

K12 学习平台个性化学习资源推荐^①



徐亚军, 郭 俭

(苏州百智通信息技术有限公司, 苏州 215000)

摘 要: 随着在线学习平台的普及, 产生了大量学习行为数据, 如何利用大数据挖掘技术分析在线学习行为, 解决学习者经常面临的“资源过载”和“学习迷航”问题, 更好地实现教学决策、学习过程优化和个性化学习方法推荐等, 已经成为研究重点. 文章基于苏州线上教育中心的学习行为数据, 研究了常用的推荐系统模型, 结合该平台的数据特点, 提出了一种基于知识图谱的协同过滤推荐算法, 利用该算法, 平台推荐的资源准确率超过了 90%, 有效解决了学生“学习迷航”的问题.

关键词: 在线教育; 知识图谱; 协同过滤; 资源推荐

引用格式: 徐亚军, 郭俭. K12 学习平台个性化学习资源推荐. 计算机系统应用, 2020, 29(7): 217-221. <http://www.c-s-a.org.cn/1003-3254/7510.html>

Recommendation of Personalized Learning Resources on K12 Learning Platform

XU Ya-Jun, GUO Jian

(Suzhou Baizhitong Information Technology Co. Ltd., Suzhou 215000, China)

Abstract: With the popularity of online learning platform, a large number of learning behavior data are generated. How to use big data mining technology to analyze online learning behavior, to solve the problem that learners often face “resource overload” and “learning confusion”, better implementation of teaching decision-making, learning process optimization, personalized learning method recommendation, etc., has become a research focus. Based on the learning behavior data of Suzhou online education center, this work studies the common recommendation system model. Combined with the data characteristics of the platform, a collaborative filtering recommendation algorithm based on knowledge map is proposed. With this algorithm, the accuracy of the platform’s recommended resources is more than 90%, which effectively solves the problem of “learning lost” for students.

Key words: online education; knowledge map; collaborative filtering; resource recommendation

1 概述

近年来, 随着在线教育和在线学习平台的不断发展, 积累了海量的学习资源. 这些学习资源极大地丰富了学生的选择, 学生在平台中查阅的资源数量也越来越多, 范围越来越广泛. 然而在为学生带来便利的同时, 也使学生需要花费更多的时间和精力去筛选符合自己学习需求的资源, 因此“资源过载”和“学习迷航”现象越来越严重^[1], 而且平台资源数量众多, 学生无法辨别资源的优劣, 另外, 由于学生的认知能力存在差异, 导致

不同学生的学习需求也是不同的. 因此, 如何能够为学生提供个性化的学习资源, 提高在线课程的学习效率是在线教育面临的一大挑战^[2].

在线学习系统积累了海量的与用户行为相关的数据, 例如资源学习、资源收藏、交流讨论、资源分享等, 隐含了大量有价值的信息. 许多在线学习平台利用数据挖掘和人工智能技术, 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资源^[3].

苏州线上教育中心是 2018 年 1 月苏州市教育局

① 收稿时间: 2019-11-04; 修改时间: 2019-12-09; 采用时间: 2020-01-22; csa 在线出版时间: 2020-07-03

面向 K12 教育领域推出的线上学习平台. 平台以“全名师、全过程、全免费”为核心理念, 统筹全市名优教师资源, 通过个人电脑、手机、平板和电视等终端向全市师生提供“名师优质资源”、“名师在线直播”、“名师在线答疑”、“在线学习行为数据分析与智能引导”等教育服务^[4], 平台上线至今用户使用活跃, 总登录人次已达到 880.1 万, 日均登录 2.1 万人次; 微课观看总数 529 万余人次, 人均观看 29.3 次; 初步形成了常态化的应用局面. 然而, 目前平台学习资源的推荐仅仅是按照浏览量推荐, 推荐的准确性太低.

