(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110569443 B
(45)授权公告日 2022.05.17

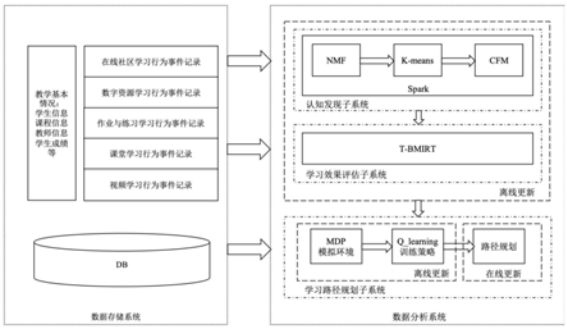
(21)申请号 201910907990.X
(22)申请日 2019.09.24
(65)同一申请的已公布的文献号
申请公布号 CN 110569443 A
(43)申请公布日 2019.12.13
(66)本国优先权数据
201910202413.0 2019.03.11 CN
(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号
(72)发明人 吴文峻 刘丽萍
(74)专利代理机构 北京科迪生专利代理有限责任公司 11251
专利代理师 安丽 邓治平
(51)Int.Cl.
G06F 16/9535(2019.01)
G06Q 10/04(2012.01)
(56)对比文件
CN 105956754 A,2016.09.21
US 2018253989 A1,2018.09.06
CN 108803313 A,2018.11.13
US 6120300 A,2000.09.19
叶露.在线学习资源个性化推荐与学习路径规划研究.《中国优秀硕士学位论文全文数据库信息科技辑》.2018,第I138-562页.
Manjesh Kumar Hanawal,等.Learning Policies for Markov Decision Processes From Data.《IEEE Transactions on Automatic Control 》.2018,第64卷(第6期),第2298-2309页.
审查员 凡保磊
权利要求书3页 说明书7页 附图4页

(54)发明名称

一种基于强化学习的自适应学习路径规划系统

(57)摘要

本发明涉及一种基于强化学习的自适应学习路径规划系统,包括环境模拟,策略训练和路径规划三个模块,整个过程根据改进后的项目反映原理得到的学生每个时刻的能力值,基于马尔科夫决策过程,模拟复杂的学习环境,并合理应用强化学习的算法结合学生历史的学习轨迹离线训练路径规划策略,最后根据训练好的策略在线为学生自适应规划学习路径。本发明最后基于强化学习的思想,将在线教育平台上学习的复杂场景构建于马尔科夫决策过程的框架中,以高效获得能力提升为目标,为学生提供学习资源的持续性推荐,规划最优的学习路径,从而提高学习者的学习效果以及学习效率。



CN 110569443 B

1.一种基于强化学习的自适应学习路径规划系统，其特征在于，包括：环境模拟模块、策略训练模块和路径规划模块；

环境模拟模块，实现将复杂的在线学习环境转化成机器能够理解的语言和文字；基于学生于在线学习平台上的历史学习记录以及学习资源的基本信息，根据改进后的项目反映原理，形式化得到马尔科夫决策过程的五元组；

策略训练模块，实现离线训练基于各能力状态下的路径规划策略的功能；根据环境模拟模块得到的马尔科夫决策过程的五元组，应用基于强化学习的Q_learning算法，离线训练得到基于各能力状态下的路径规划策略；

路径规划模块，实现为目标学生实时进行路径规划的功能；根据策略训练模块得到的策略，基于目标学生当前的能力状态，得到为该目标学生实时规划的最优学习路径，最终达到提高学习效果和效率的目标；

所述环境模拟模块步骤实现如下：

(21)S表示能力状态集合，根据改进后的项目反映原理得到学生每个时刻的能力值，即将学生的能力值定义为状态，为保证状态的离散型，需要进行能力划分，将每一维的学生能力值按学生数量正态分布比例来划分能力值区间，按照学生数量高斯分布比例来划分区间，每个区间取均值作为该区间的能力值；

(22)A表示动作集合，指智能体能够采取的行为集合，在线教育的环境下，即为学生学习的资源集合；

(23)T表示状态转移概率，基于步骤(21)中的能力划分后的状态，以及能力划分后的学生学习行为路径数据，统计计算状态转移概率T；

$T(s,a,s')=P(s_{t+1}=s'|s_t=s,a_t=a)$

$$\sum_{s'\in S} T(s,a,s')=1$$

其中$\forall s,s'\in S$表示状态实例，$\forall a\in A$表示动作实例，t表示时刻，s_t表示t时刻下的状态，a_t表示t时刻下选择的动作；

