Journal of Computer Applications 计算机应用
ISSN 1001-9081
CODEN JYIIDU
2019-10-25
http://www.joca.cn

文章编号:1001-9081(****)**-0000-00
doi:10.11772/j.issn.1001-9081.2019071222

融合知识图谱和协同过滤的学生成绩预测方法

陈曦[1]，梅广[1]，张金金[3]，许维胜[1,2*]

(1.同济大学 电子与信息工程学院，上海 201804；
2.同济大学 信息化办公室，上海 200092；
3.同济大学 教育技术与计算中心，上海 200092)

(*通信作者电子邮箱 xuweisheng@tongji.edu.cn)

摘 要: 针对高等教育本科教学场景中的学生成绩预测问题。提出了一种基于课程知识图谱的预测算法。首先构造了一个表示课程信息的课程知识图谱。然后，分别使用基于邻节点的方法和基于知识图谱表示学习的方法基于知识图谱计算课程在知识层面的相似度，其次将课程的知识相似度集成到传统的成绩预测框架协同过滤中。最后实验对比融合知识图谱的算法和常见成绩预测算法在不同数据稀疏度场景下的性能。实验结果显示在数据稀疏场景下，基于邻节点的算法和传统协同过滤算法相比均方根误差下降约 11%，平均绝对误差下降约 9%；基于图谱表示学习的算法与协同过滤相比均方根误差下降 17.55%，平均绝对误差下降 11.40%。实验结果表明，运用知识图谱的协同过滤算法可使预测误差显著下降，验证了知识图谱可以作为历史数据缺乏场景下的信息补足，从而帮助协同过滤获得更好的预测效果。

关键词: 协同过滤；知识图谱；成绩预测；教育数据挖掘；智慧校园

中图分类号:TP391.1 **文献标志码：A**

Student Grade Prediction Method based on Knowledge Graph and Collaborative Filtering

CHEN Xi[1], MEI Guang[1], ZHANG Jinjin[3], XU Weisheng[1,2*]

(1.College of Electronics and Information, Tongji University, Shanghai 201804, China;
2.Informatics Office, Tongji University, Shanghai 200092, China;
3. Education Technology and Computing Center, Tongji University, Shanghai 200092, China)

Abstract: Focused on the prediction of student grade in undergraduate teaching process of higher education, a prediction algorithm based on the curriculum knowledge graph was proposed. First, a curriculum knowledge graph representing course information was constructed. Then, the neighbor-based method and the representation-learning-based method were used to calculate the similarity of the course based on the knowledge graph, and then those similarities among courses were integrated into traditional performance prediction framework, collaborative filtering. Finally, the performance of the algorithm fusing knowledge graph and common prediction algorithm in different data sparsity were compared in experiments. The experimental results show that in the data sparse scenario, compared with the traditional collaborative filtering algorithm, the root mean square error is reduced by about 11% and the average absolute error is reduced by about 9% by neighbor-based method. And the root mean square error is reduced by about 17.55% and the average absolute error is reduced by about 11.40% by representation-learning-based method. The experimental results indicate that the collaborative filtering algorithm with knowledge graph can significantly reduce the prediction error, which proves that the knowledge graph can be used as information supplement in the lack of historical data, thus helping collaborative filtering to obtain better prediction results.

Keywords: Collaborative Filtering (CF); Knowledge Graph (KG); grade prediction; Educational Data Mining (EDM); intelligent campus;

收稿日期: 2019-07-15； 修回日期:2019-09-06； 录用日期: 2019-09-06。
基金项目：国家自然科学基金资助项目（71540022）

作者简介: 陈曦(1995—)，女，安徽芜湖人，硕士研究生，主要研究方向：数据挖掘、自然语言处理；梅广(1989—)，男，安徽天长人，博士研究生，主要研究方向：教育信息化、数据挖掘、人工智能；张金金(1994—)，女，山东临沂人，工学学士，主要研究方向：智慧教学环境设计、教育信息化；许维胜(1966—)，男，山东临沂人，教授，博士，主要研究方向：智能控制、应急管理、教育信息化。

本刊网络出版时间 yyyy-mm-dd。 本刊网络出版地址;
知网网络出版时间 yyyy-mm-dd hh:mm:ss。 知网网络出版地址。

0 引言

学生成绩预测是教育数据挖掘(Educational Data Mining，EDM)领域的研究热点之一。研究通过对课程设置、学生历史成绩或其他背景数据的分析，预测学生在未来学习阶段的表现。高等教育中日益严重的退学问题使采用更为创新有效的方法促进学生及时毕业已成为迫切需求：文献[1]分析了全美的教育数据，发现在所有 2011 年秋季入学攻读四年制学士学位的学生中，仅有 60%在 6 年内完成了学业。众多教育家认为早期成绩预测是解决该困境的一种实用的方法：文献[2][3][4]都曾通过一系列实验表明早期识别出有退学风险的学生是防止他们辍学的一个关键举措。

随着数据挖掘技术在教育领域的兴起，大量数据挖掘的方法被应用于学生成绩预测的研究中。现有的研究方法可分为两类：一类是将预测问题视为回归或分类问题，应用线性回归[5]、决策树[6][7]、支持向量机[8]、深度神经网络[9]、贝叶斯网络[10]等数据挖掘模型。另一类是将学生预测问题类比成推荐系统中的用户评价问题，借用推荐领域的技术解决问题，包括协同过滤(Collaborative Filtering, CF)、矩阵分解(Matrix Factorization, MF)等方法[11-17]。与基于回归的方法相比，基于推荐的方法因为其较高的预测精度和可解释性得到更为广泛的应用。

但是，基于推荐的方法往往在缺乏历史数据的情况下性能较差。方法主要依赖学生成绩的历史记录挖掘课程的相似性，进而对结果进行预测。在课程的历史选课人数较少时，必须采用额外的信息帮助准确刻画课程之间的相似度。例如学生的知识基础和课程的知识领域。这两者之间的重合度与课程成绩息息相关。如果能够揭示出这种关联，并运用到成绩预测中，预测的精度将有机会得到改善。但在学生成绩预测领域，大多数研究都只利用了与知识信息关联较弱的学生背景信息或课程背景信息，包括学生年级、课程难度、课程学时等。这些背景信息通常类别冗杂，对数据源的要求较高，且对知识信息的挖掘有限。目前为止，还未见依赖知识信息预测学生成绩的研究。