因此, 本文通过对现有学习推荐系统进行梳理, 针对 K12 在线学习平台的特点, 提出了基于知识图谱的协同过滤推荐方法. 本文基于学生学习资源的数据, 充分利用数据间的关系, 构建知识图谱, 并在其基础上进行协同过滤推荐.

2 相关研究

近年来, 学习资源的推荐逐渐成为在线教育领域的研究热点之一. 陈池等对主流在线教育领域的大数据应用进行了研究总结和归纳^[5], 介绍了数据挖掘、学习分析和知识图谱等相关技术, 并且描述了大数据应用的基本轮廓, 为在线教育领域大数据的研发起到了指导性的作用, 目前推荐系统的推荐方法一般分为三类: 基于内容的推荐方法^[6]、基于协同过滤的推荐方法^[7-9]和混合推荐方法^[10]. 其中基于协同过滤的推荐算法是应用最广泛的, 并且引入了数据挖掘、机器学习、本体等概念或技术^[11-14]. 目前针对基础教育的推荐系统较少, 而基础教育的用户及数据有鲜明的特点, 采用常用的推荐方法效果不好, 张海东等^[15]使用 TF-IDF 算法提取课程和资源的内容特征, 构建模型并计算资源之间的关系从而达到推荐的目的, 但是太依赖于资源的标签.

本文针对 K12 在线教育平台的数据特点, 采用基于知识图谱的协同过滤推荐方法.

3 基于知识图谱的协同过滤推荐

K12 平台的主要用户-学生的学习具有周期性特点, 基础教育学习平台都同步了学习场景, 随着学习进度的推进, 对某个知识点资源的需求会在几天之内迅速从零到达高峰, 并且又在几天之内迅速恢复到零, 所以知识点对推荐的准确率起到关键作用, 而知识图谱

具有紧密的知识相关性的特点, 因此, 知识图谱在扩展实体信息、强化实体之间的联系上具有天然的优势, 可以为推荐系统提供准确而丰富的参考作用. 在知识图谱中, 实体中包含着丰富的属性信息, 单个实体具有多个父实体, 不是所有的父实体都适合被往下遍历. 而知识图谱的属性信息恰好可以为推荐结果提供精确性、多样性与可解释性.

另外一个学校或者一个班级的教学进度是基本统一的, 所以对于资源的推荐需要针对不同的学生有不同的权重, 这样推荐会更加准确, 协同过滤推荐有按照群体的喜好推荐的特点, 所以本文采用协同过滤的推荐方法.

学习平台资源众多, 单纯的推荐算法筛选计算的资源太多, 严重影响推荐的效率和效果. 基于特征的知识图谱的协同推荐, 通过知识图谱筛选出最近学习的知识点的资源、最近同学学习的资源, 从而大大减少推荐的资源数量, 而且推荐的准确度也会更高. 先知识图谱再协同过滤推荐, 通过知识图谱, 进行数据过滤, 再采用协同过滤推荐方法, 两者相结合, 两者的可用信息可以互补, 知识图谱可以帮助协同过滤推荐摆脱局部极小值; 可以防止协同过滤推荐过拟合; 可以提高协同过滤推荐的泛化能力.

在苏州线上教育中心平台, 老师通过平台上传及共享资源, 学生通过平台学习资源, 平台与教育基础库对接, 采集了老师和学生的基础信息, 本文根据老师和学生的基础信息及学生学习的行为数据构建知识图谱, 结合协同过滤推荐向学生推荐个性化的学习资源. 首先利用知识图谱进行资源筛选, 利用协同过滤推荐算法对筛选的资源进行排序推荐, 给学生推荐个性化高质量的资源.

4 知识图谱构建

在学习平台中, 学生学习资源的需求会随着教学进度周期性的变化, 而老师在这个过程中起着主导作用, 并且对于学生的情况, 老师也是最清楚的, 所以本文根据老师提供的相关经验, 整理出知识图谱的节点及属性, 下文将对节点及属性进行详细说明.

