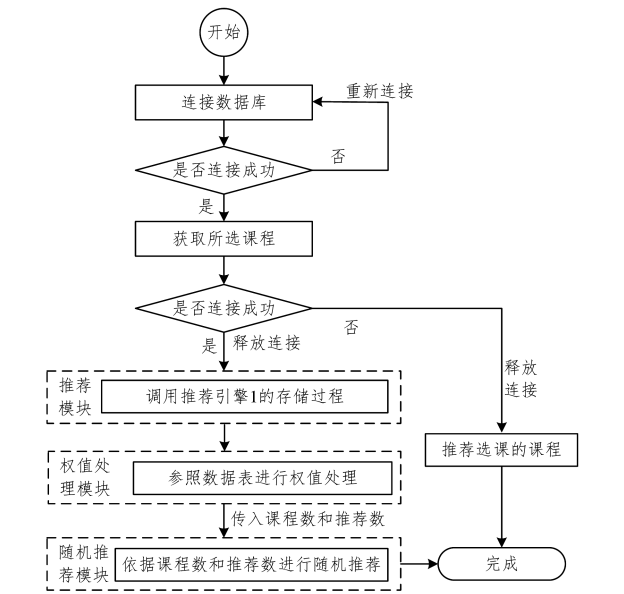


图 3 推荐引擎 1 算法

3.3 推荐引擎 2 算法

推荐引擎 2 用于判断学生感兴趣的课程,其设定的关联条件必须满足:

- 1)数据库中 存在可进行关联挖掘的字段,即存在于表 2 所列出的属性中。
- 2)关联挖掘的字段是对学生选课程有影响,或有关联的。例如,经过 SQL 语句查询出的文科类学生所选课程对比理科类学生所选课程,可以发现查询结果与全体学生选课课程次数的排列几乎相同,这说明文科类或理科类学生对选什么样的课程关联性不大,将这个条件作为关联规则去匹配推荐是无意义的或者说意义不大。若把关联条件改为文科类学生所选课程与理科类学生所选课程来进行关联挖掘,会发现文科类学生所关注的课程与理科类学生所关注的课程差别很大,将这条作为关联推荐 2 的关联条件,就很有意义。

表 2 推荐引擎 2 所用的 9 条规则

规则	是否采用	推荐数
规则 1 男生/女生选课最多的课程推荐	是	2 门课程
规则 2 文科类学生/理科类学生选课最多的课程推荐	是	2 门课程
规则 3 所有学生选课次数最多的热门课程推荐	是	2 门课程
规则 4 根据学生所在院系推荐课程	是	2 门课程
规则 5 根据所选课程的学时推荐学时相同的课程	是	2 门课程
规则 6 根据所选课程的主讲教师推荐主讲教师相同的课程	是	2 门课程
* 规则 7 根据所选课程的分类推荐课程	否	
* 规则 8 根据学生的年龄,查询相同年龄选课最多的课程	否	
* 规则 9 根据所选课程的课件院系推荐相同课件院系制作的课程	否	

推荐引擎 2 的关联规则主要为表 2 中的 9 条(加 * 号的课程虽然是较好的关联条件,但不符合上述 2 个条件,只是简单地列出来,在设计算法时候并不采用,主要关注差异化)。

规则 1 男生/女生选课最多的课程推荐。这里的焦点在于男生非常关注的课程而女生不太关注的课程和女生非常关注的课程而男生却不关注的课程。

规则 2 文科类学生/理科类学生选课最多的课程推荐。这里的焦点在于文科类学生非常关注的课程且理科类学生却不那么关注的课程和理科类学生非常关注的课程且文科类学生却不那么关注的课程。

规则 3 所有学生选课次数最多的热门课程推荐。分别包含选修课和必修课,网站学习的规则必修课是有强制任务的,相比较起来选修课的选择更能体现出学生的意愿。这里多推荐一些选修课能够达到较好的推荐效果。

