[文章编号] 1004—5856(2018)06—0113—04

面向继续教育学生的图书推荐算法研究

李晓光,孙洪庆,周雪妍

(哈尔滨学院,黑龙江 哈尔滨 150086)

[摘 要] 图书是继续教育过程中不可或缺的、重要的学习资源。如何让图书馆丰富的图书资源更好地为继续教育学生服务,针对其职业(专业)及兴趣爱好进行高质量个性化的图书推荐,有效引导其高效优质阅读,从而提高这一特殊群体的阅读兴趣,培养其终身学习的习惯已成为摆在图书工作者面前的一个课题。文章应用用户活跃度进行数据预处理,进而以用户节点的职业(专业)信息、兴趣爱好(借阅记录、荐购记录和书评)两个主要属性为参数,计算度量继续教育学生用户节点的相似度,并通过相似用户为图书打分的方法,有针对性地为继续教育学生推荐其感兴趣的图书,进而更好地为继续教育学生服务。实验证明,该算法能更精准地向目标借阅者推荐其感兴趣并有利于其职业(专业)发展和终身学习的图书。

[关键词] 继续教育学生;图书推荐;用户相似性;多属性度量
[中图分类号] G203 [文献标识码] A doi:10.3969/j.issn.1004-5856.2018.06.026

继续教育是面向学校教育之后所有社会成员特别是成人的教育活动,是终身学习体系的重要组成部分,是专业进修及普通教育后的教育进阶。这个阶段的教育从年龄上讲大多数学生已步入成年。继续教育过程更主要依靠的是学生自学来汲取对职业(专业)有益的知识。图书已经成为继续教育学生在受教育过程中不可或缺的学习工具。而在图书选择的过程中,绝大多数学生的图书选择存在盲目性、跟从性,图书选择的科学性不高,利用效率低下。如何准确地向继续教育学生进行图书推荐成为科学选择图书和提高图书利用率的关键。[1] 随着大数据技术的广泛应用,针对特定群体的多特征的个性化图书推荐算法也逐步出现,协同过滤推荐算法是其中一种比较成功的推荐算法,但其计算范围过大,算法复杂性高,同时,这一算法没有考虑借阅时段、评价等具体情感因素。[2-4] 比如在自学考试备考过程中,学生借阅图书的真实目的是突击备考;在课程结束前,学生大量借阅与某一课程相关的图书是为了结课论文的写作,这些借阅记录本身无法准确表达读者的真实爱好,因此仅从读者的借阅记录来推荐图书,其准确率并不高。[5-6] 可见,有效的图书推荐需要通过用户活跃程度过滤掉随机因素,挖掘出读者的真实兴趣,还需要综合考虑用户节点的不同属性信息来计算度量用户相似性,进而推荐给目标用户其可能感兴趣的图书。这些用户节点的不同属性主要包括两个方面,一是用户的职业属性;二是用户的兴趣爱好属性,用户兴趣爱好属性的描述依据主要是用户持续感兴趣的类别的新书书目、用户潜在有兴趣的新类别的图书书目和对职业(专业)发展有益的图书书目、图书荐购记录和书评信息等。

一、继续教育学生作为读者用户属性分析

1. 职业(专业)属性

继续教育学生作为特殊的读者群体其显著

[收稿日期] 2018-02-28
[基金项目] 黑龙江省社会科学研究规划项目,项目编号:16XWB01;黑龙江省高等教育科学研究“十三五”规划课题,课题编号:16Q170,16Q172。
[作者简介] 李晓光(1971-),男,高级工程师,哈尔滨学院校办公室主任,主要从事高等教育研究、数据挖掘研究以及党政建设研究;孙洪庆(1960-),男,山东莱州人,哈尔滨学院党委书记,高级政工师,主要从事高等教育管理、执政党建设研究;周雪妍(1981-),女,副教授,哈尔滨工程大学博士后,主要从事高等教育研究、社会计算、情报传播研究。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

