(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111428052 B
(45)授权公告日 2023.06.16

(21)申请号 202010235272.5

(22)申请日 2020.03.30

(65)同一申请的已公布的文献号
申请公布号 CN 111428052 A

(43)申请公布日 2020.07.17

(73)专利权人 中国科学技术大学
地址 230026 安徽省合肥市包河区金寨路96号

(72)发明人 刘淇 陈恩红 黄小青 王超 马建辉 苏喻

(74)专利代理机构 北京凯特来知识产权代理有限公司 11260
专利代理师 郑立明 韩珂

(51)Int.Cl.
G06F 16/36(2019.01)
G06N 20/10(2019.01)
G06N 20/00(2019.01)
G06Q 50/20(2012.01)

(56)对比文件
CN 106875014 A,2017.06.20
CN 109299282 A,2019.02.01
CN 109308323 A,2019.02.05
CN 110347894 A,2019.10.18
CN 110532328 A,2019.12.03
US 2004123237 A1,2004.06.24
US 2013138696 A1,2013.05.30
US 2015056596 A1,2015.02.26
US 2017242909 A1,2017.08.24
US 7493253 B1,2009.02.17
涂新辉;何婷婷;李芳;王建文.基于排序学习的文本概念标注方法研究.北京大学学报(自然科学版).2012,(01),全文.
向芳玉;郝建江;顾文玲;黄冬明.基于概念图的可视化教学整合研究——以地理概念为例.中国教育信息化.2018,(16),全文.

审查员 吴海旋

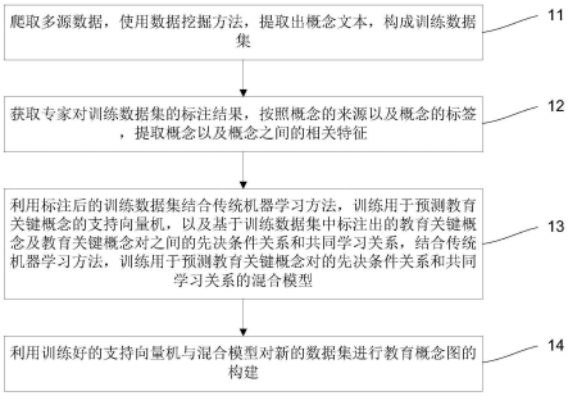
权利要求书6页 说明书10页 附图1页

(54)发明名称

一种从多源数据构建具有多重关系的教育概念图方法

(57)摘要

本发明公开了一种从多源数据构建具有多重关系的教育概念图方法,包括:爬取多源数据,使用数据挖掘方法,提取出概念文本,构成训练数据集;获取专家对训练数据集的标注结果,按照概念的来源以及概念的标签,提取概念以及概念之间的相关特征;利用标注后的训练数据集结合传统机器学习方法,训练用于预测教育关键概念的支持向量机,以及基于训练数据集中标注出的教育关键概念及教育关键概念对之间的先决条件关系和共同学习关系,结合传统机器学习方法,训练用于预测教育关键概念对的先决条件关系和共同学习关系的混合模型;利用训练好的支持向量机与混合模型对新的数据集进行教育概念图的构建。该方法可以精准地构建具有多重关系的教育概念图。



CN 111428052 B

1.一种从多源数据构建具有多重关系的教育概念图方法,其特征在于,包括:

步骤11、爬取多源数据,使用数据挖掘方法,提取出概念文本,构成训练数据集;

步骤12、获取专家对训练数据集的标注结果,标注结果包括:根据概念重要程度为各个概念标注的教育关键概念或非教育关键概念的标签,以及教育关键概念对之间的先决条件关系和共同学习关系;按照概念的来源以及概念的标签,提取概念以及概念之间的相关特征;

步骤13、利用标注后的训练数据集结合传统机器学习方法,训练用于预测教育关键概念的支持向量机,以及基于训练数据集中标注出的教育关键概念及教育关键概念对之间的先决条件关系和共同学习关系,结合传统机器学习方法,训练用于预测教育关键概念对的先决条件关系和共同学习关系的混合模型;

步骤14、利用训练好的支持向量机与混合模型对新的数据集进行教育概念图的构建;

其中,按照概念的来源,所要提取的特征包括:对于每一数据源的概念语义相似度特征,包括:标题匹配特征,用来表示概念是否出现在标题中;概念匹配特征,用来概念对之间的关系;词表征相似度,用来表示概念对在向量空间的相似性与距离;

维基百科链接特征,包括:概念对在维基百科页面中的出入度、概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离以及引用距离;

课本结构化特征与概念共现程度,其中,课本结构化特征包括:目录结构化特征以及课本间结构化特征;概念共现程度,用来表示一个概念对在一个句子中同时出现的次数;

试题答题记录特征,包括:概念频率特征、概念难度距离、试题内容分析距离以及学生答题记录特征;

上述的标题匹配特征、概念频率特征以及概念对在维基百科页面中的出入度是针对单个概念而言,不区分概念是否是教育关键概念;而其余特征是针对概念对而言,只针对教育关键概念对进行提取;

标题匹配特征表示为:

$TM(w_i,ct)\in\{0,1\}$

其中,$ct\in\{CT,p_t,q'\}$,q'表示试题q的标题,w_i表示一个概念,当概念w_i出现在相应的标题中,则$TM(w_i,ct)=1$;否则,$TM(w_i,xt)=0$;

概念匹配特征表示为:

$$CM(w_i,w_j)=\frac{||w_i\cap w_j||}{MAX\{||w_i||,||w_j||\}}$$

其中,(w_i,w_j)为一个概念对,$||.||$表示数目统计符号;

词表征相似度包括:余弦相似度$WEcs(w_i,w_j)$以及欧几里得距离$WEed(w_i,w_j)$;

余弦相似度$WEcs(w_i,w_j)$反映了概念对(w_i,w_j)之间的语义关联,表示为:

$$WEcs(w_i,w_j)=\frac{u_{w_i}\cdot u_{w_i}}{||u_{w_i}||\cdot||u_{w_j}||}$$

欧几里得距离$WEed(w_i,w_j)$表示向量空间中概念对(w_i,w_j)的欧氏距离,表示为:

$$WEed(w_i,w_j)=\sqrt{\sum_{k=1}^{P}(u_{w_{ik}}-u_{w_{jk}})^2}$$

其中，u_{w_i}、u_{w_j}分别表示概念w_i、w_j的词向量，k为向量中元素的序号，P为向量长度；

利用训练好的支持向量机与混合模型对新的数据集进行教育概念图的构建包括：

对于一个未发布的新数据集，按照步骤11的方式提取出各个概念文本，按照步骤12提取概念与概念之间的相关特征；然后，利用训练好的支持向量机与混合模型的参数及相关阈值，构造概念图G，步骤如下：