规则 4 根据学生所在院系的推荐课程。查询学生所在院系全体学生选课次数前 5 位的课程,从这 5 门课程中随机抽取 2 门作为推荐。

规则 5 根据所选课程的学时推荐学时相同的课程。将学时相同的课程作为推荐,随机抽取其中的 2 门作为推荐集合。

规则 6 根据所选课程的主讲教师推荐主讲教师相同的课程。查询所选课程主讲教师相同的课程作为推荐,随机抽取其中的 2 门作为推荐集合。

规则 7 根据所选课程的分类推荐课程。推荐出的课程与推荐 1“选这门课的学生还选择了别的课程”的推荐结果重复,不采用。

规则 8 根据学生的年龄,查询相同年龄选课最多的课程。年龄与课程关联度不高,即按年龄查询课程被选择的次数与不按年龄查询课程选课的次数基本相同。

规则 9 根据所选课程的课件院系推荐相同课件院系制作的课程,数据库中这个字段空值太多。

推荐引擎 2 通过制定的 6 条关联规则,从数据库中根据关联规则取出课程后再随机抽取 5 门课程作为推荐引擎 2 “判断感兴趣的”课程推荐。

3.4 推荐引擎 3 算法

推荐引擎 3 的逻辑结构比较简单,即直接通过数据库查询学生所修课程的成绩,再对成绩进行由高分到低分排序,抽取与课程成绩较好的相关 5 门课程作为课程推荐。

4 推荐选课关联推荐引擎的实现

4.1 关联推荐引擎推荐课程的时机选择

电子商务平台中的推荐方式是客户将某一商品加入购物车后,系统弹出推荐商品。面向知识推荐的选课中,当用户点击选课,选课成功后,系统根据登录用户的信息(如年龄、性别、兴趣爱好、所在院系等)通过关联推荐引擎 1,2,3 向学生推荐选择课程的信息^[10-12]。

4.2 关联推荐引擎实现流程

推荐系统作为一个功能模块,在面向知识推荐服务选课中的一次完整的课程推荐主要分为 7 个步骤,如图 4 所示。

为了保持数据推荐的准确性和算法的快速性,考虑到关联推荐表是使用大量历史数据关联挖掘得到的,其中表数据并不会常常变化,采用数据库 Job 的方式,每天通过 Job 自动更新一次,执行时间定在晚上 24:00,这样的好处是可以降低推荐程序的复杂性,把复杂性交到数据库端。

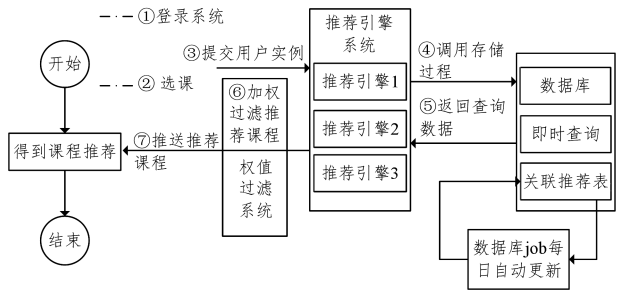


图 4 面向知识推荐服务的选课流程图

结束语 面向知识推荐服务的选课决策系统通过跟踪分析学生的课程学习实情,并针对学生的显式和隐式兴趣、个性特征等信息,对其分别进行构建模型。显式兴趣的模型构建是构建标签特征值库,即对传统协同过滤中基于项目推荐算法的改进。对学生浏览行为进行 Web 日志的序列挖掘来构建隐式兴趣模型,在研究 Apriori 算法的基础上,通过增加用户属性及对剪枝步做出优化来减小数据库规模,提高算法的效率和性能,从而最大程度地给出规则模式。学生兴趣度值由显式兴趣、隐式兴趣的权重采用加权平均法算出,针对学生兴趣构建出模型。

通过引入用户属性和剪枝技术来对 Apriori 算法进行改进,减少了候选集中元素的个数和对数据集的扫描次数,计算的空间复杂度、时间复杂度得到了有效降低。面向知识推荐服务的选课提出了基于学生兴趣的协同过滤推荐算法,有效地解决了学生从传统的零散、随意和功利性的选课方式,向具有针对性、引导性和规范性的选课模式转变,有利于学校推进个性化教育的改革,促进学生个性化的发展,真正达到了学校开设选修课的目的。

参 考 文 献

[1] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]// Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann, 1994: 487-499.