本文研究如何利用课程知识信息对高等教育中本科生在本科学位课程上所取得的成绩进行预测。研究通过 TextRank 算法[18]从课程信息中提取关键字作为知识点，再结合数据库中其他课程信息，构建了基于课程信息的知识图谱来表示课程的知识信息。在知识图谱的帮助下，本文借助节点亲密度算法（Adamic Adar[19], Preferential attachment[20], Resource Allocation[21] ）和知识图谱表示学习算法（Translating Embeddings[22]和 DistMult Model[23]）挖掘课程之间的知识相关性，并比较了它们在传统 CF 框架下的有效性。

本文的主要贡献如下:

1）基于同济大学 2013-2017 年间的本科生课程信息构建了课程知识图谱。

2）提出了一种在 CF 框架下利用课程知识信息进行成绩预测的方法。并利用同济大学的本科生成绩数据验证了方法的有效性。

本文的其余部分组织如下：第 1 节总结了相关领域的研究现状。第 2 节构建了课程知识图谱。第 3 节详细介绍了预测算法。第 4 节描述了数据集和实验结果，同时提供了实验结果的理论分析。第 5 节总结了全文并对未来工作进行展望。

1 研究现状

现有文献表明了基于推荐的算法在成绩预测领域的有效性和利用课程信息建立教育类知识图谱的可行性，为从知识层面发掘课程关系并应用于预测学生成绩提供了理论基础。

1.1 基于推荐算法的成绩预测方法

学生成绩预测问题常与推荐系统中的用户评价问题进行类比，现有的研究也将推荐领域中的相关技术用于预测学生的成绩。文献[11][12]使用 CF 方法预测成绩并证明了 CF 在学生成绩预测上的表现优于传统回归方法。文献[13]以 CF 为底层算法构建了一个选修课推荐系统并应用在中山大学。该应用使得选修课程的退课率大幅下降，进一步证明了 CF 的有效性。文献[14]扩展了传统的推荐算法，利用学生的历史成绩以及乔治梅森大学的各种课程背景资料和学生资料解决成绩预测问题；研究提出了一种混合分解机和随机森林(Factorization Machines with Random Forest，FM-RF)的方法用于准确预测学生在课堂上的表现。文献[15]在 CF 框架下开发了三种融合了时间信息的预测方法，并对明尼苏达大学的学生成绩数据进行了一系列实验，证明了方法的有效性。

上述成绩预测方法大多忽略了学生成绩随着学生努力程度而改变的事实基础。为解决上述问题，一些研究者基于学生的学习过程对成绩的影响提出了动态预测算法：文献[16]使用历史成绩信息和可用的附加信息(如期中考试成绩)来预测学生未来课程的成绩，研究采用 MF 方法并得到了较好的结果；文献[17]在评估学生行为的基础上，提出了一种基于 MF 的动态预测学生学习成绩的方法。

1.2 教育知识图谱的构建和应用

在教育领域，知识图谱也称概念图[24]或领域模型[25]，主要关注包括课程和知识在内的教育实体及其之间的连接关系。挖掘每门课程的关键知识是构建课程知识图谱的过程中必不可少的一步。挖掘关键知识的一类方法是使用关键字提取算法，包括 TextRank[18]和 TF-IDF(Term Frequency - Inverse Document Frequency)[26]。该类方法将课程信息视为普通文献，提取其中的关键字作为关键知识，文献[27]就使用关键字提取方法基于 MOOC 课程信息构建了课程知识图谱。另一种方法是利用实体链接技术识别知识点。例如：文献[28]

利用教学数据和实体识别技术，从 MOOC 平台的课程信息中提取了教学概念；文献[29]提出了一种利用 Web 知识从数字图书中提取概念层次结构的方法，并通过该方法将图书内部的知识与外部的知识资源连接起来；文献[25]提出了一种从电子教材中半自动生成知识模块的框架 DOM-Sortze。

知识图谱特殊的结构为计算节点的相似度提供了可能性。将知识图谱看作由节点和边组成的网络结构，可以使用一些链路预测方法来计算节点间的紧密度，包括 Adamic Adar[19]，Preferential attachment[20]，以及 Resource Allocation[21]。一些知识图谱表示学习算法也可以用于计算每个节点的特征向量，从而计算节点相似度，如 TransE (Translating Embeddings)[22]、DistMult[23]和 ComplEx[30]。

2 课程知识图谱的构建

本节设计了课程知识图谱的结构，并使用 TextRank[18]对知识图谱进行实体提取，以完成图谱的构建。

2.1 知识图谱结构

本研究使用同济大学的课程信息相关数据构建课程知识图谱。通过对数据的分析，本文选取了以下实体：“院系”、“课程”、“知识点”、“教材”、“参考书”和“教学模式”。“院系”实体指开设该课程的机构；“知识点”实体指学生在完成课程后应该掌握的概念或技能；“教学模式”实体指教学过程中所采用的教学方法，如讲课、讨论或实践。图 1 描述了几个实体之间的关系类型：图中节点代表实体，边缘代表实体之间的关系。图 2 展示了部分知识图谱；图谱以《模式识别》、《模式信息处理》及《模式识别及其地学应用》这三门课程为中心发散，展现了三门课程之间的联系。其中圆形节点表示“课程”实体；白色矩形节点表示“知识点”实体；灰色矩形节点表示“院系”实体；实线箭头表示“院系-OFFER-课程”关系；虚线箭头表示“课程-COVER-知识点”关系。可以看出，知识图谱可以直观地反映课程的相关特征以及不同课程之间的联系。

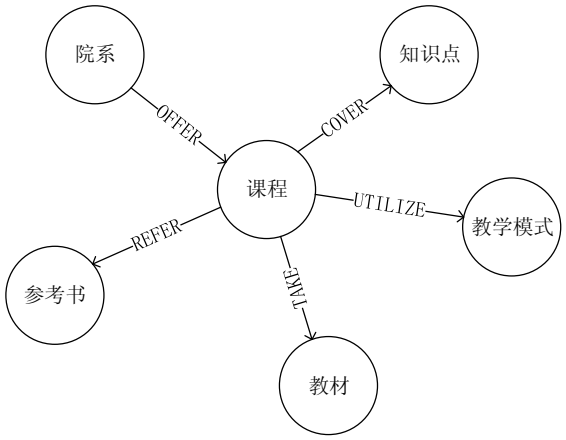

图 1 课程知识图谱的结构

Fig. 1 Structure of course knowledge graph

2.2 课程知识相关性挖掘

课程知识图谱中涉及的大部分实体和关系可以从数据库中获取。而‘知识点’实体需要从课程简介中提取。课程简介的文本往往结构统一，大都包含相似的词汇和句型。除去通用词汇，课程简介主要由专业术语组成。综上，利用简单的关键字提取方法即可提课程简介中所包含的关键知识点。本节使用 TextRank 从中提取“知识点”。