基础教育的知识点间具有较强的逻辑关系, 我们称为“前置后导”关系, 从图 1 中可以看出, 学生学习了“10 以内的数”, 才会学习“10 以内的加减法”或者“20 以内的数”. 所以知识点在资源推荐中很重要. 因此本文将知识点作为知识图谱的实体, 并且将知识点作为固定不变的实体.

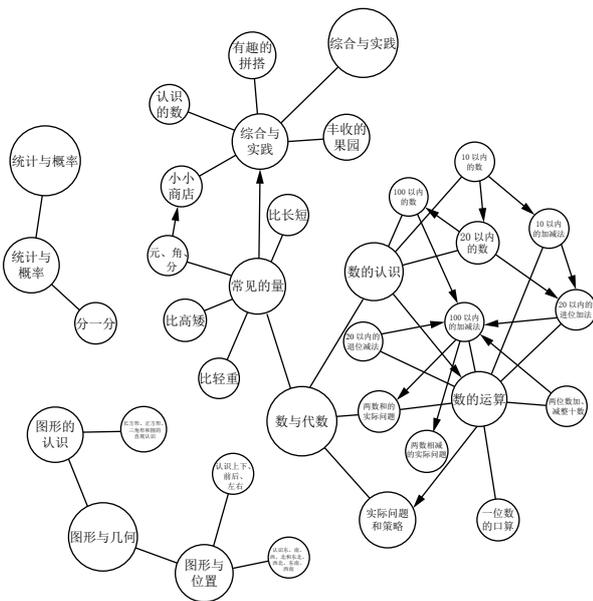


图1 知识点知识图谱

如图1所示,纯知识点构成的知识图谱;另外,资源是推荐的对象,所以将资源作为另一个实体,而知识点作为资源的属性;另外本文也将学生、班级、老师、学校作为知识图谱的实体,这些都是老师总结出的重要属性。

实体及三元组的结构如下:

- 1) 资源: 推荐实体;
< res : 认识11-20, kp, kp : 认识20以内的数 >
- 2) 知识点: 资源的属性, 与教学时间相关;
< student : 学生A, watch, res : 认识11-20 >
- 3) 学生: 推荐主体, 协同过滤推荐用户;
< student : 学生A, class, class : 班级A >
- 4) 班级: 学生属性, 协同过滤推荐参数;
< teacher : 老师A, class, class : 班级A >
- 5) 老师: 班级属性, 协同过滤推荐参数;
< class : 班级A, org, school : 学校A >
- 6) 学校: 班级属性, 协同过滤推荐参数.

综上所述,构建的知识图谱例子如图2所示。

由于K12教育具有周期性教学的特点,学生在某个日期进入平台,平台会根据日期计算出当前学习知识点,本文采用以下算法,将一学年的时间进行分割,按照表1计算一年时间里每天的半径值。

为方便计算,本文将一学年的知识点取值为0-1.0,每个知识点就得到相应的半径值,通过每天的半径值

与知识点的半径值就可以计算出每天对应学习的知识点.本文取日期半径值对应的最近的3个知识点.通过知识点及用户从知识图谱中取相应的实体资源,包括最近知识点的资源,用户观看的资源,同班同学学习的资源,同校同学学习的资源,所属班级老师上传的资源.将这些资源作为协同过滤的输入。

