知识主题间先序关系挖掘

麻珂欣[1,2]，魏笔凡[1,2]，马杰[1,2]，刘均[1,2]，黄毅[3]，胡珉[3]，冯俊兰[3]
1. 西安交通大学计算机科学与技术学院，陕西 西安 710049；
2. 陕西省天地网技术重点实验室，陕西 西安 710049；
3. 中国移动研究院，北京 100032

摘要
先序关系指知识主题之间学习的先后依赖关系。已有的先序关系挖掘方法大多是流线型的方式，易导致错误累计，且严重依赖可能导致错误先序关系的超链接。为了解决以上问题，先对知识主题间的先序关系进行统计分析，发现了先序关系的不对称性特征；接着提出从文本中挖掘知识主题间的先序关系的端到端先序关系挖掘模型。该模型基于文本中抽取出的术语间上下位关系，计算知识主题的相关术语集间先序关系的不对称性，进而预测知识主题间的先序关系。实验结果表明，该方法具有较优的先序关系抽取性能。

关键词
先序关系；不对称性；端到端模型

中图分类号：TP391.1 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020052

Mining prerequisite relations among learning objects

MA Kexin[1,2], WEI Bifan[1,2], MA Jie[1,2], LIU Jun[1,2], HUANG Yi[3], HU Min[3], FENG Junlan[3]
1. School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
2. Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Technology R&D, Xi'an 710049, China
3. China Mobile Research, Beijing 100032, China

Abstract
Prerequisite relation refers to the learning dependency between learning objects. Most previous works mined prerequisite relations in a pipelined way and heavily relied on hyperlinks, which lead to the accumulation of errors. To address these issues, prerequisite relations among knowledge topics were analyzed, and the asymmetry feature of prerequisite relation was found out. An end-to-end prerequisite relation model for mining prerequisite relations from texts was proposed. Based on the hyponymy relations between terms extracted from texts, this model calculates the asymmetry of prerequisite relation among related terms of learning objects, and then predicts the prerequisite relation betweens learning objects. The experimental results show that the proposed method achieves the state-of-the-art performance.

Key words
prerequisite relation, asymmetry, end-to-end model

1 引言

先序关系指知识主题之间学习的先后依赖顺序，即在学习一个知识主题之前必须先学习其先序知识主题[1-2]。如在“概率论”课程中，学习“联合条件概率”之前要先学习“条件概率”知识主题，“条件概率”是“联合条件概率”的先序。先序关系是导航学习[3-5]、学习计划制定[6]等教育类应用的基础。

已有先序关系挖掘工作均基于学习者行为数据或文本数据挖掘先序关系[7]。学习者行为数据指学习者的点击日志流等行为数据[7-10]，其只能在成熟的课程中获得。因此，此类方法不适用于挖掘新课程领域中的先序关系。相比于学习者行为数据，文本数据更容易获得。虽然近年来有很多从文本中挖掘知识主题间先序关系的方法[1,11-17]，但是此类方法仍然有一些问题需要被解决。

问题一：错误累积。在已有方法中，以简单规则匹配方式确定的相关术语在先序关系挖掘方法中具有重要的作用[1,12,15,17]。此类方法直接确定相关术语，这会导致错误的相关术语无法在后续阶段被修正，进而产生错误的先序结果，即错误累积问题。此类方法以流线型的方式挖掘先序关系。首先根据标题匹配等规则确定相关术语，然后基于超链接挖掘先序关系。相关术语的正确性极大地影响了先序关系的预测结果。在流线型的方法中，相关术语在确定之后，无法再根据结果进行优化。

问题二：严重依赖超链接。大多数已有方法将超链接作为挖掘先序关系的重要特征[1,11-17]。超链接仅能体现两个页面间存在某种关联，不能体现页面间有向的先序关系。以维基百科为例，“条件概率”和“联合条件概率”页面中分别存在指向彼此的超链接，但是不能根据超链接指向来判断知识主题间的先序关系。除此之外，若根据超链接判断先序关系，则在“联合条件概率”的维基百科页面上存在的指向“条件概率”的超链接，将会导致错误的先序关系，即认为“联合条件概率”是“条件概率”的先序，而事实上“条件概率”是“联合条件概率”的先序。因此，在此类方法中，超链接的使用可能会增加挖掘先序关系的难度或导致错误的先序关系结果。

为了解决以上问题，本文提出端到端先序关系挖掘模型。通过对先序关系数据集的分析，发现了先序关系的不对称性特征，即知识主题的相关术语集间的先序关系是不对称的。本文提出的端到端先序关系挖掘模型基于先序关系的不对称性特征来挖掘先序关系，使用文本中抽取出的上下位关系而不是超链接作为判断先序关系不对称性的依据。

端到端先序关系挖掘模型包含两个模块：文本中专业术语与上下位关系抽取模块和先序关系判别模块。文本中专业术语与上下位关系抽取模块可识别文本中有效文本跨距，其将作为候选专业术语，并挖掘句子中专业术语间的上下位关系[18-19]。上下位关系表明了专业术语间从属的学习依赖关系，可体现专业术语间的先序关系。该模块为先序关系的不对称性计算提供了先序关系依据，也避免了依赖超链接导致的错误。先序关系判别模块基于专业术语间的上下位关系计算知识主题的相关术语集间先序关系的不对称性，从而预测知识主题之间的先序关系。本文还提出两种不同的权重策略，以探究不同相关术语对先序关系不对称性的重要性。