TextRank 的主要思想是建立基于词之间邻接网络，并使用 PageRank[31]计算每个节点的 $Rank$。算法选择 $Rank$ 数值较大的单词作为关键词。首先将给定的课程简介文档 $\boldsymbol{D}$ 分成完整的句子 $[\boldsymbol{S_1},\boldsymbol{S_2},\ldots,\boldsymbol{S_i}]$；对句子 $\boldsymbol{S_i}$ 进行分词和词性标注。分割后从句子中过滤停止词，留下带有指定词性的单词。停止词包含一些常见但无意义的词，如“学时”、“课堂”、“理论”和“大学”。根据上述规则将 $\boldsymbol{S_i}$ 分成一组单词 $[t_{i,1},t_{i,2},\ldots,t_{i,n}]$；其中，$t_{i,n}$ 表示句子中的第 n 个候选单词。算法根据这些单词构建候选关键字网络 $G=(V,E)$；每个候选单词 $t_{i,n}$ 对应一个节点，V 是从所有节点的集合；E 是由代表节点之间共现关系的边组成的集合。共现关系是指一对节点对应的两个词在长度为 K 的文本窗口内共现。在本文中，K 设置为 30。根据式(1)迭代计算各节点的 $Rank(V_i)$ 直到收敛。再选择 $Rank(V_i)$ 的数值较大者作为关键词。

$$Rank(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}Rank(V_j)/|Out(V_j)|; \quad (1)$$

其中，d 为用于平滑的参数，$In(V_i)$ 是 V_i 的前继节点；$Out(V_j)$ 为 V_j 的后继节点。

TextRank 虽然可以有效地从课程简介中提取关键字，但无法识别知识点间的歧义现象。例如，“神经网络”一词在《模

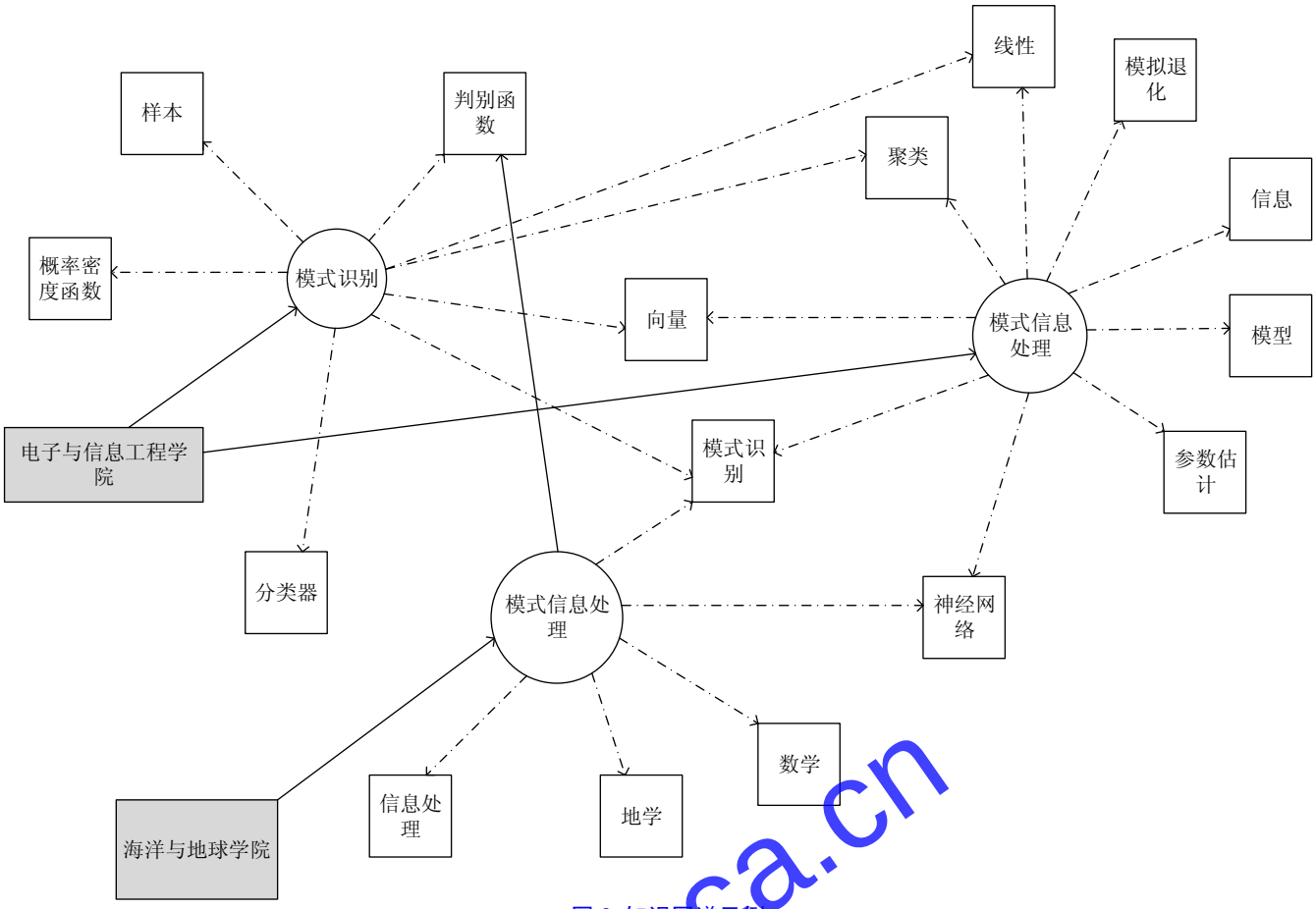

图 2 知识图谱示例

Fig. 2 Sample of knowledge graph

式识别》和《人体解剖学》两门课程中就有不同的含义。由于该词汇在这两篇文档中具有相同的词性和所处语境的相似性，很难将其区分开来。

一个术语的意义取决于它的领域：在本文中，这体现在课程所处“院系”和课程包含的“知识点”这两个实体上；即《模式识别》由“电子与信息工程学院”开设，《人体解剖学》由“医学院”开设，且这两门课程涵盖的“知识点”存在显著差异。对含有相同关键词的课程，比较其所属院系和包含的“知识点”。如果这两门课程来自不同的院系或超过一半的“知识点”是不同的，即认为两门课程不属于同一知识领域，可能在同一个关键词上有不同的含义。考虑到数据库中存在大量的交叉学科课程，本文在上述歧义检测的基础上进行人工确认，从而确保消歧过程的准确性。