(24)R表示奖赏，奖赏分为即时奖赏和累积奖赏

即时奖赏应用于学生的学习过程，理解为学生在某个时刻状态为s∈S学习了资源a∈A后转移到状态s′∈S，能够获得该时刻的即时奖赏值r(s,a,s′)，表示R在该时刻获得的奖赏实例，该奖赏值与正确完成概率，正确转移频次以及能力增量三个因素有关；

累积奖赏G，又称之为回报，被定义为奖赏序列的某一具体函数，假设当前时刻为t，则t时刻后的奖赏序列为R_{t+1}，R_{t+2}，R_{t+3}，…R_M，M为总时长则累积奖赏G表示为每一时刻即时奖赏的和，然后再加入折扣因子得到：

$$G=R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots+R_M=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$$

(25)γ表示折扣因子，在上述计算累积奖赏的表示中，γ∈[0,1]，相当于将未来的回报打了折扣，若γ趋近于0，则仅关心当前的即时奖赏，往往会执行使当前即时奖赏最大的行为，本质是一种贪心行为；若γ趋近于1，则会更多的考虑未来回报。

2.根据权利要求1所述的基于强化学习的自适应学习路径规划系统，其特征在于：所述

策略训练步骤如下：

(31)存储环境模拟步骤中的得到的马尔科夫决策过程的五元组<S,A,T,R,γ>；

(32)从能力状态集合S中随机选择一个初始的能力状态S_1；

(33)基于ε-greedy策略在能力状态S_1下选择资源A_1进行学习，再根据环境可观测下一个能力状态S_2，同时得到即时奖赏R_2，此时选择当前能力状态下最大的Q值用以更新Q值表：

$$Q_{k+1}(S_1,A_1)=(1-\alpha)Q_k(S_1,A_1)+\alpha[R_2+\gamma max_a Q_k(S_2,A_2)],\forall a\in A$$

其中Q_k表示当前的Q值表，Q_{k+1}表示更新后的Q值表，α表示更新比例，每次按新值部分更新旧值；

(34)不断循环步骤(33)，直至学习能力达到要求，即到达终止状态，循环步骤(32)，重新选择初始的能力状态；

(35)以字典的形式存储每个能力状态下的最优路径，至此策略训练完成。

3.根据权利要求1所述的基于强化学习的自适应学习路径规划系统，其特征在于：所述路径规划模块实现步骤如下：

(41)获取目标学生的当前能力状态s∈S；

(42)在策略中，寻找与能力s最接近的状态下的一个学习路径l；

(43)将所述学习路径推荐给该目标学生，并在后续的学习过程中自适应调整规划学习路径。

4.根据权利要求3所述的基于强化学习的自适应学习路径规划系统，其特征在于：所述步骤(43)中，自适应调整规划路径步骤如下：

(51)根据目标学生当前的能力s为该目标学生规划学习路径，下一个学习阶段后，目标学生的能力状态变更为s′；

(52)重复步骤(42)，依据目标学生更新后的能力状态s′，为该目标学生规划新的推荐路径l′

(53)比较步骤(42)中的一个学习路径l的后续路径与新的推荐路径l′，若不同，则用新的推荐路径l′替换所述步骤(42)中的学习路径l，若相同则不变。

5.根据权利要求1所述的基于强化学习的自适应学习路径规划系统，其特征在于：所述步骤(21)中，学生能力状态区间的离散化方法，依照学生数量1:2:5:2:1的高斯分布分布比例来划分五个区间。

6.根据权利要求1所述的基于强化学习的自适应学习路径规划系统，其特征在于：步骤(24)中，即时奖赏值与如下三个因素有关：

P(T)：正确完成概率，学生在该某个时刻能力值下能够正确完成学习资源a的概率，基于学习效果评估模型预测；

F(T)：正确转移频次，学生路径中所有从状态s转移通过a转移到状态s′的样本，其中通过正确完成学习资源而完成转移的概率，表示为：

$$F(T)=\frac{C(从s正确完成a转移到s')}{C(从s通过a转移到s')}$$，C表示样本数

$Diff(s_1,s_2)=(s'-s)\cdot difficulty_a$，转化前后能力的最大增量表示为能力前后的差值向量与学习资源难度的点积，以匹配学生的能力值与学习资源的难度，并将向量标量化，