2 相关工作

近年来，国内外研究者提出了较多的

先序关系抽取方法。根据挖掘先序关系时所依赖学习资源的不同，这些方法可分为4类：基于学习者行为数据、基于已有先序关系、基于长文本内容、基于网页信息。

（1）基于学习者行为数据

学习者行为数据通常指学习者在学习过程中的行为日志（如观看课程视频的点击日志流）或问答等互动行为[7]。这些行为数据体现了学习者的学习方法与学习者知识储备之间的重要联系。此类方法使用不同模型从学习者的行为数据中挖掘先序关系特征[7,9,20]。Chen W等人[7]通过构建知识状态转移模型来捕获学习者的参与度信息，进而分析学习者的知识状态的转变过程。该方法首先分析学习者的行为数据，如播放、暂停、快进和快退等行为，然后构建学习者行为模型，从这些数据中预测学习者转变到特定知识状态的概率，进而挖掘先序关系。Chaplot D S等人[9]综合考虑文本中概念的共现特征和学习者的行为特征（如课程的参与度以及测评分数），提出一种无监督的学习依赖图构建方法。该方法可以识别任意粒度级别（课程、单元、模块等）之间的学习依赖关系，同时证明了学生的互动行为比文本阅读更易反映学生的学习效果。此类方法不适用于新课程领域。

（2）基于已有先序关系

隐式的先序关系可从显式的关系结构中发现。已有的先序关系可构成先序关系图谱，通过分析该图谱的图特征，可预测知识主题间的先序关系。Liang C等人[21]提出从课程先序关系中恢复概念间先序关系的方法，并指出课程之间的依赖性是由课程内主要概念间的学习依赖关系引起的。该方法从课程的描述文本中抽取出代表该课程的概念集，通过对课程间先序关系以及已有概念间先序关系的分析，根据先序关系的因果性以及稀疏性两个特征构建目标函数，达到预测未知概念间先序关系的目标。Roy S等人[22]假设课程间先序关系已知，且不同的课程间具有部分共同的概念。他们使用主题模型衡量概念对之间的相关性，并根据主题词向量的聚类、稀疏性及简单性等特征训练神经网络，以识别概念之间的先序关系。

（3）基于长文本内容

在非结构化的长文本中，知识主题的分布特征可反映主题间的先序关系[23-25]。基于此，Liu J等人[23]基于从文本中发现的学习依赖关系的两个特征（学习依赖关系的局部性特征及术语分布的非对称性特征）来挖掘知识主题间的学习依赖关系。Adorni G等人[24]挖掘长文本中以线性方式分布的知识主题之间的先序关系，根据术语共现的特征筛选出长文本中可能存在先序关系的知识主题对，并根据知识主题在文本中出现的顺序识别候选知识主题对的先序关系。此类方法只能挖掘文本中以特定方式组织的知识主题间的先序关系。

（4）基于网页信息

开放知识源中的丰富信息为知识主题间先序关系的挖掘提供了极大便利。以维基百科为例，该知识源中的每个知识主题都具有对应的维基百科页面。页面中不仅包含与当前知识主题相关的完备结构化信息，同时存在指向其他相关知识主题页面的链接。主题间的目录层次关系以及链接关系能在一定程度上反映主题间的先序关系。因此，研究者考虑基于维基百科来实现先序关系的挖掘[11-14,16]。Talukdar P和Cohen W[13]通过分析维基百科页面的文本内容、超链接以及页面编辑历史等信息，使用最大熵分类器识别知识主题之间的先序关系。Gasparetti F等人[16]从维基百科的文本、超链接以及目录结构3个层次分别抽取特征，并构建分类器，以识别先序关系。Liang C等人[1]从认知的角度出发，

认为理解知识主题需要学习与该知识主题在同一认知框架中的所有相关概念，并提出仅基于相关概念间超链接的先序关系挖掘方法RefD（reference distance）。该方法考虑了知识主题的相关概念，并根据两个知识主题的相关概念集之间的超链接的差异，判断知识主题间是否存在先序关系。由于RefD可以轻量且高效地抽取出知识主题间的先序关系，其作为一个重要特征被集成到许多监督学习方法[15,17,26]中。但此类方法严重依赖开放知识源中的超链接等结构化信息。一方面，超链接并不能直接反映先序关系的方向；另一方面，此类方法大多基于流线型的方式挖掘先序关系，存在错误累积的问题。

为了使先序关系挖掘方法适用于大多数领域，本文将网页信息作为数据源来挖掘先序关系。不同的是，本文只关注网页信息中的文本内容，避免了严重依赖结构化信息的缺点。本文提出了基于不对称性的端到端先序关系挖掘方法，避免了流线型方法错误累积对先序关系结果的影响。

3 先序关系不对称性特征

通过对先序关系数据集中知识主题间先序关系的分析，发现了先序关系的不对称性特征。学习者在学习新课程的某一知识主题时，为了全面理解该主题的含义，往往需要学习和理解该主题的其他相关术语[27]。知识主题的相关术语指的是有助于学习和理解该知识主题的一些其他概念。给定某课程的两个知识主题，一个主题的大多数相关术语的学习往往依赖另一个知识主题的相关术语的学习，即知识主题的相关术语集之间的先序关系是不对称的。显然，对于知识主题对(t_a,t_b)，如果学习者在学习主题t_b的大多数相关术语之前，需要先学习主题t_a的大多数相关术语，则主题t_a更可能是主题t_b的先序[1]。

如**图1**所示，知识主题“树”的相关术语集和知识主题“堆”的相关术语集之间的先序关系是不对称的。例如，知识主题“树”的相关术语有“二叉树”“二叉搜索树”等可帮助理解“树”的专业术语；“堆”的相关术语有“斐波那契堆”“二叉堆”等可帮助理解“堆”的专业术语。而“树”的大多数相关术语先于“堆”的大多数相关术语进行学习，如“树”的相关术语“二叉搜索树”应该在学习“堆”的相关术语“二叉堆”之前学习。因此，两个知识主题的相关术语集之间存在的大量不对称的先序关系表明，知识主题“树”与知识主题“堆”之间存在先序关系，且“树”是“堆”的先序。显然，相关术语集之间先序关系的不对称性可反映出知识主题之间的先序关系。

为了验证先序关系不对称性的有效性，对CrowdComp数据集[13]中的先序关系样例进行统计分析。首先在知识主题的描述文本中标记相关术语以及术语之间的先序关系；然后，统计分析是否可通过相关术语集之间先序关系的不对称性推断出知识主题之间的先序关系。**图2**为CrowdComp数据集中是否可通过不对称