特征是具有职业(专业)的固定性。一般而言,继续教育学生在继续教育阶段学习的专业是为其当前从事的工作或意向性工作服务的,这种强目的性使得职业(专业)属性成为面向继续教育学生做为读者区别其他读者的一个显著特征。而这个特征对继续教育学生的图书借阅导向性十分明显,因此,我们将继续教育学生的职业(专业)作为针对这一特殊读者群图书推荐算法的主属性。学生的职业(专业)是其主要特征,其相当一部分图书借阅会与职业(专业)相关。基于此,职业(专业)属性是选择图书的一个潜在属性。根据《国民经济行业分类和代码》可构建职业(专业)分类树(如图 1 所示)。门类代码用一位拉丁字母表示,即用字母 A、B、C……依次代表不同门类;大类代码用两位阿拉伯数字表示,打破门类界限,从 01 开始按顺序编码;中类代码用三位阿拉伯数字表示,前两位为大类代码,第三位为中类顺序代码;小类代码用四位阿拉伯数字表示,前三位为中类代码,第四位为小类顺序代码。记职业(专业)相似度为 $C_s(u,v)$,则在同一小类相似度最高,同一中类次之,不同门类相似度为零。具体表示见图 1。

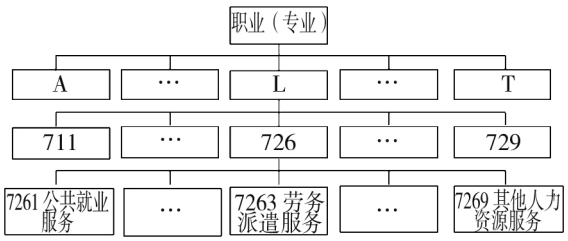

图 1 职业(专业)分类树

$$C_S(u,v)=\begin{cases}0 & u,v\text{ 属于不同门类}\\ 0.3 & u,v\text{ 仅属于相同门类}\\ 0.6 & u,v\text{ 属于相同大类}\\ 0.8 & u,v\text{ 属于相同中类}\\ 1 & u,v\text{ 属于相同小类}\end{cases} \quad (1)$$

2. 兴趣爱好属性

继续教育学生一般不再过多涉猎其他学科,其兴趣爱好相对固定并且一般和专业有关联。在这样的背景下,分析继续教育学生的兴趣爱好属性有利于提高图书推荐的准确性。用户兴趣爱好属性的描述依据除了用户一直感兴趣的类别的新书书目、用户潜在有兴趣的新类别的图书书目和对职业(专业)发展有益的图书书目外,能更大程度反映用户兴趣爱好的是图书荐购记录和书评信息。图书馆对所荐购图书受到广泛欢迎的荐购者进行奖励,因此那些荐购明星的兴趣爱好可以很容易得到,同时,所荐购图书的借阅者们也一定是该兴趣组的成员。同样,对同一本书做过书评的用户我们也认为其具有相同的爱好,其相似程度很高,当然,这里忽略了情感倾向分析,即书评中对图书的评价好坏之分。

定义:u,v 为用户节点,b 表示图书,$B_u(b)$ 为 u 节点借阅过的图书,$J_u(b)$ 为 u 节点所荐购的图书,$C_u(b)$ 为 u 节点所评论过的图书,则用户 u 和 v 的兴趣相似度 $I_s(u,v)$ 可以表示为:

$$I_s(u,v)=\begin{cases}1 & \exists b\in B_u(b)\cap B_v(b)\text{ 满足}(C_u(b)\cap C_v(b))\\ 0.8 & \exists b\in B_u(b)\cap B_v(b)\text{ 满足}(J_u(b)\cap J_v(b))\\ (B_u(b))\cap B_v(b))/(\max)(B_u(b),B_v(b)) & \text{其他}\end{cases} \quad (2)$$

其中,$B_u(b)\cap B_v(b)$ 为用户 u 和 v 共同借阅过的图书数量,如果存在图书 $b\in B_u(b)\cap B_v(b)$ 为两人都评论过的图书,或者为其中一人推荐过的图书,则认为其兴趣相似度较高。

3. 用户活跃度属性

用户活跃度特指在一定时间内图书借阅相对较多的读者,这类用户是图书馆的忠实用户,为其推荐图书更有价值。因此,用户活跃度属性是面向继续教育学生图书推荐算法的一个重要属性,可以用来对原始数据集进行数据清洗,在降低算法复杂度的同时能有效提高推荐精准度。