首先，按照步骤11的方式，提取各个概念文本，构成概念候选集合R，结合各候选概念的相关特征X_{1_t}，以及支持向量机的参数W^1以及第一阈值K^*，抽取关键概念集合C′，表示为：；

$$w_{i'} \in C', \text{if } W^{1^T} X_{1_t} > K^*$$

$$w_{i'} \notin C', \text{if } W^{1^T} X_{1_t} < K^*$$

在得到关键概念集合C′的基础上，根据混合模型的参数W^2与W^3，以及两个阈值P_2与P_3，分别预测关键概念对{$(w_{i'},w_{j'})|w_{i'},w_{j'}\in C'$}之间是否有先决条件关系以及共同学习关系：

$$< w_{i'},w_{j'} >= 0, \text{if } W^{2^T} X_{2_{l'}} < P_1, W^{3^T} X_{3_{l'}} < P_2$$

$$< w_{i'},w_{j'} >= 1, \text{if } W^{2^T} X_{2_{l'}} > P_1, W^{3^T} X_{3_{l'}} < P_2$$

$$< w_{i'},w_{j'} >= 2, \text{if } W^{2^T} X_{2_{l'}} < P_1, W^{3^T} X_{3_{l'}} > P_2$$

其中，$<w_{i'},w_{j'}>=0$表示概念$w_{i'}$和概念$w_{j'}$之间没有先决条件以及共同学习关系，$<w_{i'},w_{j'}>=1$表示概念概念$w_{i'}$和概念$w_{j'}$之间有先决条件关系，$<w_{i'},w_{j'}>=2$表示概念$w_{i'}$和概念$w_{j'}$之间有共同学习关系；$X_{2_{l'}}$、$X_{3_{l'}}$分别表示关键概念集合C′中第1′个概念对$(w_{i'},w_{j'})$之间的用于预测先决条件关系、共同学习关系的相关特征；

以筛选出的关键概念集合C′中的每一教育关键概念作为节点，根据教育关键概念对之间是否存在先决条件关系与共同学习关系，来构造相应节点之间的连接关系，从而构建教育概念图。

2.根据权利要求1所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，所爬取的多源数据至少包括：相关学科的课本数据与历史答题信息、以及相应的维基百科中的相关数据；其中：

相关学科的课本数据包含了n本相同学科的电子课本，表示为：$S=\{B_1,\cdots,B_x,\cdots,B_n\}$，其中$B_x$表示第x本电子课本；对于每一电子课本B，其包含H个子章节，表示为$B=\{C_1,\cdots,C_h,\cdots,C_H\}$，其中$C_h$表示第h个子章节；对于每一子章节包含标题CT以及Y个句子，表示为$C=\{ct,s_1,\cdots,s_y,\cdots,s_Y\}$，其中，$s_y$表示子章节C的第y个句子；

试题答题记录包括：学生答题分数、答题时间以及题目信息；一个试题答题记录是一个五元组$(u,q,s_{uq},t_{uq},con_q)$，其中，$u\in U$表示学生，U为学生集合；$q\in Q$表示试题，Q为试题集合；$s_{uq}$表示答题分数；$t_{uq}$表示答题时间；$con_q$表示试题文本，包含试题内容$con_{q^\dagger}$以及题目解析$con_{q^\ddagger}$；

维基百科中的相关数据对应了M个页面，表示为$P=\{p_1,\cdots,p_m,\cdots,p_M\}$，其中$p_m$表示第m个页面，每个页面p包含了标题$p_t$、摘要$p_{abs}$以及页面内容，表示为$p=(p_t,p_{abs},p_{con})$；

通过分词工具对数据源中的文本内容进行分词，之后将分词内容与百科标题进行匹配，从而提取出不同的数学概念，构成概念集合，从概念集合中随机挑选指定数目的概念，

构成训练数据集。

3.根据权利要求1所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，

概念对在维基百科页面中的出入度：将概念对(w_i,w_j)的出入度分别定义为，IN(w_i)、OUT(w_i)、IN(w_j)、OUT(w_j)；

概念对的公共邻居程度：对于概念对(w_i,w_j)，概念对(w_i,w_j)的公共邻居越多，则概念对(w_i,w_j)的语义相似度越高，表示为：

$$\mathrm{CN}(w_i,w_j)=\frac{OUT(w_i)\cap OUT(w_j)+IN(w_i)\cap IN(w_j)}{\max\{OUT(w_i),OUT(w_j)\}+\max\{IN(w_i),IN(w_j)\}}$$

维基百科摘要定义：如果概念w_i在概念w_j的摘要定义中，那么概念w_i为概念w_j的先序概念，表示为：

$$\mathrm{WAD}(w_i,w_j)=\begin{cases}0, w_i\text{未出现在}w_j\text{的定义中}\\ 1, w_i\text{出现在}w_j\text{的定义中}\end{cases}$$

归一化的谷歌页面距离：通过对谷歌网页中概念之间的超链接，得到概念之间的关联程度，表示为：

$$\mathrm{WAD}(w_i,w_j)=\frac{\max(\log|IN(w_i)|,\log|IN(w_j)|)-\log|IN(w_i)\cap IN(w_j)|}{logN-\min(\log|IN(w_i)|,\log|IN(w_j)|)}$$

引用距离，表示为：

$$\mathrm{RefD}(w_i,w_j)=\frac{\sum_{o_2=1}^{O_2}r(c_{o_2},w_j)\cdot w(c_{o_2},w_i)}{\sum_{o_1=1}^{O_1}w(c_{o_1},w_i)}-\frac{\sum_{o_4=1}^{O_4}r(c_{o_4},w_i)\cdot w(c_{o_4},w_j)}{\sum_{o_3=1}^{O_3}w(c_{o_3},w_j)}$$

其中，0_1表示概念w_i所在维基百科页面中其他概念的数目，0_2表示概念w_i所在维基百科页面中其他概念被概念w_j所在维基百科页面中其他概念所链接的数目，0_3表示概念w_j所在维基百科页面中其他概念的数目，0_4表示概念w_j所在维基百科页面中其他概念被概念w_i所在维基百科页面中其他概念所链接的数目；c_{o_1}、c_{o_2}、c_{o_3}与c_{o_4}均表示维基百科中相应页面的概念；$w(c_{o_1},w_i)$表示概念c_{o_1}是否指向概念w_i所在维基百科页面，1表示指向，0表示未指向；$r(c_{o_2},w_j)$表示概念c_{o_2}在概念w_j所在维基百科页面的重要程度，$w(c_{o_2},w_i)$表示概念c_{o_2}是否指向概念w_i所在维基百科页面；$r(c_{o_4},w_i)$表示概念c_{o_4}在概念w_i所在维基百科页面的重要程度，$w(c_{o_4},w_j)$表示概念c_{o_4}是否指向概念w_j所在维基百科页面。

4.根据权利要求1所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，

目录结构化特征，体现了子章节C中概念对(w_i,w_j)的关系，表示为：

$$\mathrm{TSF}(w_i,w_j)=\frac{\sum_{B\in S}(\sum_{C\in B}f(w_i,C)-f(w_j,C))\Big/|B|}{|S|}$$

其中，|B|表示课本的数量，|S|表示书本的数量，f(w_i,C)是指包含有概念w_i的子章节C

的数目，f(w$_j$,C)表示包含有概念w$_j$的子章节C的数目；

课本间结构化特征，体现了课本中概念对(w$_i$,w$_j$)的关系，表示为：

$$\mathrm{GSF}(w_i,w_j)=\frac{\sum_{B\in S}(f(w_i,B)-f(w_j,B))}{|S|}$$

其中，f(w$_i$,B)是指包含有概念w$_i$的课本B的数目；

概念共现程度，计算公式如下：

$$\mathrm{CCo}(w_i,w_j)=\frac{\sum_{B\in S}\sum_{C\in B}\sum_{s\in C}r(s,w_i)\cdot r(s,w_j)}{\sum_{B\in S}\sum_{C\in B}\sum_{s\in C}r(s,w_i)}$$