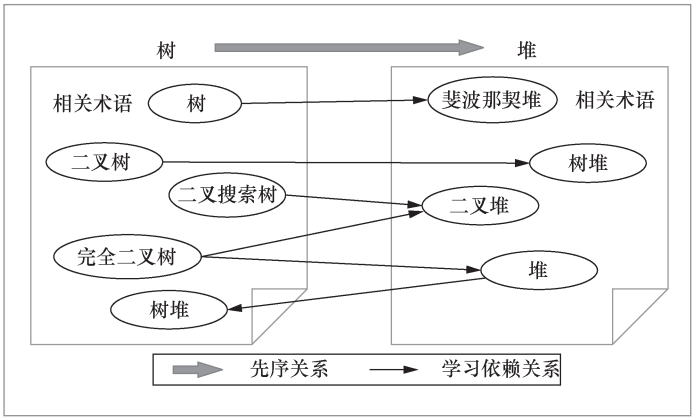

图1 先序关系不对称性实例

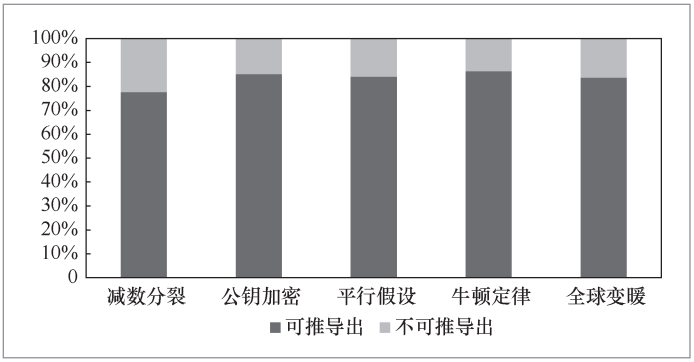

图 2 知识主题间先序关系是否可通过先序关系不对称性特征推导的统计结果

性推断出知识主题间先序关系的统计结果。从图2可以看出，大多数知识主题间的先序关系可通过不对称性推导出。知识主题的相关术语集之间极度不对称的先序关系导致了知识主题之间的先序关系。因此，本文可通过先序关系的不对称性特征有效挖掘知识主题之间的先序关系。

4 先序关系挖掘方法

基于先序关系的不对称性特征，本文提出端到端的先序关系挖掘模型，如图3所示。

对于知识主题对(t_a,t_b)，该模型将对应知识主题的原始文本描述D_a和D_b作为输入，输出一个衡量知识主题t_a和t_b之间先序关系的值v：

$$v=\begin{cases}1, & f(t_a,t_b)\in(\phi,1]\\ 0, & f(t_a,t_b)\in[0,\phi]\end{cases} \quad (1)$$

其中，ϕ为先序关系判断阈值。当v=1时，知识主题t_a是知识主题t_b的先序；当v=0时，知识主题t_a和知识主题t_b间不存在先序关系。整体来说，该模型可细分为两个模块：文本中专业术语与上下位关系抽取模块和先序关系判别模块。

文本中专业术语与上下位关系抽取模块：该模块挖掘文本描述D中术语间的上下位关系。首先，该模块将文本描述D中所有有效的文本跨距作为候选的专业术语；然后，抽取专业术语之间的上下位关系。该模块抽取出的术语间的上下位关系是先序关系判别模块衡量先序关系不对称性的基础。

先序关系判别模块：该模块预测知识主题t_a和t_b之间的先序关系。该模块首先从

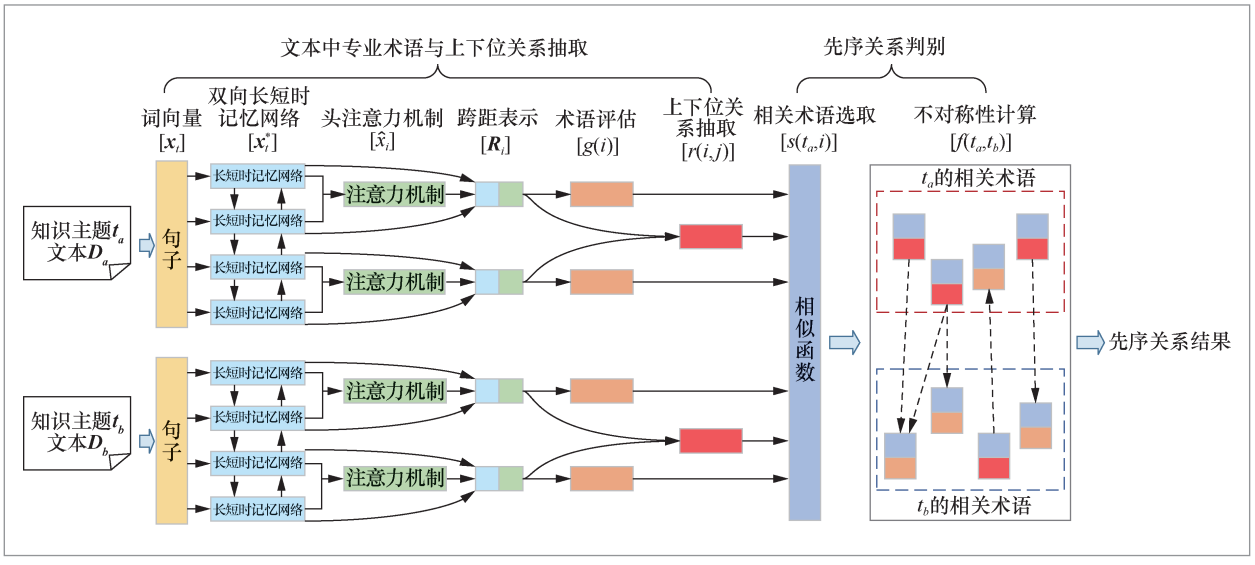

图 3 端到端先序关系挖掘模型框架

候选的专业术语集中识别出知识主题的相关术语，然后基于术语间的上下位关系计算知识主题的相关术语集之间先序关系的不对称性。

4.1 文本中专业术语与上下位关系抽取模块

在衡量先序关系的不对称性时，首先需要识别文本中与特定知识主题相关的专业术语，挖掘每个句子中术语间的上下位关系。将文本描述D中的每一个文本跨距作为候选的专业术语。文本跨距指连续的单词序列，如图4所示，“红”“红黑”“红黑树”均为语句“红黑树是一种自平衡二叉查找树”中的文本跨距。对于文本描述D，每个文本跨距i可用二元组$(i_{\text{start}},i_{\text{end}})$定位，即该文本跨距是从文本描述$D$中的第$i_{\text{start}}$个单词开始，到第$i_{\text{end}}$个单词结束。