二、图书推荐策略

基于上文分析,面向继续教育学生作为特殊读者群体的图书推荐需要应用用户活跃度属性进行数据预处理,进而综合考虑读者的职业(专业)相似性和用户兴趣相似性来完成。

1. 用户综合相似性

本文认为向一个爱读书的人推荐图书会有更好的效果。采用职业(专业)度量和兴趣相似性度量相结合的方法来进行。则读者阅读相似度 Y_s 可以表示为:

$$Y_s(u,v)=\alpha C_s(u,v)+\beta I_s(u,v) \quad (3)$$

其中,α 和 β 为调节参数,且 $\alpha+\beta=1$。

2. 用户节点相似性度量算法

输入:每个读者的图书证信息,包括姓名 M、职业(专业)C、班级 G、性别 S,每位读者借阅记录、荐购记录和书评记录等,以及图书馆馆藏图书数据信息。

输出:为每位忠实读者 r 推荐 N 本图书。

步骤 1:借阅记录信息统计分析,数据预处理得到忠实读者 r;

步骤 2:根据式(1)和读者职业(专业)信息计算读者职业(专业)相似性 $C_s(u,v)$;

步骤 3:根据式(2)和读者借阅历史计算读者兴趣相似性 $I_s(u,v)$;

步骤 4:根据式(3)计算读者基于职业(专业)和兴趣的综合相似性;

步骤 5:对于任意忠实读者 r,按其相似度较高用户的借阅图书并集中 top-N 进行推荐。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3. 准确度度量方法

对于本文提取的忠实读者 r，按照用户相似性度量方法进行图书推荐，具体推荐集合为 N 本图书，如果推荐集合中的某本图书出现在了测试数据集里，说明这是一次成功的推荐。其准确度为:

$$P = N_h / N \quad (4)$$

其中，N 为推荐图书总数，N_h 为命中推荐数量。

三、实验与仿真

1. 数据预处理与忠实读者挖掘

数据来源于哈尔滨市图书馆真实数据集，时间跨度 2015 - 03 - 01 到 2015 - 12 - 31，共包含93 142条记录，主要包含借阅、还书、续借、荐购、预约、评论等基本操作。本文采用前 7 个月的数据为实验分析数据并进行图书推荐预测，后 3 个月的数据作为测试数据集。为了分析数据特征，需要统计读者的平均借阅数量等基本信息的分布情况。图 2 为读者借阅图书数量分布图，可见占比 53% 的读者借阅图书在 12 本以内，因此原始数据集中有大量读者并不热衷于图书馆借阅，应用用户活跃度指数有效挖掘图书的忠实读者是数据预处理的重要工作。

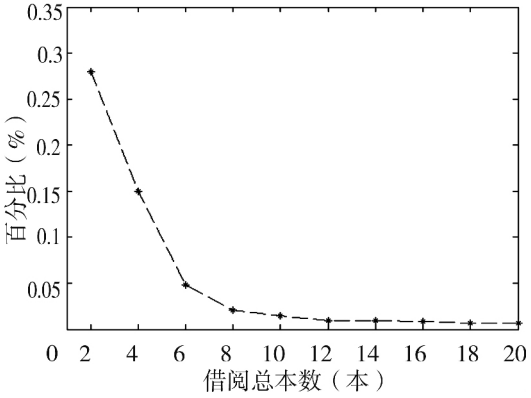

图 2　读者借阅数量统计图

为了更好地推荐图书，对原始数据进行了借阅情况统计(如图 3 所示) 。由图可见，80% 的读者借阅次数在 10 次以内，而 50% 的读者借阅次数不足 3 次，另外很少有人借阅书目超过 80 册。基于统计数据，把借阅次数在 3 到 10 次，且借阅图书总量在 15 到 50 册的读者定义为忠实读者，因为他们具有良好持续的阅读习惯和一定的阅读数量，其推荐效果会显著增强。本数据内可得到这样的读者 3 541 人，设计记录 36 415 条。