其中，r(s,w$_i$)∈{0,1表示概念w$_i$是否出现在句子s中，若出现在句子s中，则取值为1，否则，取值为0；r(s,w$_j$)∈{0,1}表示概念w$_j$是否出现在句子s中，若出现在句子s中，则取值为1，否则，取值为0。

5.根据权利要求1所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，

概念频率特征，表示概念w$_i$的出现频率，表示为：

$$\mathrm{CMIE}(w_i)=\frac{n_{w_i}}{\max\{n_{w_1},\dots,n_{w_i},\dots,n_{w_{|W|}}\}}$$

其中，n_{w_i}是试题内容中出现的概念w$_i$的次数；

概念难度距离，表示包含概念w$_i$试题的平均难度与包含概念w$_j$试题的平均难度的距离，表示为：

CDD(w$_i$,w$_j$)＝CD(w$_i$)CD(w$_j$)

其中，CD(w$_i$)、CD(w$_j$)表示概念w$_i$、w$_j$的平均难度；CD(w$_i$)的计算公式如下：

$$\mathrm{CD}(w_i)=\frac{\sum_{q\in L}f(w_i,Con_q\dagger)\cdot dif_q}{|\mathrm{L}|}$$

其中，$f(w_i,Con_q\dagger)$表示试题内容$Con_q\dagger$中概念w$_i$出现的次数，反映了试题q中概念w$_i$的重要程度；dif$_q$为试题q的难度，L表示试题集合Q中包含概念w$_i$的试题集合，|L|表示L的数目；

试题内容分析距离，计算公式为：

Qcad(w$_i$,w$_j$)＝Qcaw(w$_j$,w$_i$)-Qcaw(w$_i$,w$_j$)

其中：

$$\mathrm{Qcaw}(w_i,w_j)=\frac{\sum_{q\in L}f(w_i,Con_q\dagger)\cdot r(w_j,Con_q\ddagger)}{\sum_{q\in L}f(w_i,Con_q\dagger)}$$

$$\mathrm{Qcaw}(w_j,w_i)=\frac{\sum_{q\in L}f(w_j,Con_q\dagger)\cdot r(w_i,Con_q\ddagger)}{\sum_{q\in L}f(w_j,Con_q\dagger)}$$

其中，$f(w_j,Con_q\dagger)$表示试题内容$Con_q\dagger$中概念w$_j$出现的次数，$r(w_j,Con_q\ddagger)$表示概念w$_j$是否出现在试题分析$Con_q\ddagger$中，$r(w_j,Con_q\ddagger)$表示概念w$_i$是否出现在试题分析$Con_q\ddagger$中，出现取值为1，否则取值为0；表示

学生答题记录特征，表示为：

$$\mathrm{SER}(w_i, w_j) = \frac{\sum_{u \in U} \sum_{(i_1, j_1) \in S(Q)} s_{ui_1} - s_{uj_1}}{|U|}$$

其中，s_{ui_1}、s_{uj_1}分别为学生u在试题i_1、试题j_1上的得分，$S(Q)=\{(i_1,j_1)\,|\,i_1 \in I(Q;w_i), j_1 \in I(Q;w_j), i_1 < j_1\}$、$I(Q;w_i)$、$I(Q;w_j)$各自为试题集合Q中包含概念$w_i$、$w_j$的试题索引，U为学生集合。

6.根据权利要求1所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，训练用于预测教育关键概念的支持向量机的方式包括：

利用标注后的训练数据集，根据各个概念的标签，以及提取的概念特征，即标题匹配特征、以及根据概念对来源提取的概念频率特征、和/或概念对在维基百科页面中的出入度，对支持向量机进行训练，获得支持向量机的完整参数W^1，以及第一阈值K^*；训练的目标是最小化预测标签$W^{1_i^T} X_{1_i}$与实际标签X_i间的误差：

$$min \sum_{i=0}^{M_1} \left| |X_i - W^{1_i^T} X_{1_i}| \right|^2 + \lambda_1 ||W^1||^2$$

其中，M_1表示训练数据集中概念的数目，$W^{1_i^T} X_{1_i}$表示支持向量机预测到的第i个概念的标签，X_{1_i}为第i个概念的相关特征，W_i^1为对于第i个概念的参数，角标Y为矩阵转置符号，M_1个参数W_i^1构成支持向量机的完整参数W^1；X_i表示专家为第i个概念标注的标签；$\lambda_1||W^1||^2$是正则化项，λ_1是手动调节的参数。

7.根据权利要求6所述的一种从多源数据构建具有多重关系的教育概念图方法，其特征在于，混合模型包括用于预测先决条件关系的二分类器、以及用于预测共同学习关系的二分类器；其中：

训练用于预测先决条件关系的二分类器包括：

训练阶段，根据训练数据集中概念的标签选出其中的教育关键概念，利用专家标注的教育关键概念对之间的先决条件关系，结合教育关键概念对之间的概念匹配特征与词表征相似度，以及根据概念对来源提取的概念难度距离、试题内容分析距离与学生答题记录特征，目录结构化特征与课本间结构化特征，和/或概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离与引用距离，来训练用于预测先决条件关系的二分类器，获得二分类器的完整参数W^2及第二阈值P_1；训练的目标是最小化预测标签$W^{2_l^T} X_{2_l}$与实际标签X'_l之间的误差：

$$min \sum_{l=0}^{M_2} \left| |X'_l - W^{2_l^T} X_{2_l}| \right|^2 + \lambda_2 ||W^2||^2$$

其中，M_2表示教育关键概念对的数目，$W^{2_l^T} X_{2_l}$表示对于二分类器预测到的第l个教育关键概念对的标签，即教育关键概念对是否存在先决条件关系，X_{2_l}为第l个教育关键概念对的相关特征，W^2_l为对于第l个教育关键概念对的参数，M_2和参数W^2_l构成了二分类器的完整参数W^2；X'_l表示专家为第l个教育关键概念对标注的先决条件关系，$\lambda_2||W^2||^2$是正则化项，λ_2

是手动调节的参数；

训练用于预测共同学习关系的二分类器的方式包括：

训练阶段，根据训练数据集中概念的标签选出其中的教育关键概念，利用专家标注的教育关键概念对之间的共同学习关系，结合教育关键概念对之间的概念匹配特征与词表征相似度，以及根据概念对来源提取的概念共现程度，概念难度距离，和/或概念对的公共邻居程度以及维基百科摘要定义，来训练二分类器，获得用于预测共同学习关系的二分类器的完整参数W^3及第二阈值P_3；训练的目标是最小化预测标签$W^{3T}_l X_{3l}$与实际标签X''_1之间的误差：

$$min\sum_{l=0}^{M_2}\left|\left|X''_l - W^{3T}_l X_{3l}\right|\right|^2 + \lambda_3||W^3||^2$$