该模块包含3个部分：跨距表示、术语评估及上下位关系抽取[18,28]。其中，跨距表示部分将每个语句中可能的专业术语表示为具有一定语义的跨距词向量；术语评估部分根据跨距词向量的语义表征进一步判定其是否为真正的专业术语；上下位关系抽取部分衡量同一语句中的不同专业术语间是否存在上下位关系。

（1）跨距表示

对于文本中的每个单词，用预训练好的ELMo（embeddings from language model）词向量来表征其高层语义，则文本中每个单词的词向量表示为$\{x_1,\cdots,x_t\}$。考虑到语句中的上下文信息，本节采用双向长短时记忆（bi-directional long short-term memory，Bi-LSTM）网络[29]对文本中的每个语句进行重编码，进一步获得单词t在当前语境下的词向量$\boldsymbol{x}_t^*$。

任一文本跨距与其所在语句中的很多其他单词存在语义关联[18]，其中，第一个关联单词称为该文本跨距的语义头单词。文本跨距和其语义头单词之间通常存在上下位关系。为此，本文使用头注意力机制[18]来预测文本跨距i的语义头单词$\hat{x}_i$。具体来说：

$$\beta_t = \text{FFNN}_\beta(\boldsymbol{x}_t^*) \tag{2}$$

$$\alpha_{i,t} = \frac{\exp(\beta_t)}{\sum_{m=i_{\text{start}}}^{i_{\text{end}}}\exp(\beta_m)} \tag{3}$$

$$\hat{x}_i = \sum_{t=i_{\text{start}}}^{i_{\text{end}}}\alpha_{i,t}\boldsymbol{x}_t^* \tag{4}$$

其中，β_t为单词t的得分，$\alpha_{i,t}$为文本跨距i的单词t的概率分布。$\text{FFNN}_\beta(\cdot)$表示前馈神经网络。

在获得每个文本跨距的上下文表征以及语义头单词的词向量之后，将它们聚合，以获得最终文本跨距的词向量$\boldsymbol{R}_i$：

$$\boldsymbol{R}_i = \left[\boldsymbol{x}_{i_{\text{start}}}^*, \boldsymbol{x}_{i_{\text{end}}}^*, \hat{x}_i\right] \tag{5}$$

（2）术语评估

在对每个文本跨距进行语义表征后，需要准确判断该文本跨距是否为专业术语，以达到识别专业术语间是否存在上下

图4 文本跨距实例

位关系的目的。考虑到专业术语的单词数一般不会过长，因此过滤文本中长度大于L个单词的文本跨距。对于剩余的文本跨距i，根据式（6）估算其属于专业术语的得分值$g(i)$。

$$g(i)=\boldsymbol{W}_m\mathrm{FFNN}_m(\boldsymbol{R}_i) \quad (6)$$

其中，$\boldsymbol{W}_m$表示学习的权重矩阵，$\mathrm{FFNN}_m(\cdot)$表示前馈神经网络，m表示术语评估模块。为使本文端到端先序关系抽取模型更加关注有价值的文本跨距，对术语得分值$g(i)$从高到低进行排序，选取得分高的前λT个文本跨距作为专业术语，记作$Y=\{i:g(i)\geqslant\varepsilon\}$，其中，$\varepsilon$表示第$\lambda T$个术语得分值，$\lambda$为保留的文本跨距的比例，$T$为文本描述$D$中包含的单词个数。

（3）上下位关系抽取

给定文本描述D中的任一语句，对于该语句中的文本跨距对(i,j)，当$i\in Y$且$j\in Y$时，文本跨距i与j都被判定为专业术语。在此基础上，通过计算文本跨距对(i,j)的函数值$r(i,j)$来判定是否存在上下位关系，具体如下：

$$r(i,j)=\boldsymbol{W}_r\cdot\mathrm{FFNN}_r([\boldsymbol{R}_i,\boldsymbol{R}_j,\boldsymbol{R}_i\cdot\boldsymbol{R}_j]) \quad (7)$$

其中，$\boldsymbol{W}_r$表示权重参数矩阵，$\mathrm{FFNN}_r(\cdot)$表示前馈神经网络，r表示属于上下位关系抽取模块。通常，上下位关系只存在于有一定语义关联的专业术语之间，且与某一术语存在上下位关系的其他术语是有限的。为此，在计算上下位关系得分$r(i,j)$时，考虑了两个专业术语特征向量间的语义相似性$\boldsymbol{R}_i\cdot\boldsymbol{R}_j$（其中，$\cdot$表示两个向量的点乘操作）。同时，对于语句中的任一文本跨距i来说，最多考虑K个在当前语句中与其具有上下位关系的专业术语。

4.2 先序关系判别模块

对于知识主题对(t_a,t_b)，该模块首先从文本D中识别出的专业术语集Y中选取出知识主题t_a、t_b的相关术语，然后进一步根据相关术语间的上下位关系来判断t_a、t_b之间是否存在先序关系。

知识主题的相关术语选取：将知识主题t_a表征为知识主题词向量$\boldsymbol{R}_{t_a}$。基于相似函数$s(t_a,i)$来衡量知识主题t_a与文本中任意专业术语i之间的相似性。使用曼哈顿相似性定义的相似函数$s(t_a,i)$，如下：

$$s(t_a,i)=\left|\boldsymbol{R}_{t_a}-\boldsymbol{R}_i\right| \quad (8)$$

当相似函数值$s(t_a,i)$大于相似阈值θ时，知识主题t_a与专业术语i相关。同理，使用相似函数$s(t_b,i)$选取与知识主题t_b相关的专业术语。

权重策略：不同的相关术语在计算知识主题间先序关系的不对称性时具有不同的作用。为此，使用权重函数衡量不同相关术语在计算知识主题间不对称性的重要性。提出以下两种不同的权重策略。