3 成绩预测方法

基于已构建的课程知识图谱，分别采用基于邻节点的方法和基于知识图谱表示学习的方法从知识图谱中挖掘课程相似度，该相似度揭露了课程在知识领域的关系。在缺乏历史数据的场景下，课程在知识层面的关联为 CF 框架提供了相似度的计算途径；知识相似度与基于历史纪录的相似度之间互为补充，使得预测结果更接近真实数据。

3.1 算法框架

算法首先生成课程相似度矩阵，再选取 k 个与目标课程相似度最大的课程作为相似课程。学生在相似课程上分数的加权平均值即为学生在目标课程获得的分数。当预测学生 s 在课程 c 上所取得的成绩时，根据课程知识图谱计算 c 和 s 上过的历史课程 $[c_1,c_2,\ldots,c_i]$ 的知识相似度。基于知识相似度筛选出相似度高的 k 个课程。s 在这 k 门课程上所得成绩的加权平均即为目标课程 c 的成绩估计 $est1$，计算加权平均值时以知识相似度为权重。本文也使用基于历史记录的传统 CF 生成估计值 $est2$。对这两种预测模型做线性集成。得到最终预测 $ScoreEst$。图 3 给出了算法流程图。

3.2 相似度计算

3.2.1 基于邻节点的相似度计算

基于邻节点的相似度计算方法将知识图谱看作由节点和边构成的网络，用课程对应节点之间的亲密度衡量课程相似度。本节采用了多种基于邻节点的节点亲密度算法来计算课程间的知识相似度，并按照 3.1 所述与 CF 框架融合。

在基于邻节点的方法中，邻节点的数量对于确定一对节点的相似性起着至关重要的作用。两个节点共享的邻节点越多，关系就越亲密。本文使用了一些经过链路预测领域验证的节点亲密度计算方法。包括 Adamic Adar[19]、共享邻节点数量 (Common Neighbors)、Preferential Attachment[20]、Resource Allocation[21]、同属社区(Same Community)和邻节点总数量 (Total Neighbors)。具体的计算公式如（2）-（6）所示。

对于 Adamic Adar:

$$AA_{sim(x,y)}=\sum_{u\in N(x)\cap N(y)}1/ln|N(u)|\ ; \tag{2}$$

其中 $N(u)$ 表示 u 的邻节点，$|N(u)|$ 表示 $N(u)$ 的节点数量。

对于 Common Neighbors:

$$CN_{sim(x,y)}=|N(x)\cap N(y)|; \tag{3}$$

对于 Preferential Attachment:

$$PA_{sim(x,y)}=|N(x)|\times|N(y)|; \tag{4}$$

对于 Resource Allocation:

$$RA_{sim(x,y)}=\sum_{u\in N(x)\cap N(y)}1/|N(u)|\ ; \tag{5}$$

对于 Total Neighbors:

$$TN_{sim(x,y)}=|N(x)\cup N(y)|; \tag{6}$$

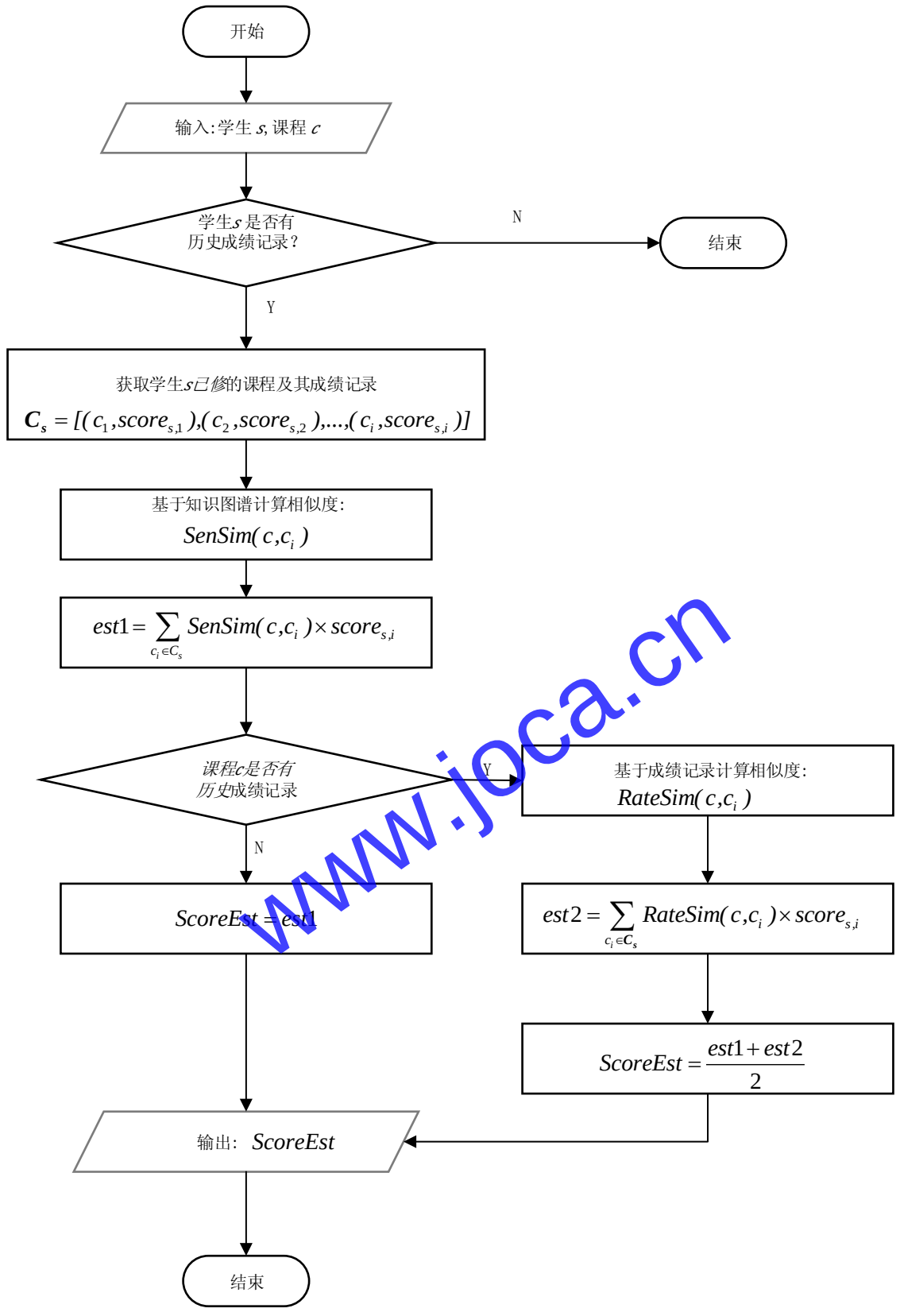

图 3 预测算法流程图

Fig. 3 Flowchart of prediction algorithm

Same Community 是通过确定两个节点是否属于同一社区来决定两节点关系的一种方法。算法将网络划分为不同的社区，值为 0 表示两个节点不在同一个社区中，值为 1 表示同属一个社区。本文定义课程知识图谱中同一个连通域内的点属于同一个社区。

利用上述公式计算课程所对应节点的亲密度，并将其作为课程之间的相似度应用于后续计算中。

3.2.2 基于图谱表示学习的相似度计算

知识图谱的表示学习是一种将知识图谱的实体和关系转化为低维向量的方法。为了将实体和关系嵌入到低维向量空间中，使用三元组的集合表示知识图谱。以本文为例，课程知识图谱 G 可以看作三元组 $(sub, pred, obj)$ 的集合，每个三元组包含一个主体 $sub \in Entity$，一个谓词 $pred \in Relation$，以及一个对象 $obj \in Entity$。$Entity$ 和 $Relation$ 分别是所有实体和关系类型的集合。例如，“课程”实体 C 以及“院系”实体 D 连接形成一个三元组：$(D, offer, C)$。推断三元组中的一对节点在知识图谱的语义上是相似的。有效的知识图谱表示形式应该能够将图谱中存在的三元组(正三元组)和不存在的三元组(负三元组)区分开；即正三元组中，实体所对应的嵌入向量相似，负三元组中，实体所对应的嵌入向量差异大。本文采用了 TransE[22]和 DistMult[23]对图谱进行低维嵌入。得到的嵌入向量之后，使用 Pearson 距离来度量向量之间的相似度，从而得到课程之间的相似矩阵。

TransE 和 Distmult 使用评分函数 $S(t)$ 对正三元组 t^+ 和负三元组 t^- 评分；再通过合适的损失函数尽可能的让负三元组的得分显著低于正三元组。

本文中，TransE 的评分函数采用 L_2 范数：

$$S_{TransE}(t) = -||\boldsymbol{e}_{sub} + \boldsymbol{e}_{pred} + \boldsymbol{e}_{obj}||_2; \quad (7)$$

其中，$\boldsymbol{e}_{sub}$、$\boldsymbol{e}_{pred}$、$\boldsymbol{e}_{obj}$ 分别表示 sub、$pred$、obj 的嵌入向量。

DistMult 模型采用三线性点积作为评分函数:

$$S_{DistMult}(t) = \langle \boldsymbol{e}_{sub} + \boldsymbol{e}_{pred} + \boldsymbol{e}_{obj} \rangle; \quad (8)$$

本文中采用了 pairwise 损失函数和 negative log-likelihood 损失函数训练 TransE 和 DistMult，计算公式如(9)和(10)。

$$L_{pairwise}(G,N) = \sum_{t^+ \in G} \sum_{t^- \in N} max(0, [\gamma + S(t^-) - S(t^+)]); \quad (9)$$

$$L_{NLL}(G,N) = \sum_{t \in G \cup N} ln(1 + exp(-I(t \in G) S(t))); \quad (10)$$

其中，γ 为边缘参数，表示正负三元组的区分度；G 是正三元组的集合，N 是负三元组的集合，由替换正三元组的 sub 或 obj 而生成；$I(.)$ 是指示函数，$I(t \in G)$ 在 $t \in G$ 的时候取 1，其余为 0。

使用上述方法计算得到的 k 维向量表示课程，并生成相似度矩阵。

相对于基于邻节点的方法，基于知识图谱表示学习的方法考虑了不同关系具有的不同意义。例如，来自同一“院系”的课程比只有一个共同“知识点”的课程在知识层面上更相似。但在以邻节点为核心的方法中，相似度只与共同邻节点的数量相关。

4 实验与讨论

为验证学生知识基础和课程知识信息在学生成绩预测中的有效性，本文进行了一系列对比实验来衡量提出的预测算法在不同场景下的预测精度；实验场景包括冷启动问题、数据稀疏场景和数据密集场景。

4.1 实验设置

4.1.3 数据集

实验中采用的数据集是来自同济大学的 1217086 条课程成绩记录。数据集涉及 23903 名本科生和 5378 门课程，涵盖了 2013 年至 2017 年所有在校本科生的课程记录。每一项成绩记录描述了学生、课程以及相应的课程成绩（5 分制）。图 4 描绘了课程数量关于选课人数的分布。图 5 则是学生数量关于选课数量的分布情况。

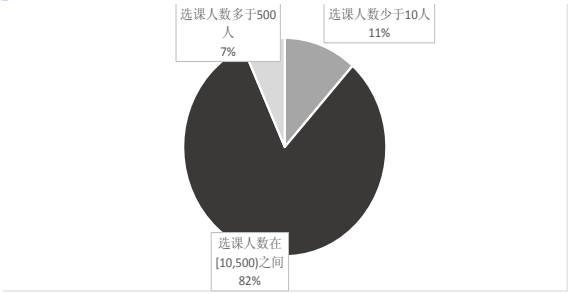

图 4 课程数量关于选课人数的分布

Fig. 4 Distribution of number of courses according to number of students

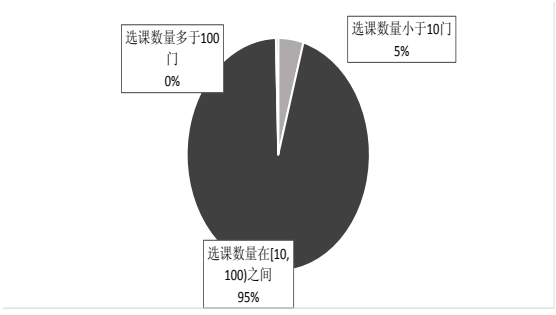

图 5 学生人数关于选课门数的分布

Fig. 4 Distribution of number of students according to number of taken courses

如图 4 所示，大部分的课程选课人数在[10,500)之间。仅有 11.64%的课程选课人数少于 10 人，即 11.64%的课程会出现数据稀疏或冷启动问题。使用传统 CF 对这部分课程进行成绩预测的效果有限。