2. 读者相似性度量及图书推荐

算法在忠实读者数据范围内来度量节点相似性，采用职业(专业) 度量和兴趣相似性度量相结合的方法来进行。按公式(3) 对读者进行

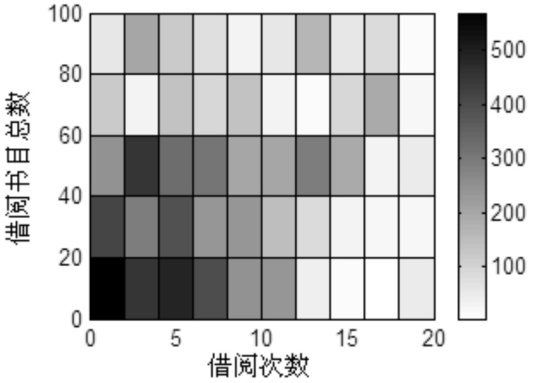

图 3　借阅情况统计

两两交叉匹配可以发现相似度超过 0.2 的节点有 11 032 对，其分布情况如图 4 所示。

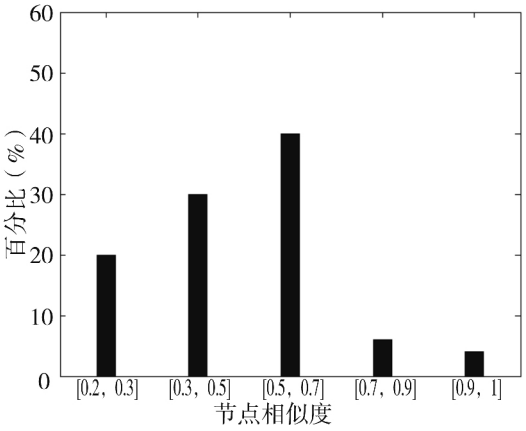

图 4　节点相似度情况统计

可见，大部分用户的相似度在 0.2 到 0.7 之间，只有 11% 的节点相似度超过 70% 。图书推荐的候选数据集为相似度高的节点图书的并集，并结合相似度进行打分的方式进行，推荐图书示意图如图 5 所示。

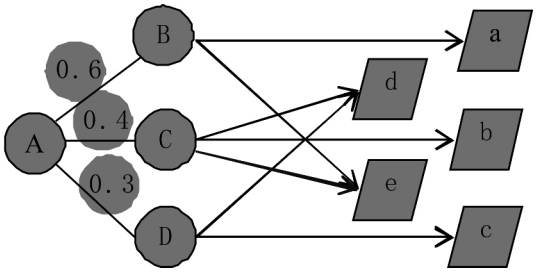

图 5　用户节点 A 推荐图书打分示意图

图 5 中节点 A 有 3 个相似节点分别为 B、C 和 D，对应的相似度分别为 0.6、0.4 和 0.3，箭头关联的图书为对应用户借阅过的图书，则按照打分规则进行推荐，图书 e 得分为 1，最高，图书 e 由 B 和 C 推荐，B 为 e 贡献了 0.6，C 为 e 贡献了 0.4。这种推荐方法既可以包含所有相关图书，又充分考虑了用户相似性，可以在一定程度上提高推荐精准度。本文实验中为每个

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

忠实用户推荐得分最高的 5 本图书。

3. 算法评价

对实验数据集分析后为每位忠实用户推荐 top-5 本图书，并用测试数据集来验证其准确性。对不同的参数值进行分析得到基于参数的推荐精度如图 6 所示。可见，职业(专业) 精度在本数据集中作用较大，在取值 0.6 时达到最高推荐精度，即算法的推荐效果最好。进而，将本文算法与 CBDR 算法[7] 的推荐效果进行对比如图 7 所示。