其中，M_2表示教育关键概念对的数目，$W^{3T}_l X_{3l}$表示对于二分类器预测到的第l个教育关键概念对的标签，即教育关键概念对是否存在共同学习关系，X_{3l}为第l个教育关键概念对的相关特征，W^3_1为对于第l个教育关键概念对的参数，M_2和参数W^3_1构成了二分类器的完整参数W^3；X''_1表示专家为第l个教育关键概念对标注的共同学习关系，$\lambda_3||W^3||^2$是正则化项，λ_3是手动调节的参数。

一种从多源数据构建具有多重关系的教育概念图方法

技术领域

[0001] 本发明涉及教育数据挖掘技术领域，尤其涉及一种从多源数据构建具有多重关系的教育概念图方法。

背景技术

[0002] 概念图由各种概念及其关系组成，是一种广泛使用的组织和表示知识的图形工具。在各种概念图中，教育概念图主要关注概念之间的教学关系。因此，它有利于学生组织和获得一个学科的知识。构建教育概念图不仅有利于学生增强自主学习策略，而且在很大程度上有助于教师提高科学教育、教学评价、课程规划等任务，还可以根据教育概念图为学生实现试题或者学习资源的推荐任务(统称为后续任务)。

[0003] 教育概念图能帮助学生高效的、个性化的学习，是智能化个性教学的重要基石。自动准确的构建概念图，可以帮助学生清楚地了解自身的学习路径，同时可以辅助家长和老师为学生制定个性化的学习策略。因此，如何自动的、准确的构建概念图，一直是教育数据挖掘领域探索的一个重要问题。

[0004] 在目前的研究工作和专利中，关于教育概念图构建的方法主要有以下方法：

[0005] 1)基于人工构建的教育概念图方法。

[0006] 目前，基于人工构建的教育概念图方法主要着重于不同学科，由教师或助教提供。

[0007] 2)基于机器学习的教育概念图构建方法。

[0008] 基于机器学习的教育概念图构建方法结合了传统机器学习中常用的分类(如支持向量机)算法，有学者利用此方法抽取维基百科中的概念图。

[0009] 上述两种方法都存在着一些不足，第一种方法费时的，而且，教师和助教只能根据自己的经验为学生开发个性化的概念图。因此，手工概念图难免存在一些错误和遗漏。第二种方法并没有考虑多源信息对构建教育概念图的帮助，而且它们均只关注一种教育学关系，因此构建的图谱是不完善的。教育概念图做后续任务的参考数据，当教育概念图不够准确时，也将影响后续任务的效果。

发明内容

[0010] 本发明的目的是提供一种从多源数据构建具有多重关系的教育概念图方法，通过对不同数据源进行准确的建模分析处理，从而提高预测结果的准确性，进而可以精准地构建具有多重关系的教育概念图。

[0011] 本发明的目的是通过以下技术方案实现的：

[0012] 一种从多源数据构建具有多重关系的教育概念图方法，包括：

[0013] 步骤11、爬取多源数据，使用数据挖掘方法，提取出概念文本，构成训练数据集；

[0014] 步骤12、获取专家对训练数据集的标注结果，标注结果包括：根据概念重要程度为各个概念标注的教育关键概念或非教育关键概念的标签，以及教育关键概念对之间的先决条件关系和共同学习关系；按照概念的来源以及概念的标签，提取概念以及概念之间的相

关特征；

[0015] 步骤13、利用标注后的训练数据集结合传统机器学习方法，训练用于预测教育关键概念的支持向量机，以及基于训练数据集中标注出的教育关键概念及教育关键概念对之间的先决条件关系和共同学习关系，结合传统机器学习方法，训练用于预测教育关键概念对的先决条件关系和共同学习关系的混合模型；

[0016] 由上述本发明提供的技术方案可以看出，该方法针对多种不同的数据源，通过不同的数据集特点，提取出不同的特征；在此基础上，对于三大不同的任务，首先基于相关特征对关键概念进行抽取，之后对分别对两种不同的关系：先决条件关系以及共同学习关系进行抽取。通过对多种数据源的利用以及对多种关系的抽取，弥补了现有方法关系单一以及分类效果不理想的问题，进而更加准确的构建了教育概念图，进而可以更为准确的实现学生个性化试题或者学习资源的推荐。

附图说明

[0017] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他附图。

[0018] 图1为本发明实施例提供的一种从多源数据构建具有多重关系的教育概念图方法的流程图。

具体实施方式

[0019] 下面结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明的保护范围。

[0020] 本发明实施例提供一种从多源数据构建具有多重关系的教育概念图方法的流程图，如图1所示，其主要包括如下步骤：

[0021] 步骤11、爬取多源数据，使用数据挖掘方法，提取出概念文本，构成训练数据集。

[0022] 本发明实施例中，所爬取的多源数据至少包括：相关学科的课本数据与历史答题信息、以及相应的维基百科中的相关数据。

[0023] 1)相关学科的课本数据包含了n本相同学科的电子课本，表示为：$S=\{B_1,\cdots,B_x,\cdots,B_n\}$，其中$B_x$表示第x本电子课本；对于每一电子课本B，其包含H个子章节，表示为$B=\{C_1,\cdots,C_h,\cdots,C_H\}$，其中$C_h$表示第h个子章节；对于每一子章节包含标题CT以及Y个句子，表示为$C=\{ct,s_1,\cdots,s_y,\cdots,s_Y\}$，其中，$s_y$表示子章节C的第y个句子。

[0024] 示例性的，电子课本可以通过互联网下载，再通过OCR工具将下载的课本数据(小学、初中和高中的电子课本)转换为txt格式。

[0025] 2)试题答题记录包括：学生答题分数、答题时间以及题目信息；一个试题答题记录是一个五元组$(u,q,s_{uq},t_{uq},con_q)$，其中，$u\in U$表示学生，U为学生集合；$q\in Q$表示试题，Q为试题集合；$s_{uq}$表示答题分数；$t_{uq}$表示答题时间；$con_q$表示试题文本，包含试题内容$con_{q^\dagger}$以及题

目解析$con_{q^{\ddagger}}$。

[0026] 示例性的，每一学生的试题答题记录可以从在线学习平台智学网获得。

[0027] 3)维基百科中的相关数据对应了M个页面，表示为$P=\{p_1,\cdots,p_m,\cdots,p_M\}$，其中$p_m$表示第m个页面，每个页面p包含了标题$p_t$、摘要$p_{abs}$以及页面内容，表示为$p=(p_t,p_{abs},p_{con})$。

[0028] 通过分词工具对数据集中的文本内容进行分词，之后将分词内容与百科标题进行匹配，从而提取出不同的概念文本，构成概念集合，从概念集合中随机挑选指定数目的概念(具体数目可以根据实际需要来设定)，构成训练数据集。

[0029] 本领域技术人员可以理解，概念主要是指数学上通用的概念形式，例如“一元二次方程”、“函数”、“小数”等。

[0030] 步骤12、获取专家对训练数据集的标注结果，标注结果包括：根据概念重要程度为各个概念标注的教育关键概念或非教育关键概念的标签，以及教育关键概念对之间的先决条件关系和共同学习关系；按照概念的来源以及概念的标签，提取概念以及概念之间的相关特征。