- 相同权重：当术语与知识主题相关时，所有相关术语具有相同的重要性。权重策略$w_e(t_a,i)$定义为：

$$w_e(t_a,i)=\begin{cases}0, & s(t_a,i)<\theta\\ 1, & s(t_a,i)\geqslant\theta\end{cases} \quad (9)$$

- 不同权重：在衡量知识主题对之间先序关系的不对称性时，给予不同相关术语不同的重要性。术语与知识主题越相似，则该术语对知识主题越重要。使用相似函数$s(t_a,i)$衡量相关术语对知识主题的重要性$w_d(t_a,i)$：

$$w_d(t_a,i)=\begin{cases}0, & s(t_a,i)<\theta\\ s(t_a,i), & s(t_a,i)\geqslant\theta\end{cases} \quad (10)$$

不对称性计算：知识主题的相关术语集之间的先序关系是不对称的，该模块根据相关术语集之间上下位关系指向的差异来衡量知识主题之间的先序关系。提出不对称性函数$f(t_a,t_b)$，以衡量先序关系指向的不对称性。

$$f_{t_a}=\frac{\sum_{i=1}^{K}r(i,j)\cdot w(t_a,i)\cdot w(t_b,j)\cdot g(j)}{\sum_{i=1}^{K}w(t_a,i)\cdot g(i)\cdot w(t_b,j)\cdot g(j)} \quad (11)$$

$$f_{t_b}=\frac{\sum_{i=1}^{K} r(i,j)\cdot w(t_b,i)\cdot w(t_a,j)\cdot g(j)}{\sum_{i=1}^{K} w(t_b,i)\cdot g(i)\cdot w(t_a,j)\cdot g(j)} \quad (12)$$

$$f(t_a,t_b)=f_{t_a}-f_{t_b} \quad (13)$$

其中，j为与文本跨距i具有上下位关系的文本跨距。f_{t_a}用于计算知识主题t_a先于知识主题t_b学习的概率，即t_a是t_b的先序的概率。f_{t_b}用于计算知识主题t_b先于知识主题t_a学习的概率，即t_b是t_a的先序的概率。不对称性函数$f(t_a,t_b)$用于衡量t_a的大多数相关术语是否为t_b的相关术语的先序，即t_a和t_b之间是否存在先序关系的不对称性。因此不对称性函数$f(t_a,t_b)$用于计算t_a和t_b之间存在先序关系的概率。

4.3 损失函数

由于先序关系的稀疏性，正例先序关系的数量远小于候选知识主题对的数量。本文使用了交叉熵损失函数$L(t_a,t_b)$，使得本文提出的端到端先序关系抽取模型更加关注正例先序关系。

$$L(t_a,t_b)=-\boldsymbol{W}_{\text{pos}}u(t_a,t_b)\log\hat{u}(t_a,t_b)-(1-u(t_a,t_b))\log(1-\hat{u}(t_a,t_b)) \quad (14)$$

其中，$\boldsymbol{W}_{\text{pos}}$是正例先序关系样本的权重矩阵，$u(t_a,t_b)$是知识主题对$(t_a,t_b)$的真实先序关系标签，$\hat{u}(t_a,t_b)=\text{sigmoid}(f(t_a,t_b))$为模型预测的知识主题对$(t_a,t_b)$的先序关系。当$t_a$是$t_b$的先序时，$u(t_a,t_b)$=1。

该模型优化了损失函数$L(t_a,t_b)$，使得模型可以更加准确地识别相关术语及抽取术语间的上下位关系。

5 实验与分析

5.1 实验数据集

本文在CrowdComp数据集上进行实验，以验证本文所提端到端先序关系抽取模型的有效性。CrowdComp数据集包含5个不同领域的先序关系数据（见**表1**）。在该数据集中，每对知识主题对(t_a,t_b)的先序关系有4种可能：t_a是t_b的先序；t_b是t_a的先序；知识主题t_a与t_b不相关；知识主题t_a与t_b间的先序关系未知。本实验将第一类先序关系作为知识主题对先序关系的正例数据，其他类作为先序关系的负例数据，并使用留一法验证本文方法在不同领域的实验效果。

在该数据集中，每个知识主题对应一个维基百科页面。本文将每个知识主题的维基百科页面中的文本内容作为知识主题的描述文本D。

5.2 模型参数

经过多次实验发现，以下参数取得了最优效果：使用1 024维ELMo词向量以及8维卷积神经网络（convolutional neural network，CNN）词向量。前馈神经网络FFNN(·)为两层的神经网络。有效文本跨距的最大长度L=15，且λ=0.4。每个知识主题的描述文本中，最多包含K=50个上下位关系。知识主题的相关术语相似性阈值θ=0.3，先序关系判别阈值ϕ=0.3。

5.3 对比实验

选取CrowdComp数据集上3个经典的先序关系抽取方法作为本文端到端先序关系抽取模型的对比方法。实验结果见**表2**。

表1 CrowdComp 数据集

领域	知识主题对数量/个	先序关系对数量/个
减数分裂	400	67
公钥加密	200	27
平行假设	200	25
牛顿定律	400	44
全球变暖	400	43

表2 对比实验结果（准确率）

领域	MaxEnt	RefD	MLP	端到端模型	
				相同权重策略	不同权重策略
减数分裂	51.0%	55.7%	**79.0%**	65.7%	77.12%
公钥加密	67.1%	57.7%	58.0%	72.4%	**85.93%**
平行假设	64.7%	67.9%	**85.0%**	51.7%	80.79%
牛顿定律	53.9%	64.6%	68.0%	45.5%	**86.99%**
全球变暖	56.8%	60.1%	82.0%	33.5%	**84.26%**
平均	58.7%	61.2%	74.4%	53.8%	**83.02%**