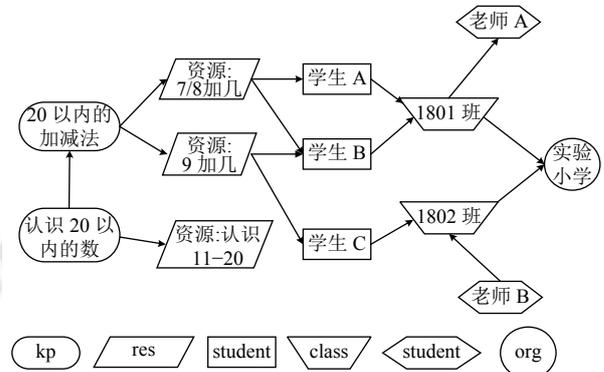


图2 知识图谱

表1 时间取值对照表

时间段	取值期间	说明
寒假	[0,0]	寒假不分日期,不设区间统一为0
春季学期	(0,0.5)	寒假与暑假间的日期,根据一年的天数进行计算,精确到小数后三位
暑假	[0.5,0.5]	暑假不分日期,不设区间,统一为0.5
秋季学期	(0.5,1)	暑假与寒假间的日期,根据一年的天数进行计算,精确到小数后三位

5 协同过滤

经过知识图谱过滤的资源作为协同过滤的样本,推荐的实体是资源,而资源的观看,点赞,收藏作为用户对资源的评分,综合资源的知识点属性,同班同学,同校同学作为不同的权重综合计算出分数作为用户对资源的评价。

5.1 数据预处理

协同过滤推荐需要将学生对资源的打分作为推荐依据,而基础教育平台没有学生对资源的打分系统,而且平台的主要用户是中小学生,他们对资源的打分不完全客观.本文根据用户的行为记录,学生属性等数据计算出评价分数,本文将评价指数,知识点指数,协同指数3个指数作为学生对资源的评价分数。

5.1.1 评价指数

学生在观看资源后没有对资源的直接打分评价,所以本文综合学生对资源的观看,点赞,收藏等行为作为学生对资源的评分,例如如果学生观看视频在很短

的时间内就关掉了视频,很大可能就是视频对学生用途不大,本文将这个行为作为负向评分,如果是负向评分本文就取值为0,另外学生在选择资源的时候带有倾向性,对于自己薄弱的知识点会更多的关注,并且会选择难度较低的资源,对于自己已经掌握的知识点会选择难度较高的资源.另外学生可以对选择的资源进行点赞或者收藏,所以点赞和收藏指标可以很好地表明学生对于资源的喜好,公式如下:

$$score_{evaluate} = f_{view} \times \max(w_{collect} \times f_{collect}, w_{good} \times f_{good}, w_{view} \times f_{view})$$

其中, w 表示各个行为权重, $f_{collect}$ 、 f_{good} 表示是否收藏、点赞, 值分别为1或者0. f_{view} 表示有效观看, 如果观看时长超过了30%, $f_{view}=1$, 否则 $f_{view}=0$.

5.1.2 知识点指数

基础教育平台观看资源的热度与学习进度密切相关, 本文取日期半径值临近的3个知识点的资源, 并计算知识点指数, 公式如下:

$$score_{kp} = 1 - |r_{date} - r_{kp}|$$

其中, r_{date} 表示日期半径, r_{kp} 表示知识点半径.

5.1.3 协同指数

基础教育中, 一个校的教学进度基本相似, 而一个班就是一个教学进度, 所以同班同学或者同校同学的观看资源作为协同指数, 公式如下:

$$score_{collaborative} = \begin{cases} w_{class} \times f_{class} \\ w_{org} \times f_{org} \end{cases}$$

其中, w 表示班级或者学校权重, f 表示是否是同班同学或者同校同学, 其值是1或者0.

综合上面3个指数, 学生对资源的评分公式如下:

$$score = w_{evaluate} \times score_{evaluate} + w_{kp} \times score_{kp} + w_{collaborative} \times score_{collaborative}$$

其中, $score$ 为各个指数的数值, w 为各个指数的权重.

5.2 相似度计算

本文利用用户的协同过滤推荐资源, 所以需要计算用户的相似度, 本文对比常见的几种相似度算法, 选择了斯皮尔曼等级相关系数计算方法.

欧几里德距离是计算空间中两个点的距离, 利用欧式距离定义相似度, 相似度没有考虑重叠数对结果的影响.