[0031] 本发明实施例中，以概念的重要程度为指标来衡量一个概念是教育关键概念或非教育关键概念，重要程度可以多种常规方式来确定，例如，可以通过概念在数学教材标题中出现的次数来判别，如果出现此处超过规定数值，则认为其重要程度较高，属于教育关键概念；例如，前文提到的“小数”等，还可以由专家根据经验来确定。

[0032] 本发明实施例中，通过多源数据集的特点，根据概念的数据来源，分别提取以下特征：

[0033] (1)对于每一数据源的概念语义相似度特征，包括：标题匹配特征，用来表示概念是否出现在标题中；概念匹配特征，用来表示概念对之间的关系；词表征相似度，用来表示概念对在向量空间的相似性与距离。

[0034] (2)维基百科链接特征，包括：概念对在维基百科页面中的出入度、概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离以及引用距离。

[0035] (3)课本结构化特征以及概念共现程度，其中，课本结构化特征包括：目录结构化特征以及课本间结构化特征，概念共现程度用来表示一个概念对在一个句子中同时出现的次数。

[0036] (4)试题答题记录特征，包括：概念频率特征、概念难度距离、试题内容分析距离以及学生答题记录特征。

[0037] 以上各项特征中，标题匹配特征、概念频率特征以及概念对在维基百科页面中的出入度是针对单个概念而言，因而无需区分概念是否是教育关键概念，而其余特征是针对概念对而言，因此，只针对教育关键概念对进行提取(同样考虑数据来源)；为了便于说明，下面统一使用w_i,w_j来表示训练数据集中的概念，不区分数据来源，也不区分对应的标签。

[0038] 下面针对每一类型的特征做详细的介绍。

[0039] 1、概念语义相似度特征。

[0040] 1)标题匹配特征。

[0041] 标题是对分章内容的总结，指出了分章的要点。如果一个概念出现在标题中，它很可能是一个关键的概念。标题匹配特征表示为：

[0042] $TM(w_i,ct)\in\{0,1\}$

[0043] 其中，ct∈{CT，p_t，q′}，q′表示试题q的标题，w_i表示一个概念，当概念w_i出现在相应的标题中，则TM(w_i，ct)＝1；否则，TM(w_i，ct)＝0。

[0044] 2)概念匹配特征。

[0045] 给定一个概念对＜w_i，w_j＞，如果概念w_i出现在概念w_j中，则w_i更有可能与w_j存在先决条件关系。概念匹配特征表示为：

[0046] $$\mathrm{CM}(w_i,w_j)=\frac{||w_i\cap w_j||}{\mathrm{MAX}\{||w_i||,||w_j||\}}$$

[0047] 其中，||.||表示数目统计符号；

[0048] 3)概念共现程度。

[0049] 4)词表征相似度。

[0050] 词表征相似度包括：余弦相似度WEcs(w_i，w_j)以及欧几里得距离WEed(w_i，w_j)：

[0051] 余弦相似度WEcs(w_i，w_j)反映了概念对(w_i，w_j)之间的语义关联，表示为：

[0052] $$\mathrm{WEcs}(w_i,w_j)=\frac{u_{w_i}\cdot u_{w_i}}{||u_{w_i}||\cdot||u_{w_j}||}$$

[0053] 欧几里得距离WEed(w_i，w_j)表示向量空间中概念对(w_i，w_j)的欧氏距离，表示为：

[0054] $$\mathrm{WEed}(w_i,w_j)=\sqrt{\sum_{k=1}^{P}(u_{wik}-u_{wjk})^2}$$

[0055] 其中，u_{w_i}、u_{w_j}分别表示概念w_i、w_j的词向量，k为向量中元素的序号，P为向量长度。

[0056] 2、维基百科链接特征。

[0057] 1)概念对在维基百科页面中的出入度。

[0058] 通过维基百科页面计算概念的出入度，将概念对(w_i，w_j)的出入度分别定义为：IN(w_i)、OUT(w_i)、IN(w_j)、OUT(w_j)。

[0059] 2)概念对的公共邻居程度。

[0060] 概念对的公共邻居程度：对于概念对(w_i，w_j)，概念对(w_i，w_j)的公共邻居越多，则概念对(w_i，w_j)的语义相似度越高，表示为：

[0061] $$\mathrm{CN}(w_i,w_j)=\frac{OUT(w_i)\cap OUT(w_j)+IN(w_i)\cap IN(w_j)}{\max\{OUT(w_i),OUT(w_j)\}+\max\{IN(w_i),IN(w_j)\}}$$

[0062] 3)维基百科摘要定义。

[0063] 维基百科摘要定义：如果概念w_i在概念w_j的摘要定义中，那么概念w_i为概念w_j的先序概念，表示为：

[0064] $$\mathrm{WAD}(w_i,w_j)=\begin{cases}0,w_i未出现在w_j的定义中\\1,w_i出现在w_j的定义中\end{cases}$$

[0065] 4)归一化的谷歌页面距离。

[0066] 归一化的谷歌页面距离：通过对谷歌网页中概念之间的超链接，得到概念之间的关联程度，表示为：

[0067] $$\mathrm{WAD}(w_i,w_j)=\frac{\max(\log|IN(w_i)|,\log|IN(w_j)|)-\log|IN(w_i)\cap IN(w_j)|}{logN-\min(\log|IN(w_i)|,\log|IN(w_j)|)}$$

[0068] 5)引用距离。

[0069] 引用距离:如果与w_i最关联的概念都指向w_j,那么w_i更有可能是w_j的先序概念,表示为:

[0070] $$\mathrm{RefD}(w_i,w_j)=\frac{\sum_{o_2=1}^{O_2}r(c_{o_2},w_j)\cdot w(c_{o_2},w_i)}{\sum_{o_1=1}^{O_1}w(c_{o_1},w_i)}-\frac{\sum_{o_4=1}^{O_4}r(c_{o_4},w_i)\cdot w(c_{o_4},w_j)}{\sum_{o_3=1}^{O_3}w(c_{o_3},w_j)}$$

[0071] 其中,O_1表示概念w_i所在维基百科页面中其他概念的数目,O_2表示概念w_i所在维基百科页面中其他概念被概念w_j所在维基百科页面中其他概念所链接的数目,O_3表示概念w_j所在维基百科页面中其他概念的数目,O_4表示概念w_j所在维基百科页面中其他概念被概念w_i所在维基百科页面中其他概念所链接的数目;c_{o_1}、c_{o_2}、c_{o_3}与c_{o_4}均表示维基百科中相应页面的概念;$w(c_{o_1},w_i)$表示概念c_{o_1}是否指向概念w_i所在维基百科页面,1表示指向,0表示未指向;$r(c_{o_2},w_j)$表示概念c_{o_2}在概念w_j所在维基百科页面的重要程度,$w(c_{o_2},w_i)$表示概念c_{o_2}是否指向概念w_i所在维基百科页面;$r(c_{o_4},w_i)$表示概念c_{o_4}在概念w_i所在维基百科页面的重要程度,$w(c_{o_4},w_j)$表示概念c_{o_4}是否指向概念w_j所在维基百科页面。