● 最大熵（maximum entropy，MaxEnt）[13]方法是第一个在CrowdComp数据集上挖掘先序关系的方法。它同时考虑了基于图的特征以及基于文本的特征，如PageRank分值、编辑历史信息、超链接信息以及概念的长度等。使用最大熵分类器识别概念对的先序关系。

● RefD[1]方法是一种仅根据引用信息衡量先序关系的方法。引用信息即页面中存在的超链接或者页面中提及的另一专业术语。RefD方法首先根据标题匹配的规则获得知识主题的相关术语；然后，通过衡量知识主题的相关术语集之间引用的差异，判断主题之间的先序关系。实验证明，该单一的衡量规则可以简单有效地衡量出概念间的先序关系。

● 多层感知机（multilayer perceptron，MLP）[16]方法从文本资源中抽取全面的特征以识别先序关系。它从维基百科的3个层次（文本、超链接、目录）分别提取特征，如文本中概念出现的次数、概念间存在超链接的数量、概念间是否存在目录层级关系等；并使用所提出的特征训练分类器有效识别出概念间的先序关系。

表2中，加粗字体表示该领域最优先序关系挖掘性能。本文提出的使用不同权重策略的端到端模型在平均性能上最优，且在不同领域的性能差异较小。详细分析如下。

使用不同权重策略的端到端模型的平均性能较使用相同权重策略的端到端模型提高了29.22%。在衡量相关术语集之间先序关系的不对称性时，相同权重策略赋予每个相关术语相同的权重。而不同的相关术语对知识主题的重要性不同，因此在不对称性衡量中的影响也不同。当赋予弱相关的相关术语与紧密联系的相关术语相同的权重时，将导致最终的先序关系结果产生偏差。不同权重策略则赋予不同相关术语不同的权重，使得紧密联系的相关术语在判断先序关系结果时产生较大的影响。因此，不同权重策略使得端到端模型更关注可体现知识主题间先序关系的术语之间的关系，有助于端到端模型更加准确地计算各术语间关系对衡量先序关系不对称性的重要性，进而使得端到端模型取得更优的性能。

显然，基于不同权重策略的端到端模型的性能优于对比方法RefD。端到端模型与RefD均通过衡量知识主题的相关术语集之间互相引用的差异来预测知识主题间的先序关系。端到端模型和RefD的性能差异主要由以下两个原因引起。

● RefD将超链接等引用信息作为计算知识主题相关术语间先序关系差异的依据，而端到端模型将从文本中挖掘的相关术语间的上下位关系作为判断知识主题相关术语间先序关系的依据。超链接等引用信息不能反映知识主题间的先序关系，仅能体现知识主题间存在某种联系。因此，超链接不能作为判断知识主题间先序关系的依据，甚至可能导致错误判断先序关系。而端到端模型使用的文本中专业术语之间有向的上下位关系则是判断知识主题间先序关系不对称性的有力证据，其正确反映了知识主题间的不对称性。因此，端到端模型中挖掘的文本中术语间的上下位关系有力支撑了对知识主题间先序关系不对称性的计算。

● RefD使用流线型的方式挖掘先序关系。其将知识主题的相关术语的确定以及相关术语集之间引用的差异视为两个独立的模块进行。RefD直接确定知识主题的相关术语，并且不在后序计算过程中对相关术语进行优化，即错误识别的相关术语不会被改正，该方法会造成错误的累积。端到端模型将整个先序关系挖掘过程视为一个整体，模型可根据最终预测出的先序关系与真实标签之间的偏差调整对文本中术语的检测以及术语间上下位关系抽取的正确性。即端到端模型通过不断地迭代学习，可以更准确地识别文本中的术语及术语间的上下位关系，并为计算先序关系的不对称性提供了有力的证据。因此，端到端模型的性能优于RefD。

本文所提的基于不同权重策略的端到端模型的性能优于MaxEnt和MLP。MaxEnt和MLP均根据大量的从结构化信息中提取的与先序关系直接相关的特征来预测先序关系。结构化信息在不同的学习资源中是不易获得的。而本节所提的端到端模型仅将知识主题的文本信息作为输入，使得端到端模型被广泛应用到更多的领域中。**表2**中，MLP方法在平行假设领域的性能高于端到端模型。对平行假设领域的数据集进行分析，该领域在维基百科上存在丰富的结构化信息，而MLP方法基于从维基百科中提取的综合的特征，获得了全面的信息，并表现出很好的性能。虽然端到端模型在该领域的性能稍差于MLP方法，但是在平均性能上优于MLP方法。MLP方法中的特征需由领域专家构建，该特征构建过程耗时且领域通用性差。而端到端模型并不使用人工提取的特征，具有更优异的性能。

5.4 相似函数对模型的影响

由于相似函数会影响相关术语以及权重策略的确定，本文进行了对比实验，以验证不同相似函数对模型效果的影响，即在使用不同权重策略的端到端模型上，探究不同相似函数对模型效果的影响。使用余弦相似函数和欧几里得相似函数进行对比实验。

图5为在CrowdComp数据集上使用不同相似函数的模型的实验结果。端到端模型使用不同相似函数对模型效果影响较小，这表明先序关系判别模块可稳定地判别知识主题间是否存在先序关系，该模块具有鲁棒性。在精确率和召回率上，不同相似函数可能降低正例先序关系对被正确预测的概率。不同的相似函数会影响先序关系判别模块正确地识别知识主题的相关术语，使得该模块在计算先序关系的不对称性时产生偏差，最后影响本文端到端模型的先序关系挖掘效果。当相似函数可准确识别出知识主题的相关术语时，本文所提的端到端模型可取得优异的性能。