从图 5 的数据可以得出，95.14%的学生的选课数量在 10 到 100 之间，数据表明大部分学生拥有足够的成绩记录来保证预测的准确性。只有 0.41%的学生选课门数少于 10 门，这些学生多是交流生或联合培养项目参与者，他们的成绩计算方法及成绩记录仍保留在原学校，因此不属于本研究的探究范畴。

在实验过程中，按照 3：1 的比例将数据集划分为已知成绩数据集和测试数据集。实验将记录每一种算法在已知部分成绩数据的基础上预测成绩的误差。

4.1.4 课程知识图谱

本文构建的课程知识图谱组成如表 1-2 所示。

表 1 实体类型及其数量

Tab. 1 Types and number of entities

实体名称	数量
课程	5378
院系	601
教材	2187
参考书	2063
知识点	7779
教学模式	3

表 2 关系类型及其数量

Tab. 2 Types and number of relationships

关系名称	数量
院系-OFFER-课程	5378
课程-COVER-知识点	58939
课程-UTILIZE-教学模式	336
课程-TAKE-教材	2581
课程-REFER-参考书	2063

课程知识图谱共有 69297 个三元组。选取其中的 2000 个作为训练嵌入向量模型的测试集；并通过嵌入向量模型在测试集上的表现评价生成的嵌入向量。

4.1.5 评价指标

实验采用均方根误差(Root Mean Square Error，RMSE)和平均绝对误差(Mean Absolute Error，MAE)两个指标对预测结果进行了评估。计算公式如式(11)-(12)。

$$RMSE(D_{test})=\sqrt{\frac{1}{|D_{test}|}\sum_{y\in D_{test}}(y^{'}-y)^2}\ ; \tag{11}$$

$$MAE(D_{test})=\frac{1}{D_{test}}\sum_{y\in D_{test}}|y^{'}-y|\ ; \tag{12}$$

其中，D_{test} 表示测试集；$y^{'}$ 和 y 分别表示样本的预测结果和该样本的实际得分。

本文使用平均互反排名(Mean Reciprocal Rank, MRR)和 Hit@10 评价嵌入向量的准确性。对于测试集中的每个正三元组，实验通过替换它的主体或对象来生成一系列的负三元组。模型使用得分函数计算这些正负三元组的得分并按得分降序排列三元组；其中正三元组在其生成的一系列负三元组中排名为 $rank(s^+,p,o^+)$。Hit@10 是指排在前 10 位的正三元组的比例。MRR 则按照式（13）计算。

$$MRR(G_{test})=\frac{1}{|G_{test}|}\sum_{i=1}^{|G_{test}|}1/rank(s^+,p,o^+)\ ; \tag{13}$$

其中 G_{test} 表示测试三元组的集合。

MRR 和 Hit@10 的数值越大，说明测试集中排名靠前的正三元组数量越多；即嵌入向量对知识图谱的描述能力越强。

4.1.6 基准

实验采用三种常用成绩预测算法在数据集上的实验结果作为基准。一是基于正态分布的成绩预测方法（Normal Prediction），该方法将所有学生在某一门课程上获得的成绩视为正态分布，通过随机取样获得待预测成绩；二是基于奇异值分解（Singular Value Decomposition，SVD）的矩阵分解方法[32]。三是基于项目的协同过滤方法（Item-Based CF），该方法采用 Pearson 距离来衡量课程之间的相似性。并选取学生在 40 个相似课程上的成绩加权平均值作为预测值。实验通过对比每种预测方法得出的实验结果与三种基准算法中的最优结果，检验知识图谱对预测算法的优化程度。

4.2 基于邻节点的方法

实验融合了传统 CF 和基于邻节点的相似度，并从整个数据集中选取了 3 段具有代表性的数据来检验邻节点法的有效性：1）选课人数少于 10 人的课程；2）选课人数在[10, 500)之间的课程；3）选课人数大于 500 人的课程。场景 1）的课程选课人数较少，冷启动问题和数据稀疏问题较严重；场景 2）几乎不存在冷启动问题，数据稀疏问题有所减轻；场景 3）为数据密集场景。表 3 记录了算法在数据稀疏性不同的场景下的性能。

表 3 基于邻节点的算法多场景下的性能

Tab. 3 Performance of neighbor-based method in multiple scenarios

场景序号	算法名称	RMSE	RMSE 下降率/%	MAE	MAE 下降率/%
场景 1）	Normal Prediction	1.1751		0.9317	
	MF	0.8898		0.6788	
	Item-Based CF	**0.8215**		**0.4159**	
	Same Community	0.7795	5.11	0.3979	4.33
	Adamic Adar	0.7293	11.22	**0.3773**	**9.28**
	Common Neighbor	**0.729**	**11.26**	**0.3773**	**9.28**
	Prefer Attachment	0.8576	-4.39	0.4771	-14.72
	Resource Allocation	0.7298	11.16	0.3781	9.09
	Total Neighbors	0.8459	-2.97	0.4702	-13.06
场景 2）	Normal Prediction	0.9782		0.8218	
	MF	0.7378		0.4002	
	Item-Based CF	**0.6884**		**0.3515**	
	Same Community	0.6519	5.30	0.3331	5.23
	Adamic Adar	0.6266	8.98	0.3186	9.36
	Common Neighbor	0.6259	9.08	0.3183	9.45
	Prefer Attachment	0.73	-6.04	0.3977	-13.14
	Resource Allocation	0.6299	8.50	0.3214	8.56
	Total Neighbors	0.7205	-4.66	0.3926	-11.69
场景 3）	Normal Prediction	0.8873		0.7906	
	MF	0.6818		0.4176	
	Item-Based CF	**0.5497**		**0.3412**	
	Same Community	0.5842	-6.28	0.3843	-12.63
	Adamic Adar	**0.5296**	**3.66**	**0.3314**	**2.87**
	Common Neighbor	0.5316	3.29	0.3321	2.67
	Prefer Attachment	0.6018	-9.48	0.3676	-7.74
	Resource Allocation	0.5935	-7.97	0.3606	-5.69
	Total Neighbors	0.5518	-0.38	0.3396	0.47