皮尔逊相关系数是反映两个变量线性相关程度的统计量, 该方法也没考虑重叠数对结果的影响.

Cosine 相似度通过多维空间两点与所设定的点形成夹角的余弦值计算相似度, 该算法对用户的绝对的数值不敏感. 而本文中, 用户对资源的喜好程度对推荐的准确性很关键, 不同类型的学生对不同类型的资源需求不同, 比如, 对某个知识点掌握较好的学生会对难度较低的资源评价较低而掌握较差的学生对难度低的资源评价较高, 所有 Cosine 相似度不适合.

Tanimoto 系数是 Cosine 相似度的扩展, 它不关心学生对资源的具体评分值是多少, 它关心学生与资源之间是否存在关联关系.

综合本文对比, 本文采用斯皮尔曼等级相关系数. 假设两个随机变量分别为 X 、 Y , 它们的元素个数均为 N , 两个随即变量取的第 i ($1 \leq i \leq N$) 个值分别用 X_i 、 Y_i 表示. 对 X 、 Y 进行排序 (同时为升序或降序), 得到两个元素排行集合 x 、 y , 其中元素 x_i 、 y_i 分别为 X_i 在 X 中的排行以及 Y_i 在 Y 中的排行. 将集合 x 、 y 中的元素对应相减得到一个排行差分集合 d , 其中 $d_i = x_i - y_i$, $1 \leq i \leq N$. 随机变量 X 、 Y 之间的斯皮尔曼等级相关系数可以由 x 、 y 或者 d 计算得到, 其计算方式如下所示:

$$\rho = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)}$$

通过实验证明斯皮尔曼等级相关系数推荐准确率是最高的.

5.3 协同过滤推荐

基于用户的协同过滤推荐主要分为3个步骤:

- (1) 从用户列表中获取当前学生 U_i 最相似的 K 个用户集合 $\{U_1, U_2, \dots, U_k\}$;
- (2) 从这 K 个学生集合排除 U_i 的偏好的资源, 剩下的 $Item$ 集合为 $\{Item_0, Item_1, \dots, Item_n\}$;
- (3) 对 $Item$ 集合里每个 $Item_j$ 计算 U_i 可能偏好的程度值 $pref(U_i, Item_j)$, 并把 $Item$ 按此数值从高到低排序, 前 N 个 $Item$ 推荐给用户 U_i .

偏好程度值 $pref$ 计算公式:

$$pref(U_i, Item_j) = \frac{\sum_{l=1}^k sim(U_l, U_i) \times p(U_i, Item_j)}{\sum_{l=1}^k sim(U_l, U_i)}$$

6 实验

本文采用采用离线测评的方法, 将苏州线上教育

中心运营数据分成训练集合和测试集,将2018年1月到2018年12月的数据构建基于知识点的知识图谱,然后对学生进行协同过滤推荐,将推荐结果与2019年1月到2019年6月的数据进行对比,计算出推荐的准确率和召回率。本文随机选取了使用量较多(观看记录超过1000条)的100位学生对其每周观看的资源进行预测。

$$\text{准确率: } P = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{|R(u)|}$$

$$\text{召回率: } R = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{|T(u)|}$$

$$\text{综合评价: } F\text{-Measure} = \frac{2 \times P \times R}{P + R}$$

实验主要测试协同过滤算法3个指数的权重及其指数下面分别的参数的权重。

综合调整,协同指数的各个参数权重取值如下:

$w_{\text{collect}}=1$ 、 $w_{\text{good}}=0.8$ 、 $w_{\text{view}}=0.7$ 、 $w_{\text{class}}=1$ 、 $w_{\text{org}}=0.92$ 、

$w_{\text{evaluate}}=0.7$ 、 $w_{\text{kp}}=0.11$ 、 $w_{\text{collaborative}}=0.19$ 。

而不同相似度算法的准确率如表2所示,根据不同相似度算法的计算对比,斯皮尔曼等级相关系数的综合准确率最高。

表2 相似度准确率对照表(单位: %)