便于奖赏的计算与比较；

即时奖赏r表示为：

$r(s,a,s')=\omega\times Diff(s,s')$

$\omega=P(T)\times F(T)+(1-P(T))\times(1-F(T))$

其中，ω作为最大能力增量的系数。

7.根据权利要求2所述的基于强化学习的自适应学习路径规划系统，其特征在于：所述步骤(33)中，ε-greedy策略的具体步骤如下：

(71)指定ε∈[0,1]值，并随机一个0-1之间的随机数；

(72)若随机数小于ε则随机选择当前能力状态下可选择的资源进行学习，每个资源被选择的概率均为$\frac{\varepsilon}{|A1|}$，其中|A1|为当前状态下可选择的资源个数；

(73)若随机数大于等于ε则选择当前状态下具有最大状态-动作值Q的资源进行学习。

一种基于强化学习的自适应学习路径规划系统

技术领域

[0001] 本发明涉及一种基于强化学习的自适应学习路径规划系统,属于计算机应用技术领域。

背景技术

[0002] 随着在线教育的日益普及,学生可以使用各种电子学习资源,包括电子书,课后练习和学习视频,鉴于学生的背景,学习方式与知识水平的多样性和差异性,在线教育平台需要引入个性化的学习资源推荐工具,以方便学生选择自己的学习路径,满足他们个性化的学习需求。

[0003] 现有的个性化学习资源推荐算法,基本可以分为两类,基于规则的推荐和数据驱动的推荐,大多数智能导学系统(Intelligent Tutoring System,ITS),大多采用基于规则的方法来进行学习资源的推荐,这就需要领域专家来评估不同类型学生的学习场景,并定义相应广泛的推荐规则。很显然,而这种劳动密集型的方法只能应用于特定的学习领域,可扩展性不强。对于现代的大规模在线教育系统,设计者通常采用数据驱动的推荐方法,如协同过滤方法来实现学习推荐算法。这些数据驱动的推荐算法试图通过比较学生和学习对象之间的相似性来为学生推荐合适的学习资源。

[0004] 尽管数据驱动的推荐方法比基于规则的方法更具可扩展性和通用性,但是目前已有的解决方案在实现对学生进行自适应学习资源推荐方面都存在着相同的问题,即往往只能够根据学习资源的内容或学生的学习行为,来检索相似内容的学习资源或相似学习行为的学生群里,并未考虑到学习资源的难度以及学生学习状态的动态变化情况的影响。

[0005] 基于目前推荐算法的研究现状,传统的推荐算法如协同过滤、隐语义模型等,主要面向于商品推荐或自媒体内容的分发,主要目标是猜测用户的喜好,为用户推荐其感兴趣的商品或内容,不论是用户方面还是内容方面,都更加侧重于相似性的计算;而面向学习资源的推荐,更看重的是学习资源能够为学生带来的能力提升,这不是传统推荐算法中简单基于相似性的计算就可以做到的,且学生能力的提升是过程化的,并非一蹴而就的,这其中便涉及到了学习路径的规划。由此本发明提出一种基于强化学习的自适应学习路径规划方法,有效解决上述问题,并使学生获得最大最快能力提升的策略。

发明内容

[0006] 本发明的技术解决问题:克服现有技术的不足,提供一种基于强化学习的自适应学习路径规划系统,基于强化学习的思想,将在线教育平台上学习的复杂场景构建于马尔科夫决策过程的框架中,以高效获得能力提升为目标,为学生提供学习资源的持续性推荐,规划最优的学习路径,能够提高学习者的学习效果以及学习效率。

[0007] 本发明的技术解决方案:一种基于强化学习的自适应学习路径规划系统,包括:包括环境模拟模块,策略训练模块和路径规划模块。

[0008] 环境模拟模块,实现了将复杂的在线学习环境转化成机器能够理解的语言和文

字;基于学生于在线学习平台上的历史学习记录以及学习资源的基本信息,根据改进后的项目反映原理,形式化得到马尔科夫决策过程的五元组;