从实验结果可得，在冷启动和数据稀疏的场景下，基于邻节点的方法显著降低了预测误差。表 3 中场景 1）数据显示，在数据稀疏场景下，Resource Allocation, Adamic Adar, 和 Common Neighbor 与结果最优的基准算法相比都在 RMSE 下降了超过 10%，在 MAE 上的下降约 9%。此外，场景 2）和场景 3）的数据表明，知识图谱在选课人数较多的情况下仍对预测结果有改善。对于选课人数在[10, 500)之间的课程，与 Item-based CF 相比，基于邻节点的方法使 RMSE 和 MAE 分别下降了 9%；对于选课人数大于 500 人的课程，性能最优的算法 Adamic Adar 与 Item-Based CF 相比 RMSE 下降了 3.66%、MAE 下降了 2.87%。综合表 3 的数据，可以发现传统 CF 的性能随着历史数据的丰富而逐渐变好，而知识图谱的作用随着数据稀疏程度的减弱而减弱。

4.3 基于图谱表示学习的方法

本节在传统 CF 中融合通过 TransE 和 DistMult 计算的相似度。实验首先利用课程知识图谱生成嵌入向量，并用 MRR 和 Hit@10 对嵌入向量进行评价。经过训练和验证，设置嵌入向量的维度为 200；本文使用 Pairwise 损失函数训练 TransE，使用 negative log-likelihood 损失函数训练 DistMult。表 4 给出了两种嵌入向量的详细评价。

表 4 TransE 和 DistMult 的评价

Tab. 4 Evaluation of TransE and DistMult

方法	MRR (训练集/测试集)	Hit@10(%/%) (训练集/测试集)
TransE	0.1962/0.1462	90.00/69.80
DistMult	0.7541/0.4992	98.00/84.65

利用 Pearson 距离计算嵌入向量的相似性。为了验证基于知识图谱表示学习的方法的有效性，从整个数据集中选取与 4.2 同样的三段数据进行实验。表 5 呈现了实验结果。

表 5 基于图谱表示学习的算法在多场景下的性能

Tab. 7 Performance of representation-based method in multiple scenarios

场景序号	算法名称	RMSE	RMSE 下降率/%	MAE	MAE 下降率/%
场景 1）	Normal Prediction	1.1751		0.9317	
	MF	0.8898		0.6788	
	Item-Based CF	**0.8215**		**0.4159**	
	TransE	**0.6773**	**17.55**	**0.3685**	**11.40**
	DistMult	0.7713	6.11	0.4013	3.51
场景 2）	Normal Prediction	0.9782		0.8218	
	MF	0.7378		0.4002	
	Item-Based CF	**0.6884**		**0.3515**	
	TransE	**0.5920**	**14.00**	**0.3126**	**11.07**
	DistMult	0.6559	4.72	0.3319	5.58
场景 3）	Normal Prediction	0.8873		0.7906	
	MF	0.6818		0.4176	
	Item-Based CF	**0.5497**		**0.3412**	
	TransE	**0.5218**	**5.08**	**0.3196**	**6.33**
	DistMult	0.5237	4.73	0.3005	5.98

表 5 说明了基于图谱表示学习的方法对传统 CF 的预测结果有显著的改善。场景 1）数据显示在数据稀疏场景中，性能最优的算法与 Item-Based CF 相比在 RMSE 和 MAE 中分别下降了 17.55%和 11.40%。随着数据的丰富，基于图谱表示学习的方法相比于传统 CF 依然有优势，且在各个场景下的预测性能优于基于邻节点的方法。比较 TransE 和 DistMult，尽管 DistMult 在描述知识图谱(包括 MRR 和 Hit@10)方面的性能优于 TransE。但 TransE 在上述几种情况下的表现都优于 DistMult。

4.4 分析

本研究结果表明，知识图谱可以帮助传统 CF 实现更准确的学生成绩预测。在冷启动和稀疏数据的情况下，基于邻节点的方法和基于图谱表示学习的方法均使 RMSE 和 MAE 显著下降。实验结果显示使用 Adamic Adar、Common Neighbors、Resource Allocation、Same Community、Total Neighbors、TransE 和 DistMult 等算法计算的知识相似性有助于预测结果的改善。这种改善可以归因于知识图谱提供的语义信息。传统 CF 往往通过历史数据评估相似度，不同课程之间对学生能力要求的共性和学科之间思维模式的相通性确保了传统 CF 能够有效地刻画课程间的联系，从而取得不错的预测效果。但在历史记录缺乏的场景下，小数据量不足以支持 CF 准确地刻画课程间的关系。而利用课程知识信息构建的知识图谱可以作为相似度计算的另一种途径；知识图谱更偏重于从教学内容挖掘课程之间的关系，它刻画了不同课程在知识领域上的交集；从学生的先验知识和课程的教学内容出发，提供预测结果。

实验还表明，随着数据稀疏程度的减弱，知识图谱对预测精度的改善逐渐减弱。对此可能的解释是信息的冗余。以往的文献都证明了 CF 在历史评分数据充足的场景下可以有效发掘课程之间的关联，这种关联既包括学科之间逻辑思维层面的相通性，又涵盖了知识层面的共同性。在历史数据不足的情况下，知识图谱提供的信息揭露了课程在知识层面的交叉，从而有助于表示课程关系，帮助 CF 框架更好地预测成绩。在密集数据情况下，知识图谱所包含的信息和历史数据本身发生冗余；因此，在历史数据密集的场景下，知识图谱对预测性能的提升有限。

5 结语

本研究通过结合关键字提取算法和消歧方法构建了一个课程知识图谱模型；并从图谱结构和语义信息两个角度出发，分别使用基于邻节点的方法和基于图谱表示学习的方法发掘了课程在知识层面的关系；本文随后对其在学生成绩预测中的应用进行了探讨。实验表明，知识图谱可以从知识领域的层面有效计算课程相关度；对传统 CF 在历史记录基础上得出的课程关联做了信息补充，对课程关联做了更加完善的刻画，从而得到了比传统 CF 更好的预测性能。