[0072] 3、课本结构化特征以及概念共现程度。

[0073] 概念共现程度用来表示一个概念对在一个句子中同时出现的次数,计算公式如下:

[0074] $$\mathrm{CCo}(w_i,w_j)=\frac{\sum_{B\in S}\sum_{C\in B}\sum_{s\in C}r(s,w_i)\cdot r(s,w_j)}{\sum_{B\in S}\sum_{C\in B}\sum_{s\in C}r(s,w_i)}$$

[0075] 其中,$r(s,w_i)\in\{0,1\}$表示概念w_i是否出现在句子s中,若出现在句子s中,则取值为1,否则,取值为0。$r(s,w_j)$的含义也是如此。

[0076] 课本目录(TOC)和教材结构表明了概念之间的内在联系,因为教师的课程规划是基于这些信息。定义了两种教科书的层次结构特征,包括目录化结构特征和课本间结构化特征,以帮助推断概念之间的关系。

[0077] 1)目录结构化特征。子章节C中概念对(w_i,w_j)的关系,表示为:

[0078] $$\mathrm{TSF}(w_i,w_j)=\frac{\sum_{B\in S}(\sum_{C\in B}f(w_i,C)-f(w_j,C))\Big/|B|}{|S|}$$

[0079] 其中,|B|表示课本的数量,|S|表示书本的数量,$f(w_i,C)$是指包含有概念w_i的子章节C的数目,最终得到的结果是一个数目;同理,$f(w_j,C)$表示包含有概念w_j的子章节C的数目。

[0080] 2)课本间结构化特征。

[0081] 与目录结构化特征类似的,课本间结构化特征,体现了课本中概念对(w_i,w_j)的关系,表示为:

[0082] $$\mathrm{GSF}(w_i,w_j)=\frac{\sum_{B\in S}(f(w_i,B)-f(w_j,B))}{|S|}$$

[0083] 其中，f(w_i,B)是指包含有概念w_i的课本B的数目。

[0084] 4、试题答题记录特征。

[0085] 1)概念频率特征。

[0086] 如果概念w_i经常被试题内容提到，那么w_i更有可能是一个关键的概念。在此假设的基础上，可以通过该特征来提取关键概念。

[0087] 概念频率特征定义为概念w_i的出现频率，表示为：

[0088] $$\mathrm{CMIE}(w_i)=\frac{n_{w_i}}{\max\{n_{w_1},\ldots,n_{w_i},\ldots,n_{w_{|W|}}\}}$$

[0089] 其中，n_{w_i}是试题内容中出现的概念w_i的次数。

[0090] 2)概念难度距离。

[0091] 概念难度距离表示包含概念w_i试题的平均难度与包含概念w_j试题的平均难度的距离，表示为：

[0092] CDD(w_i,w_j)＝CD(w_i)-CD(w_j)

[0093] 其中，CD(w_i)、CD(w_j)表示概念w_i、w_j的平均难度；一般来说，试题难度是指答对试题的学生所占的比例，概念w_i的平均概念难度CD(w_i)是包含概念w_i的所试题的平均难度，CD(w_i)的计算公式如下：

[0094] $$\mathrm{CD}(w_i)=\frac{\sum_{q\in L} f(w_i,Con_{q^\dagger})\cdot dif_q}{|\mathrm{L}|}$$

[0095] 其中，$f(w_i,Con_{q^\dagger})$表示试题内容$Con_{q^\dagger}$中概念w_i出现的次数，反映了试题q中概念w_i的重要程度；dif_q为试题q的难度；L表示试题集合Q中包含概念w_i的试题集合，|L|表示L的数目。

[0096] 同理，CD(w_j)也是类似计算方式，区别仅在将下标i更换为j。

[0097] 3)试题内容分析距离：一般试题内容出现的概念会在试题分析出现的概念之后学，基于这种特性，使用试题内容分析距离来衡量两个概念的先后序关系。

[0098] 试题内容分析距离，计算公式为：

[0099] Qcad(w_i,w_j)＝Qcaw(w_j,w_i)-Qcaw(w_i,w_j)

[0100] 其中：

[0101] $$\mathrm{Qcaw}(w_i,w_j)=\frac{\sum_{q\in L} f(w_i,Con_{q^\dagger})\cdot r(w_j,Con_{q^\ddagger})}{\sum_{q\in L} f(w_i,Con_{q^\dagger})}$$

[0102] $$\mathrm{Qcaw}(w_j,w_i)=\frac{\sum_{q\in L} f(w_j,Con_{q^\dagger})\cdot r(w_i,Con_{q^\ddagger})}{\sum_{q\in L} f(w_j,Con_{q^\dagger})}$$

[0103] 其中，$f(w_j,Con_{q^\dagger})$表示试题内容$Con_{q^\dagger}$中概念w_j出现的次数；$r(w_j,Con_{q^\ddagger})$表示概念w_j是否出现在试题分析$Con_{q^\ddagger}$中，$r(w_j,Con_{q^\ddagger})$表示概念w_i是否出现在试题分析$Con_{q^\ddagger}$中，出现取值为1，否则取值为0；当然，如果w_i(或者w_j)出现在试题内容中，而w_j(或者w_i)出现在试题分析中，那么Qcaw(w_i,w_j)(Qcaw(w_j,w_i))就会变大，这符合实际的情况。

[0104] 4)学生答题记录特征。

[0105] 定义学生u的试题集合为Q，将I(Q;w_i)定义为试题集合Q中包含概念w_i的试题索引，

$I(Q;w_j)$为试题集合Q中包含概念w_j的试题索引。例如，w_i出现在试题集合Q第一个和第三个试题中，则$I(Q;w_i)\in\{1,3\}$。假设w_j是w_i的先序概念，在学生u的答案序列中，如果学生答错了包含概念w_i的试题，那么学生u更有可能回答错包含概念w_j的试题。基于这一观察，对于给定的概念对$\langle w_i,w_j\rangle$，定义$S(Q)=\{(i_1,j_1)\mid i_1\in I(Q;w_i),j_1\in I(Q;w_j),i_1<j_1\}$，学生答题记录特征如下：

[0106] $$\mathrm{SER}(w_i,w_j)=\frac{\sum_{u\in U}\sum_{(i_1,j_1)\in S(Q)}s_{ui_1}-s_{uj_1}}{|U|}$$

[0107] 其中，s_{ui_1}、s_{uj_1}分别为学生u在试题i_1、试题j_1上的得分，U为学生集合，|U|表示U的数目。

[0108] 步骤13、利用标注后的训练数据集结合传统机器学习方法，训练用于预测教育关键概念的支持向量机，以及基于训练数据集中标注出的教育关键概念及教育关键概念对之间的先决条件关系和共同学习关系，结合传统机器学习方法，训练用于预测教育关键概念对的先决条件关系和共同学习关系的混合模型。

[0109] 由于概念图构建中缺少大规模标签数据集，本发明实施例中，基于传统机器学习方法训练三个二元分类器；使用第一个分类器(即支持向量机)结合标题匹配特征、概念频率特征以及概念对在维基百科页面中的出入度，来抽取教育关键概念集合C'；将另外两个二元分类器作为混合模型，在得到教育关键概念集合C'的基础上，预测教育关键概念集合C'中关键概念对$(w_{i'},w_{j'})$之间的先决条件关系和共同学习关系，训练阶段的优选实施方式如下：