6 结束语

本文对先序关系数据集进行分析，并发现了先序关系的不对称性特征。基于先序关系的不对称性，本文提出一种从文本中挖掘知识主题间先序关系的端到端模型。该模型包含两个模块，文本中专业术语与上下位关系抽取模块和先序关系判别模块。文本中专业术语与上下位关系抽取模块挖掘文本中专业术语间的上下位关系，上下位关系是一类有向的学习依赖关系。先序关系判别模块在上下位关系的基础上，识别知识主题的相关术语，并计算知识主题的相关术语集间先序关系的不对称性，从而预测知识主题间的先序关系。在CrowdComp数据集上进行实验，并验证了本文所提端到端模型的性能，相比于其他算法，本文所提

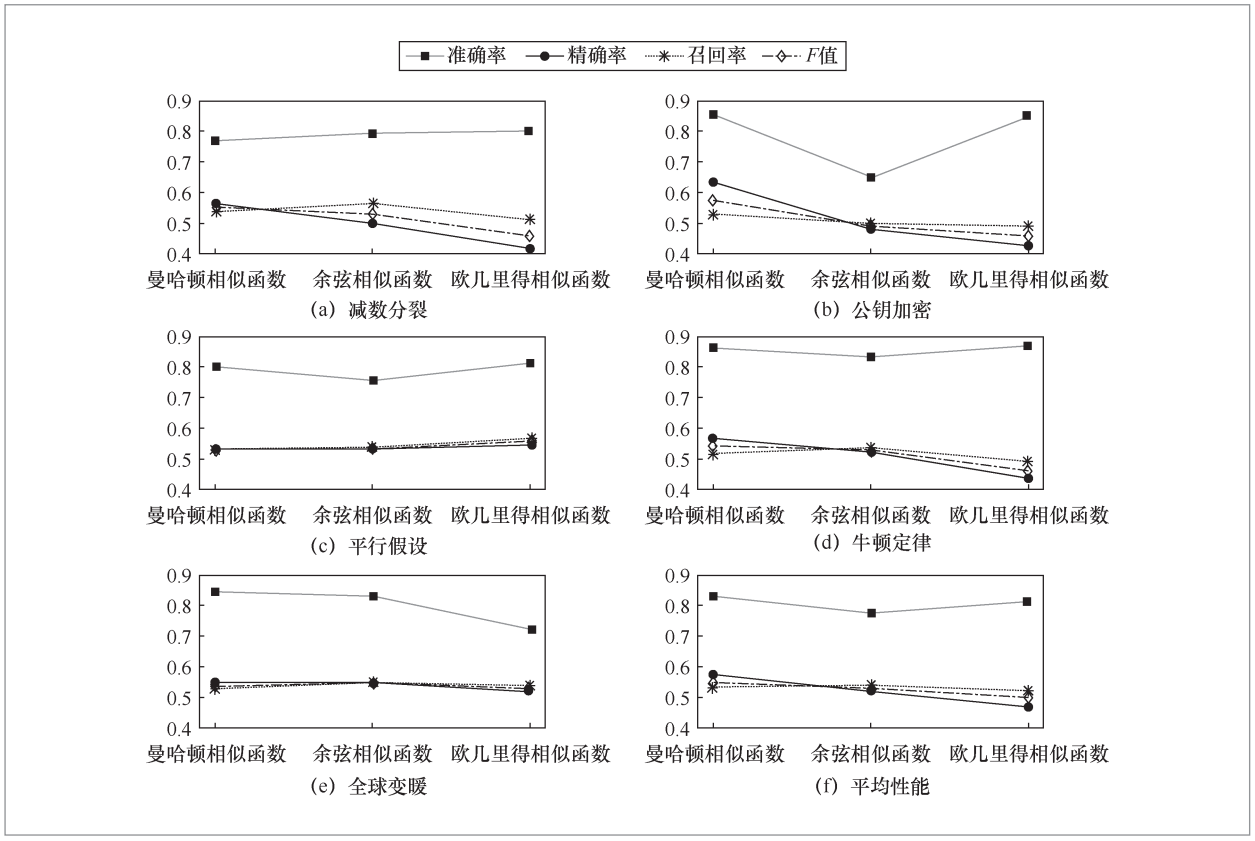

图 5 不同相似函数在不同领域的实验结果

方法取得了最优的性能。

由于部分专业术语间的先序关系需进行跨句子的关系推理才可得出，而本文仅考虑了单一句子中存在的专业术语间先序关系。因此在未来的工作中，需进一步考虑跨句子的专业术语间先序关系，为知识主题间先序关系判断提供更多更有利的关系依据，从而更准确地挖掘知识主题间的先序关系。

参考文献:

[1] LIANG C, WU Z, HUANG W, et al. Measuring prerequisite relations among concepts[C]// 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2015: 1668-1674.

[2] WILEY D A. Learning object design and sequencing theory[D]. Provo : Brigham Young University, 2000.

[3] ZHU H, TIAN F, WU K, et al. A multi-constraint learning path recommendation algorithm based on knowledge map[J]. Knowledge-Based Systems, 2018, 143: 102-114.

[4] AGRAWAL R, GOLSHAN B, PAPALEXAKIS E. Toward data-driven design of educational courses: a feasibility study[J]. Journal of Educational Data Mining, 2016, 8(1): 1-21.

[5] CHEN P, LU Y, ZHENG V W, et al. Prerequisite-driven deep knowledge tracing[C]// 2018 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE Press, 2018: 39-48.

[6] LYU P, WANG X, XU J, et al. Utilizing

knowledge graph and student testing behavior data for personalized exercise recommendation[C]// ACM Turing Celebration Conference-China. New York: ACM Press, 2018: 53-59.

[7] CHEN W, LAN A S, CAO D, et al. Behavioral analysis at scale: learning course prerequisite structures from learner click streams[C]// The 11th International Conference on Educational Data Mining. [S.l.:s.n.], 2018: 66-75.

[8] ALSAAN F, BOUGHOULA A, GEIGLE C, et al. Mining MOOC lecture transcripts to construct concept dependency graphs[C]// The 11th International Conference on Educational Data Mining. [S.l.:s.n.], 2018: 467-473.