本文探索了知识图谱在学生成绩预测中的应用，并证明了其可行性和有效性。与传统的成绩预测研究相比，本文提出的方法融合了学生的知识基础和课程的教学内容，为后续解读预测结果提供了更多角度。

然而在本研究中，知识图谱的结构不够细化，限制了语义信息进一步的挖掘。例如，本文提出的“知识点”实体可以分为几个子类型，如技能、概念、公式和理论。更详细的知识图谱将会暴露更多的语义信息。后续将对知识图谱的结构做进一步的优化。此外，本研究只是将知识图谱与 CF 框架进行了简单的整合，未来的研究可以考虑将知识图谱应用于更多的推荐算法框架，进一步优化预测性能。

参考文献

[1] MCFARLAND J, HUSSAR B, ZHANG J et al. The Condition of Education 2019[EB/OL].[2019-05-01]. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144

[2] GRAYSON A, MILLER H, CLARKE D D. Identifying barriers to help-seeking: a qualitative analysis of students' preparedness to seek help from tutors[J]. British Journal of Guidance and Counselling, 1998, 26(2): 237-253.

[3] ROMERO C, VENTURA S. Educational data mining: A survey from 1995 to 2005[J]. Expert systems with applications, 2007, 33(1): 135-146.

[4] CASTRO F, VELLIDO A, NEBOT A, et al. Applying data mining techniques to e-learning problems[M]//Evolution of teaching and learning paradigms in intelligent environment. Springer, Berlin, Heidelberg, 2007: 183-221.

[5] MEIER Y, XU J, ATAN O, et al. Predicting grades[J]. IEEE Transactions on Signal Processing, 2015, 64(4): 959-972.

[6] MARQUEZ-VERA C, ROMERO C, VENTURA S. Predicting school failure using data mining[C]// Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands: International Educational Data Mining Society, 2011:271-276 .

[7] 刘志妩.基于决策树算法的学生成绩的预测分析[J].计算机应用与软件,2012,29(11):312-314+330.(LIU Z W. Predictive Analysis of Students' Achievements Based on Decision Tree Algorithm [J]. Computer Applications and Software, 2012, 29(11): 312-314+330.)

[8] BURMAN I, SOM S. Predicting Students Academic Performance Using Support Vector Machine[C]//Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI). Piscataway, NJ: IEEE, 2019: 756-759.

[9] CAZAREZ R L U, MARTIN C L. Neural networks for predicting student performance in online education[J]. IEEE Latin America Transactions, 2018, 16(7): 2053-2060.

[10] 黄建明.贝叶斯网络在学生成绩预测中的应用[J].计算机科学,2012,39(S3):280-282.(HUANG J M. Application of Bayesian Network in Student Score Prediction[J]. Computer Science, 2012, 39(S3): 280-282.)

[11] BYDŽOVSKÁ H. Are collaborative filtering methods suitable for student performance prediction?[C]//Proceedings of the 2015 Portuguese Conference on Artificial Intelligence. Berlin: Springer, 2015: 425-430.

[12] BYDŽOVSKÁ H. A Comparative Analysis of Techniques for Predicting Student Performance[C]//Proceedings of the 2016 International Conference on Educational Data Mining. Raleigh, NC, USA: International Educational Data Mining Society, 2016, s.306-311.

[13] HUANG L, WANG C D, CHAO H Y, et al. A Score Prediction Approach for Optional Course Recommendation via Cross-User-Domain Collaborative Filtering[J]. IEEE Access, 2019, 7: 19550-19563.

[14] SWEENEY M, RANGWALA H, Lester J, et al. Next-term student performance prediction: A recommender systems approach[J]. arXiv preprint arXiv:1604.01840, 2016.

[15] ALMUTAIRI F M, SIDIROPOULOS N D, KARYPIS G. Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(5): 729-741.p729-741.

[16] ELBADRAWY A, POLYZOU A, Ren Z, et al. Predicting student performance using personalized analytics[J]. Computer, 2016, 49(4): 61-69.

[17] XU J, MOON K H, VAN DER SCHAAR M. A machine learning approach for tracking and predicting student performance in degree programs[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(5): 742-753.

[18] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. Stroudsburg, PA: Association for Computational Linguistics, 2004: 404-411.

[19] ADAMIC L A, ADAR E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230.

[20] JEONG H, NÉDA Z, BARABÁSI A L. Measuring preferential attachment in evolving networks[J]. EPL (Europhysics Letters), 2003, 61(4): 567.

[21] ZHOU T, LÜ L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[22] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of the 2013 Advances in neural information processing systems. New York: ACM, 2013: 2787-2795.

[23] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.

[24] YANG Y, LIU H, CARBONELL J, et al. Concept graph learning from educational data[C]//Proceedings of the 8th ACM International Conference on Web Search and Data Mining. New York: ACM, 2015: 159-168.

[25] LARRANAGA M, CONDE A, CALVO I, et al. Automatic generation of the domain module from electronic textbooks: method and validation[J]. IEEE transactions on knowledge and data engineering, 2013, 26(1): 69-82.

[26] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

[27] 侯俊萌. 基于 MOOC 的高等教育知识图谱的构建[D].北京：北京邮电大学信息与通信工程学院,2017.(HOU J M. Construction of Higher Education Knowledge Map Based on MOOC. Beijing: Beijing University of Posts and Telecommunications Institute of Information and communication, 2017.)

[28] CHEN P, LU Y, ZHENG V W, et al. KnowEdu: A System to Construct Knowledge Graph for Education[J]. IEEE Access, 2018, 6: 31553-31563.

[29] WANG S, LIANG C, WU Z, et al. Concept hierarchy extraction from textbooks[C]//Proceedings of the 2015 ACM Symposium on Document Engineering. New York: ACM, 2015: 147-156.

[30] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]//Proceedings of the 2016 International Conference on Machine Learning. New York: ACM, 2016: 2071-2080.

[31] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: Bringing order to the web[R]. Brisbane, Australia. Stanford Info-Lab, 1999.

[32] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009 (8): 30-37.

This work is partially supported by the National Natural Science Foundation of China (71540022).

CHEN Xi, born in 1995, M.S. candidate. Her research interests include data mining and natural language processing.
MEI Guang, born in 1989, Ph.D. candidate. His research interests include education informatization, data mining and artificial intelligence.
ZHANG Jinjin, born in 1994, B.E.. Her research interests include design of intelligent teaching environment and education informatization.
XU Weisheng, born in 1966, Ph.D., professor. His research interests include intelligent control, emergency management and education informatization.