[0110] 1)训练支持向量机。

[0111] 利用标注后的训练数据集，根据各个概念的标签，以及之前提取的概念特征，即标题匹配特征、以及根据概念对来源提取的概念频率特征、和/或概念对在维基百科页面中的出入度，对支持向量机进行训练，获得支持向量机的完整参数W^1，以及第一阈值K^*；训练的目标是最小化预测标签$W^{1T}_i X_{1_i}$与实际标签X_i间的误差：

[0112] $$min\sum_{i=0}^{M_1}\left|\left|X_i-W^{1T}_i X_{1_i}\right|\right|^2+\lambda_1||W^1||^2$$

[0113] 其中，M_1表示训练数据集中概念的数目，$W^{1T}_i X_{1_i}$表示支持向量机预测到的第i个概念的标签(即概念为教育关键概念或非教育关键概念)，X_{1_i}为第i个概念的相关特征，W_i^1为对于第i个概念的参数，角标T为矩阵转置符号，M_1个参数W_i^1构成支持向量机的完整参数W^1；X_i表示专家为第i个概念标注的标签(即实际标签)；$\lambda_1||W^1||^2$是正则化项，λ_1是手动调节的参数。

[0114] 2)训练用于预测先决条件关系的二分类器。

[0115] 关键概念对$(w_{i'},w_{j'})$之间的先决条件关系通过概念匹配特征、词表征相似度、概念难度距离、试题内容分析距离、学生答题记录特征、目录结构化特征、课本间结构化特征、概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离以及引用距离来预测。

[0116] 训练阶段，根据训练数据集中概念的标签选出其中的教育关键概念，利用专家标

注的教育关键概念对之间的先决条件关系，结合教育关键概念对之间的概念匹配特征与词表征相似度，以及根据概念对来源提取的概念难度距离、试题内容分析距离与学生答题记录特征，目录结构化特征与课本间结构化特征，和/或概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离与引用距离，来训练用于预测先决条件关系的二分类器，获得二分类器的完整参数W^2及第二阈值P_1；训练的目标是最小化预测标签$W^{2T}_l X_{2l}$与实际标签X'_1之间的误差：

[0117] $$min\sum_{l=0}^{M_2}\left||X'_l - W^{2T}_l X_{2l}|\right|^2 + \lambda_2||W^2||^2$$

[0118] 其中，M_2表示教育关键概念对的数目，$W^{2T}_l X_{2l}$表示对于二分类器预测到的第1个教育关键概念对的标签，即教育关键概念对是否存在先决条件关系，X_{2l}为第1个教育关键概念对的相关特征，W^2_1为对于第1个教育关键概念对的参数，M_2和参数W^2_1构成了二分类器的完整参数W^2；X'_1表示专家为第1个教育关键概念对标注的先决条件关系(即实际标签)，$\lambda_2||W^2||^2$是正则化项，λ_2是手动调节的参数。

[0119] 3)训练用于预测共同学习关系的二分类器。

[0120] 如果概念对(w_i,w_j)具有共同学习关系，则它应具有以下属性：

[0121] 语义相似性：它们共享相同的语义信息；

[0122] 共现：它们可能出现在同一个句子中；

[0123] 概念匹配：它们可能包含常用词；

[0124] 类似的难度：包含w_i的问题A和包含w_j的问题B可能具有相同的难度；

[0125] 类似的邻居：他们可能在维基百科链接中共享相同的邻居；

[0126] 共享定义：w_i可能出现在w_j的定义中，反之亦然。

[0127] 基于这些假设，教育关键概念对($w_{i'}$,$w_{j'}$)之间的共同学习关系通过概念匹配特征、词表征相似度、概念共现程度、概念难度距离、概念对的公共邻居程度以及维基百科摘要定义来预测。

[0128] 训练阶段，根据训练数据集中概念的标签选出其中的教育关键概念，利用专家标注的教育关键概念对之间的共同学习关系，结合教育关键概念对之间的概念匹配特征与词表征相似度，以及根据概念对来源提取的概念共现程度，概念难度距离，和/或概念对的公共邻居程度以及维基百科摘要定义，来训练二分类器，获得用于预测共同学习关系的二分类器的完整参数W^3及第二阈值P_3；训练的目标是最小化预测标签$W^{3T}_l X_{3l}$与实际标签X''_1之间的误差：

[0129] $$min\sum_{l=0}^{M_2}\left||X''_l - W^{3T}_l X_{3l}|\right|^2 + \lambda_3||W^3||^2$$

[0130] 其中，M_2表示教育关键概念对的数目，$W^{3T}_l X_{3l}$表示对于二分类器预测到的第1个教育关键概念对的标签，即教育关键概念对是否存在共同学习关系，X_{3l}为第1个教育关键概念对的相关特征，W^3_1为对于第1个教育关键概念对的参数，M_2和参数W^3_1构成了二分类器的完

整参数W^3;X''_1表示专家为第1个教育关键概念对标注的共同学习关系(即实际标签),$\lambda_3||W^3||^2$是正则化项,λ_3是手动调节的参数。

[0131] 本发明实施例中,第一阈值K^*的数值可以根据需要做适当调整;例如,想要筛选出较多教育关键概念时,可以适当降低第一阈值K^*的数值;反之,可以适当增加第一阈值K^*的数值。

[0132] 本领域技术人员可以理解,概念对的各项特征是根据其所在数据源的相关信息来计算的,因此,此处提到的概念对主要是指相同数据源中的两个概念。在大多数情况下,相同的一个概念对,在三个数据源都存在,也就是说,一个相同内容的概念对,可以根据三个数据源中的相关信息计算出步骤12所提到的四类特征;但是,还考虑概念对只出现在一个或者两数据源的情况,此时,一个相同内容的概念对,只能够提取出步骤12所提到的两类或者三类特征,因此,上述训练过程中,根据概念对来源提取的特征之间使用了“和/或”的描述形式。

[0133] 步骤14、利用训练好的支持向量机与混合模型对新的数据集进行教育概念图的构建。

[0134] 对于一个未发布的新数据集,按照步骤11的方式提取出各个概念文本,按照步骤12提取概念与概念之间的相关特征;然后,利用训练好的支持向量机与混合模型的参数及相关阈值,构造概念图G,步骤如下:

[0135] 首先,按照步骤11的方式(即基于分词技术),提取各个概念文本,构成概念候选集合R,结合各候选概念的相关特征X_{1_t},以及支持向量机的参数W^1以及第一阈值K^*,抽取关键概念集合C′,表示为:

[0136] $$w_{i'} \in C', \text{if } W^{1^T} X_{1_t} > K^*$$

[0137] $$w_{i'} \notin C', \text{if } W^{1^T} X_{1_t} < K^*$$

[0138] 其中,相关特征X_{1_t}是指第t个概念的特征(与步骤13中的X_{1_i}是类似的含义),即标题匹配特征、以及根据概念对来源提取的概念频率特征、或概念对在维基百科页面中的出入度,