相似度算法	准确率	召回率	综合
欧几里德距离	75.1	21.5	33.4
皮尔逊相关系数	81.9	23.4	36.4
Cosine 相似度	41.2	11.8	18.3
Tanimoto 系数	55.3	15.4	20.1
斯皮尔曼等级相关系数	90.4	25.8	40.1

7 结束语

本文在苏州线上教育中心学生学习视频的数据基础上,结合数据特点,提出了基于知识图谱的协同过滤推荐方法,实验证明该方法相对于按照热度的推荐方法准确率大幅度提高。

本文提出的方法虽然提高了资源推荐的准确率,但召回率有待提高,本文没有考虑学生的用户画像,对于学生个性化的需求没有考虑全面,需要日后对平台更多的数据进行分析,从而完善算法,提高推荐算法的召回率。

参考文献

- 赵呈领, 陈智慧, 黄志芳. 适应性学习路径推荐算法及应用研究. 中国电化教育, 2015, (8): 85-91. [doi: 10.3969/j.issn.1006-9860.2015.08.014]
- 赵铮, 李振, 周东岱, 等. 智慧学习空间中学习行为分析及

推荐系统研究. 现代教育技术, 2016, 26(1): 100-106. [doi: 10.3969/j.issn.1009-8097.2016.01.015]

- 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 (India): Series B, 2014, 95(1): 43-54. [doi: 10.1007/s40031-014-0074-3]
- 郭俭, 徐亚军. K-12 区域级学习平台个性化推荐系统资源画像模型建设研究. 中国教育信息化, 2019, (12): 77-80.
- 陈池, 王宇鹏, 李超, 等. 面向在线教育领域的大数据研究及应用. 计算机研究与发展, 2014, 51(S1): 67-74.
- Zhao WD, Wu R, Liu HT. Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target. Information Processing & Management, 2016, 52(5): 976-988.
- Bogers T, Van Den Bosch A. Recommending scientific articles using citeulike. Proceedings of 2008 ACM Conference on Recommender Systems. New York, NY, USA, 2008. 287-290.
- Tian G, Jing LP. Recommending scientific articles using bi-relational graph-based iterative RWR. Proceedings of the 7th ACM Conference on Recommender Systems. Hong Kong, China, 2013. 399-402.
- Lai CH, Liu DR, Lin CS. Novel personal and group-based trust models in collaborative filtering for document recommendation. Information Sciences, 2013, 239: 31-49. [doi: 10.1016/j.ins.2013.03.030]
- Sun JS, Ma J, Liu ZY, et al. Leveraging content and connections for scientific article recommendation in social computing contexts. The Computer Journal, 2014, 57(9): 1331-1342. [doi: 10.1093/comjnl/bxt086]
- Liu J, Wang XL, Liu XZ, et al. Analysis and design of personalized recommendation system for university physical education. Proceeding of 2010 International Conference on Networking and Digital Society. Wenzhou, China, 2010. 472-475.
- Souali K, El Afia A, Faizi R. et al. A new recommender system for e-learning environments. Proceedings of 2011 International Conference on Multimedia Computing and Systems. Ouarzazate, Morocco, 2011. 1-4.
- Yang Q, Yuan Y, Sun JL, et al. Semantic P2P-based learning Resources personalized recommendation system design. Proceeding of the Third Pacific-Asia Conference on Circuits, Communications and System. Wuhan, China, 2011. 1-4.
- Vesin B, Ivanović M, Klačnja-Milićević A, et al. Protus 2. 0: Ontology-based semantic recommendation in programming tutoring system. Expert Systems with Applications, 2012, 39(15): 12229-12246.
- 张海东, 倪晚成, 赵美静, 等. 面向基础教育阶段的教学资源推荐系统. 计算机应用, 2014, 34(11): 3353-3356, 3364.