[2] 李贵林, 杨禹琪, 高星, 等. 企业搜索引擎个性化表示与结果排序算法研究[J]. 计算机研究与发展, 2014, 51(1): 206-214.

[3] 许文艳, 刘三阳. 知识库系统的逻辑基础[J]. 计算机学报, 2009, 32(11): 2123-2129.

[4] 钟秀琴, 刘忠, 丁盘苹. 基于混合推理的知识库的构建及其应用研究[J]. 计算机学报, 2012, 35(4): 761-766.

[5] AGRAWAL R, SRIKANT R. Mining sequential patterns[C]// Proceedings of the 11th International Conference on Data Engineering (ICDE'95). Piscataway, NJ: IEEE, 1995: 3-14.

[6] 周水庚, 周傲英, 曹晶, 等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11): 1287-1292.

[7] DAVIDOV D, TSUR O, RAPPOPORT A. Enhanced sentiment le learning using Twitter hashtags and smileys[C]// Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10). Stroudsburg, PA: ACL, 2010: 241-249.

[8] BIFET A, FRANK E. Sentiment knowledge discovery in twitter streaming data[C]// Proceedings of the 13th International Conference on Discovery Science (DS'10). Berlin: Springer, 2010: 1-15.

[9] 贾彩燕, 陆汝钤. 关联规则挖掘的取样误差量化模型和快速估计算法[J]. 计算机学报, 2006, 29(4): 625-634.

[10] 单留举, 王晓东, 马英运. 基于大数据的用户学习偏好建模及应用[J]. 计算机应用与软件, 2016, 33(1): 77-80.

[11] 李肇明. 基于个人兴趣的用户偏好建模[D]. 昆明: 云南大学, 2013.

[12] 刘慧婷, 周开申, 赵鹏. 基于滑动窗口的不确定性数据流频繁项集挖掘算法[J]. 计算机应用与软件, 2015, 32(12): 12-15.

[13] ZHU H, XIONG H, GE Y, et al. Mobile app recommendations with security and privacy awareness[C]// Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining. 2014: 951-960.

[14] TANG L, LIU E Y. Joint User-Entity Representation Learning for Event Recommendation in Social Network[C]// 2017 IEEE 33rd International Conference on Data Engineering (ICDE). IEEE, 2017.

[15] XIA P, ZHAI S, LIU B, et al. Design of reciprocal recommendation systems for online dating[J]. Social Network Analysis & Mining, 2016, 6(1): 32-33.

[16] TATAR A, AMORIM M D D, FDIDA S, et al. A survey on predicting the popularity of web content[J]. Journal of Internet Services and Applications, 2014, 5(1): 8-9.

[17] GUILLE A, FAVRE C. Event Detection, Tracking and Visualization in Twitter: A Mention-anomaly-based Approach[J]. Social Network Analysis & Mining, 2015, 5(1): 18-19.

[18] XU J, ZHONG Y, ZHU W, et al. Trust-based context-aware mobile social network service recommendation [J]. Wuhan University Journal of Natural Sciences, 2017, 22(2): 149-156.

[19] GARG S, GENTRY C, HALEVI S, et al. Functional encryption without obfuscation [M] // Theory of Cryptography. Springer Berlin Heidelberg, 2016.

[20] 任星怡, 宋美娜, 宋俊德. 基于用户签到行为的兴趣点推荐[J]. 计算机学报, 2017, 40(1): 28-51.

[21] LIAN D, ZHENG K, GE Y, et al. GeoMF++: Scalable Location Recommendation via Joint Geographical Modeling and Matrix Factorization[J]. ACM Transactions on Information Systems, 2018, 36(3): 1-29.

[22] LIN H T. A traveling memory based upon location-based services and social network sites[J]. Journal of the Chinese Institute of Engineers, 2015, 38(2): 181-190.

[23] MEMON I, LING C, MAJID A, et al. Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist [J]. Wireless Personal Communications, 2015, 80(4): 1347-1362.

[24] SUN Y, FAN H C, BAKILLAH M, et al. Road-based travel recommendation using geo-tagged images [J]. Computers, Environment and Urban Systems, 2015, 53(9): 110-122.