[0139] 在得到关键概念集合C′的基础上,根据混合模型的参数W^2与W^3,以及两个阈值P_2与P_3,分别预测关键概念对$\{(w_{i'},w_{j'})|w_{i'},w_{j'} \in C'\}$之间是否有先决条件关系以及共同学习关系:

[0140] $$< w_{i'},w_{j'} >= 0, \text{if } W^{2^T} X_{2_{l'}} < P_1, W^{3^T} X_{3_{l'}} < P_2$$

[0141] $$< w_{i'},w_{j'} >= 1, \text{if } W^{2^T} X_{2_{l'}} > P_1, W^{3^T} X_{3_{l'}} < P_2$$

[0142] $$< w_{i'},w_{j'} >= 2, \text{if } W^{2^T} X_{2_{l'}} < P_1, W^{3^T} X_{3_{l'}} > P_2$$

[0143] 其中,$<w_{i'},w_{j'}>=0$表示概念$w_{i'}$和概念$w_{j'}$之间没有先决条件以及共同学习关系,$<w_{i'},w_{j'}>=1$表示概念概念$w_{i'}$和概念$w_{j'}$之间有先决条件关系,$<w_{i'},w_{j'}>=2$表示概念$w_{i'}$和概念$w_{j'}$之间有共同学习关系;$X_{2_{l'}}$、$X_{3_{l'}}$分别表示关键概念集合C′中第1′个概念对($w_{i'}$,$w_{j'}$)之间的用于预测先决条件关系、共同学习关系的相关特征,与步骤13中的X_{2_l}、X_{3_l}是类

似的含义，即$X_{2l'}$包含的特征有：概念匹配特征与词表征相似度，以及根据概念对来源提取的概念难度距离、试题内容分析距离与学生答题记录特征，或者目录结构化特征与课本间结构化特征，或者概念对的公共邻居程度、维基百科摘要定义、归一化的谷歌页面距离与引用距离；X_{3l}包含的特征有：概念匹配特征与词表征相似度，以及根据概念对来源提取的概念共现程度，或者概念难度距离，或者概念对的公共邻居程度以及维基百科摘要定义；以筛选出的关键概念集合C′中的每一教育关键概念作为节点，根据教育关键概念对之间是否存在先决条件关系与共同学习关系，来构造相应节点之间的连接关系，从而构建教育概念图。

[0144] 由于未发布的新数据集通常是与学生对应的，因此，在教育概念图可以反应学生的知识掌握情况，将教育概念图与试题进行链接后，根据教育概念图上的信息，可以生成试题推荐列表，并推荐给相应的学生。比如，通过教育概念图上的信息，发现学生对于二次函数这个教育关键概念的理解能够不足，则可以生成相应的试题推荐列表，来测试学生对二次函数的先序概念(一次函数)以及共同学习概念(二次方程)是否理解，通过这种方式可以对学生的能力进行层层排查，最终找到学生不明白的症结，再通过这些症结来实现试题或者学习资源的个性化推荐等。

[0145] 本发明实施例上述方案，针对多种不同的数据源，通过不同的数据集特点，提取出不同的特征；在此基础上，对于三大不同的任务，首先基于相关特征对关键概念进行抽取，之后对分别对两种不同的关系：先决条件关系以及共同学习关系进行抽取。通过对多种数据源的利用以及对多种关系的抽取，弥补了现有方法关系单一以及分类效果不理想的问题，从而更加准确的构建了教育概念图。

[0146] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例可以通过软件实现，也可以借助软件加必要的通用硬件平台的方式来实现。基于这样的理解，上述实施例的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0147] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明披露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求书的保护范围为准。

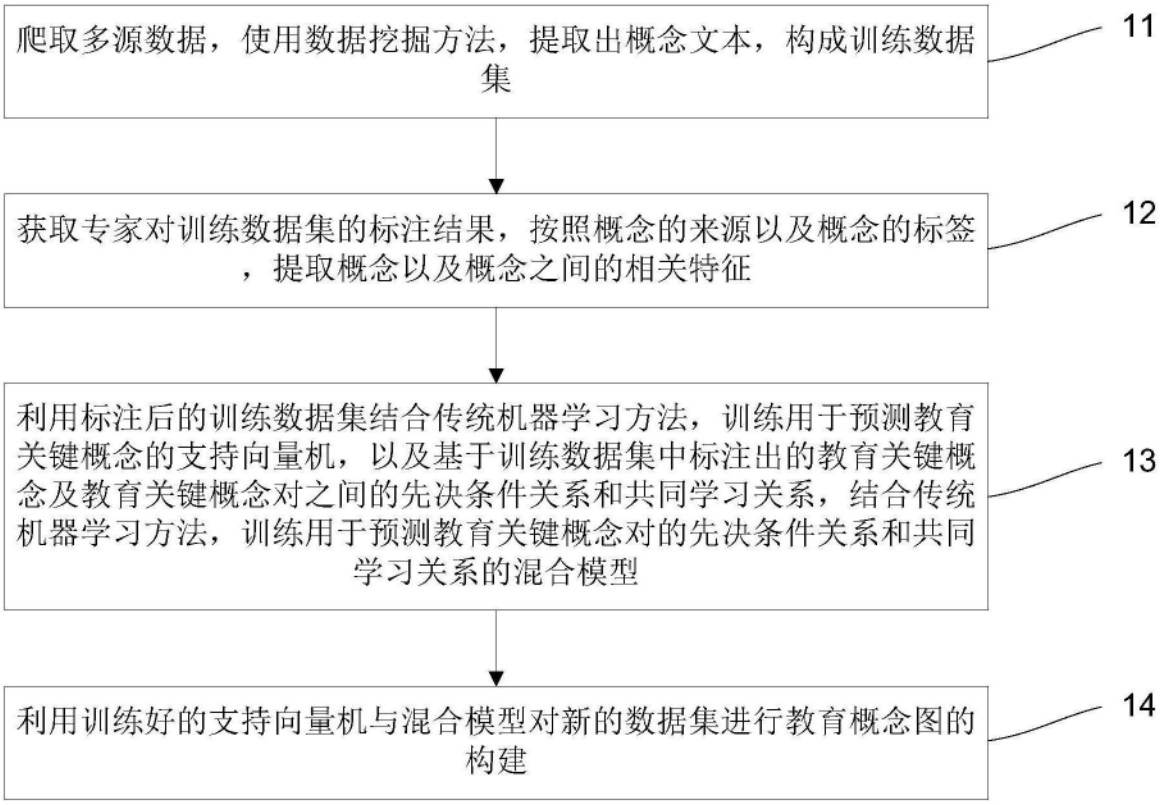

爬取多源数据，使用数据挖掘方法，提取出概念文本，构成训练数据集
11
获取专家对训练数据集的标注结果，按照概念的来源以及概念的标签，提取概念以及概念之间的相关特征
12
利用标注后的训练数据集结合传统机器学习方法，训练用于预测教育关键概念的支持向量机，以及基于训练数据集中标注出的教育关键概念及教育关键概念对之间的先决条件关系和共同学习关系，结合传统机器学习方法，训练用于预测教育关键概念对的先决条件关系和共同学习关系的混合模型
13
利用训练好的支持向量机与混合模型对新的数据集进行教育概念图的构建
14

图1