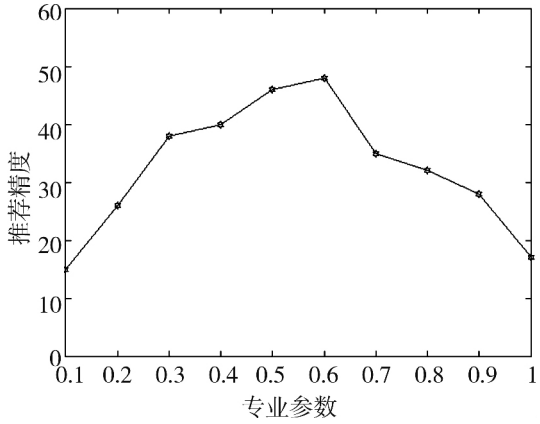

图 6 不同取值对应的推荐精度

其中，相似节点数据为按照节点相似度进行排序后的 top-N 数据，推荐精度随着相似节点的增多而降低，这是因为大量相似度低的节点会对推荐结果产生影响。本文引入用户的多个特征并做综合加权后提出的算法，在相似度节点取值下，推荐精确度较高。

四、结论

本文基于更好地为继续教育学生提供有利其职业(专业) 发展、帮助其养成终身学习习惯的相关图书为宗旨，为提高图书推荐的准确率和有效性而提出一种基于用户节点相似性的图书推荐算法。算法应用用户活跃度属性进行数据预处理，利用用户节点的职业(专业) 信息和兴趣爱好(借阅记录、荐购记录和书评等) 两个基本属性计算度量用户节点的相似度，进而有针对性地为继续教育学生推荐其感兴趣的图书。实验证明，本文算法能更精准地向目标借阅用户推荐其感兴趣的图书、潜在有兴趣的新类别图书和行业内有利于其发展的高质量图书。

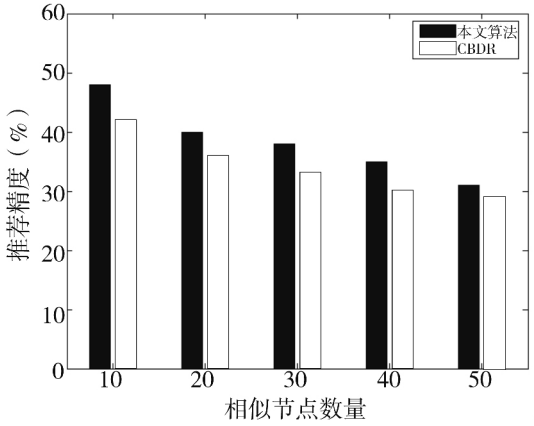

图 7 本文算法与其他算法的推荐精确度比较

参 考 文 献

[1] 李克潮，梁正友. 基于多特征的个性化图书推荐算法 [J]. 计算机工程，2012，(11) .

[2] 马炎. 一种自适应的协作过滤图书推荐系统研究 [J]. 情报杂志，2008，(5) .

[3] 武建伟，俞晓红，陈文清. 基于密度的动态协同过滤图书推荐算法 [J]. 计算机应用研究，2010，(8) .

[4] 王茜，王均波. 一种改进的协同过滤推荐算法 [J]. 计算机科学，2010，(6) .

[5] 田野，祝忠明. 关联数据驱动的数字图书推荐模型 [J]. 图书情报工作，2013，(17) .

[6] Jin S，Fan C，Meng Y，etal. The Design of a New Book Auto Recommendation System Based on Readers' Interest [C]. Electrical and Control Engineering(ICECE) ，2011 International Conference on.

[7] Sohail S S，Siddiqui J，Ali R. Book recommendation System Using Opinion Mining Technique [J]. Proceedings of the IEEE，2013.

责任编辑: 李新红

Arithmetic Study of the Book Recommendation for Adult Education Students

LI Xiao-guang，SUN Hong-qing，ZHOU Xue-yan

(Harbin University，Harbin 150086，China)

Abstract: Books are necessary resources of learning in the process of adult education. It aims to work an efficient way to make the abounding library resources serve the students better. It is suggested that to recommend books according to their specialty and interest and provide guidance of efficient reading. It should be studied how to cultivate students to develop a lifelong learning habit. By analyzing the user's data，the information of the profession and interest is measured to compute the similarities. The score of similarity will be a reference to work out the recommending book list. The experiment shows that it is an effective way to recommend books to the target readers and may contribute to developing their lifelong learning habit.

Key words: the adult education student; book recommendation; the similarity of the users; multi-attribute measurement

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net