[9] CHAPLOT D S, YANG Y, CARBONELL J, et al. Data-driven automated induction of prerequisite structure graphs[C]// The 9th International Conference on Educational Data Mining. [S.l.:s.n.], 2016: 318-321.

[10] PIECH C, BASSEN J, HUANG J, et al. Deep knowledge tracing[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 505-513.

[11] DE MEDIO C, GASPARETTI F, LIMONGELLI C, et al. Automatic extraction and sequencing of Wikipedia pages for smart course building[C]// 2017 21st International Conference Information Visualisation (IV). Piscataway: IEEE Press, 2017: 378-383.

[12] LIANG C, YE J, WANG S, et al. Investigating active learning for concept prerequisite learning[C]// The 32nd AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018: 7913-7919.

[13] TALUKDAR P, COHEN W. Crowdsourced comprehension: predicting prerequisite structure in Wikipedia[C]// The 7th Workshop on Building Educational Applications Using NLP. Stroudsburg: ACL Press, 2012: 307-315.

[14] UPADHYAY P, BINDAL A, KUMAR M, et al. Construction and applications of TeKnowbase: a knowledge base of computer science concepts[C]// The Web Conference 2018. Canton of Geneva: International World Wide Web Conferences Steering Committee, 2018: 1023-1030.

[15] WANG S, ORORBIA A, WU Z, et al. Using prerequisites to extract concept maps from textbooks[C]// The 25th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2016: 317-326.

[16] GASPARETTI F, DE MEDIO C, LIMONGELLI C, et al. Prerequisites between learning objects: automatic extraction based on a machine learning approach[J]. Telematics and Informatics, 2018, 35(3): 595-610.

[17] MANRIQUE R. Towards automatic learning content sequence via linked open data[C]// The International Conference on Web Intelligence. New York: ACM Press, 2017: 1230-1233.

[18] LEE K, HE L, LEWIS M, et al. End-to-end neural coreference resolution[C]// The 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2017: 188-197.

[19] MA J, LIU J, LI Y, et al. Jointly optimized neural coreference resolution with mutual attention[C]// The 13th International Conference on Web Search and Data Mining. New York: ACM Press, 2020: 402-410.

[20] VUONG A, NIXON T, TOWLE B. A method for finding prerequisites within a curriculum[C]// The 4th International Conference on Educational Data Mining. [S.l.:s.n.], 2011: 211-216.

[21] LIANG C, YE J, WU Z, et al. Recovering concept prerequisite relations from university course dependencies[C]// The 31st AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2017: 4786-4791.

[22] ROY S, MADHYASTHA M, LAWRENCE S, et al. Inferring concept prerequisite relations from online educational resources[C]// The 33rd AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2019: 9589-9594.
[23] LIU J, JIANG L, WU Z, et al. Mining learning-dependency between knowledge units from text[J]. The VLDB Journal, 2011, 20(3): 335-345.
[24] ADORNI G, DELL' ORLETTA F, KOCEVA F, et al. Extracting dependency relations from digital learning content[C]// Italian Research Conference on Digital Libraries. Heidelberg: Springer, 2018: 114-119.
[25] NAFA F, KHAN J I, OTHMAN S, et al. Mining cognitive skills levels of knowledge units in text using graph tringluarity mining[C]// 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW). Piscataway: IEEE Press, 2016: 1-4.
[26] MIASCHI A, ALZETTA C, CARDILLO F A, et al. Linguistically-driven strategy for concept prerequisites learning on Italian[C]// The 14th Workshop on Innovative Use of NLP for Building Educational Applications. Stroudsburg: ACL Press, 2019: 285-295.
[27] FILLMORE C J. Frame semantics[J]. Cognitive Linguistics: Basic Readings, 2006, 34: 373-400.
[28] LEE K, HE L, ZETTLEMOYER L. Higher-order coreference resolution with coarse-to-fine inference[C]// The 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Stroudsburg: ACL Press, 2018: 687-692.
[29] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

作者简介

麻珂欣(1995-),女,西安交通大学计算机科学与技术学院硕士生,主要研究方向为先序关系抽取。

魏笔凡(1977-),男,博士,西安交通大学计算机科学与技术学院高级工程师,主要研究方向为Web信息抽取、教育知识图谱构建及应用。

马杰(1993-),男,西安交通大学计算机科学与技术学院博士生,主要研究方向为知识图谱、机器学习、文本挖掘。

刘均（1973– ），男，博士，西安交通大学计算机科学与技术学院教授，主要研究方向为自然语言处理、计算机视觉、智慧教育。

黄毅（1989– ），男，中国移动研究院研究员，主要研究方向为自然语言处理和人机对话。

胡珉（1981– ），男，中国移动研究院主任研究员，主要研究方向为信息检索和知识库。

冯俊兰（1974– ），女，博士，中国移动研究院首席科学家，主要研究方向为语音识别、语言理解和数据挖掘。

收稿日期： 2020-08-31

基金项目： 国家重点研发计划基金资助项目（No.2017YFB1401300, No.2017YFB1401302）；国家自然科学基金资助项目（No.61672419, No.61672418, No.61532015, No.61937001）；“人工智能”教育部-中国移动建设资助项目（No.MCM20190701）；中国工程院咨询研究资助项目“基于MOOC中国的‘一带一路’人才培养的线上线下混合教学支撑信息化平台与服务体系”；国家自然科学基金创新研究群体资助项目（No.61721002）；教育部创新团队资助项目（No.IRT_17R86）；中国工程科技知识中心资助项目

Foundation Items: National Key Research and Development Program of China (No.2017YFB1401300, No.2017YFB1401302), The National Natural Science Foundation of China (No.61672419, No.61672418, No.61532015, No.61937001), MoE-CMCC “Artificial Intelligence” Project (No.MCM20190701), The Consulting Research Project of Chinese Academy of Engineering “The Online and Offline Mixed Educational Service System for ‘The Belt and Road’ Training in MOOC China”, Innovative Research Group of the National Natural Science Foundation of China (No.61721002), Innovation Research Team of Ministry of Education (No. IRT_17R86), Project of China Knowledge Centre for Engineering Science and Technology