[0009] 策略训练模块,实现了离线训练基于各能力状态下的路径规划策略的功能;根据环境模拟模块得到的马尔科夫决策过程的五元组,应用基于强化学习的$Q_learnin_g$算法,离线训练得到基于各能力状态下的路径规划策略;

[0010] 路径规划模块,实现了为目标学生实时进行路径规划的功能;根据策略训练模块得到的策略,基于目标学生当前的能力状态,得到为该目标学生实时规划的最优学习路径。最终达到提高学习效果和效率的目标。

[0011] 所述环境模拟模块步骤如下:环境模拟需要基于马尔科夫决策过程,将复杂的在线学习场景形式化为马尔科夫决策过程的五元组<S,A,T,R,γ>;

[0012] (11)S表示状态,根据改进后的项目反映原理得到的学生每个时刻的能力值,将学生的能力值作为状态S,将每一维的学生能力值按学生数量正态分布比例来划分能力值区间,按照学生数量1:2:5:2:1的分布比例来划分五个区间,每个区间取均值作为该区间的能力值;

[0013] (12)A表示动作,指智能体能够采取的行为集合,在线教育的环境下,即为学生可以学习的资源集合;

[0014] (13)T表示状态转移概率,基于(11)中的状态划分,以及大量能力划分后的学生学习行为路径数据,统计计算状态转移概率T;

[0015] $T(s,a,s')=P(s_t=s'|s_t=s,a_t=a)$

[0016] $$\sum_{s'\in S} T(s,a,s')=1,\forall s\in S,\forall a\in A$$

[0017] (14)R表示奖赏,奖赏分为即时奖赏和累积奖赏;

[0018] 即时奖赏应用于学生的学习过程,可理解为学生在状态s时刻学习了资源a后转移到状态s′,能够获得即时奖赏值r(s,a,s′),该奖赏值与如下三个因素有关:

[0019] • P(T):正确完成概率,学生在该时刻能力值下能够正确完成学习资源a的概率,基于学习效果评估模型预测。

[0020] • F(T):正确转移频次,学生路径中所有从状态s转移通过a转移到状态s′的样本,其中通过正确完成学习资源而完成转移的概率,可表示为:

[0021] $$F(T)=\frac{C(从s正确完成a转移到s')}{C(从s通过a转移到s')}$$,C表示样本数

[0022] • $Diff(s_1,s_2)=(s'-s)\cdot difficulty_a$,转化前后能力的最大增量表示为能力前后的差值向量与学习资源难度的点积,目的是为匹配学生的能力值与学习资源的难度,并将向量标量化,便于奖赏的计算与比较。

[0023] 由此,即时奖赏可表示为:

[0024] $r(s,a,s')=\omega\times Diff(s,s')$

[0025] $\omega=P(T)\times F(T)+(1-P(T))\times(1-F(T))$

[0026] 其中,ω作为最大能力增量的系数,目的是根据学生能力及已知的样本分布,差异化大最大能力增量,学生能从正确完成学习资源中获得能力的增长,反之亦可得到训练,比

如学生答错某道题之后根据反馈意识到了其中蕴含的知识点,对学生而言同样也是一种成长。这样的表示方式也保持了P(T)与F(T)的一致性。

[0027] 累积奖赏

[0028] 累积奖赏(Return,G),又称之为回报,被定义为奖赏序列的某一具体函数,若第t步后的奖赏序列为R_{t+1},R_{t+2},R_{t+3},...R_T,T为总步长,则回报G可简单地表示为每一步即时奖赏的和:

[0029] $$G = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T = \sum_{k=0}^{\infty} R_{t+k+1}$$

[0030] 但由于学生的路径长度不尽相同,若仅以找到最大累积奖赏为目标,随着学生路径长度的增长,G值也会越来越大,而这并不符合本文为学生推荐最优且最短路径的目标,因此此处应该加入折扣因子,来削弱未来回报的影响。

[0031] $$G = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + R_T = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

[0032] (15)γ表示折扣因子,在上述计算累积奖赏的表示中,γ∈[0,1],相当于将未来的回报打了折扣,若γ趋近于0,则仅关心当前的即时奖赏,往往会执行使当前即时奖赏最大的行为,本质是一种贪心行为;若γ趋近于1,则会更多的考虑未来回报。

[0033] 所述策略训练模块步骤如下:

[0034] (21)存储环境模拟步骤中的得到的马尔科夫决策过程的五元组<S,A,T,R,γ>;

[0035] (22)从能力集合S中随机选择一个初始的能力状态S_1;

[0036] (23)基于ε-greedy策略在S_1能力状态下选择了资源A_1进行学习,学习了A_1之后,根据环境可观测下一个能力状态S_2,同时得到了即时奖赏R_2(完成行为策略),此时选择当前能力状态下最大的Q值用以更新Q函数(完成目标策略):

[0037] $Q_{k+1}(S_1,A_1) = (1-\alpha)Q_k(S_1,A_1) + \alpha[R_2 + \gamma \max_a Q_k(S_2,A_2)]$

[0038] (24)不断循环(23),直至学习能力达到要求,即到达终止状态,循环(22),重新选择初始的能力状态;

[0039] (25)以字典的形式存储每个能力状态下的最优策略。

[0040] 进一步的,ε-greedy策略的具体步骤如下:

[0041] (1)指定ε∈[0,1]值,并随机一个0-1之间的随机数;

[0042] (2)若随机数小于ε则随机选择当前能力状态下可选择的资源进行学习(每个资源被选择的概率均为$\frac{\varepsilon}{|A1|}$,其中|A1|为当前状态下可选择的资源个数);

[0043] (3)若随机数大于等于ε则选择当前状态下具有最大状态-动作值Q的资源进行学习;

[0044] 所述路径规划模块步骤如下:

[0045] (31)获取目标学生的当前能力状态s;

[0046] (32)在步骤(25)存储的策略中,寻找与s最接近的状态下的学习路径l;

[0047] (33)将路径l推荐给该目标学生,并在其后续的学习过程中自适应调整规划学习路径。

[0048] 进一步的,自适应调整规划路径步骤如下:

[0049] (1)前序步骤(31,32)根据目标学生当前的能力s可为其规划学习路径l,下一个学习阶段后,目标学生的能力状态变更为s′;

[0050] (2)重复步骤(32),依据目标学生更新后的能力状态s′,为其规划新的推荐路径l′比较l的后续路径与l′,若不同,则用l′替换l,若相同则不变。

[0051] 本发明与现有技术相比的优点在于:现有的学习资源推荐技术主要分为基于规则的推荐和数据驱动的学习资源推荐技术,基于规则的方法来进行学习资源的推荐,需要领域专家来评估不同类型学生的学习场景,并定义相应广泛的推荐规则。是一种劳动密集型的方法,只能应用于特定的学习领域,可扩展性不强,而本发明基于强化学习技术,采用自动化规划学习路径,与基于规则的推荐方法相比,大大节约了人工成本;对于现代的大规模在线教育系统,设计者通常采用数据驱动的推荐方法,这些数据驱动的推荐算法大多通过比较学生和学习对象之间的相似性来为学生推荐合适的学习资源,造成学习路径中存在大量相似冗余的学习资源,未考虑到学生能力提升的效率,本发明以大量学生的历史学习轨迹为样本,提取出学生的学生的能力状态,以最终状态为目标训练推荐策略,实现最快最大地提升学生的能力;本发明采用离线训练策略与在线推荐路径相结合的方式,解决了推荐的响应速度问题,以实现自适应规划学习路径。

附图说明

[0052] 图1为学习路径规划方法的系统结构图;

[0053] 图2为环境模拟的流程示意图;

[0054] 图3为策略训练的流程示意图;

[0055] 图4为学习路径合理性评估的示意图;

[0056] 图5为本技术与现有技术的推荐路径与非推荐路径平均长度对比图;

[0057] 图6为学习路径有效性评估的示意图;

[0058] 图7为本技术路径匹配程度与能力增益数据图。

具体实施方式

[0059] 下面结合附图详细解释本发明提出的基于强化学习的自适应学习路径规划方法。

[0060] 本发明提出的基于强化学习的自适应学习路径规划方法,整体系统架构如图1,基于学生与学习资源的历史数据,教师及学生的用户基本信息,不同学习资源的内容数据(课程视频,课后系统,讨论区等),以及学生与学习资源的交互行为数据,将原始数据存储定期传输至HDFS中长期保存,由于学习路径规划系统在运行过程中也会产生学生与学习资源的交互行为数据,同样也需要将这批数据进行定期更新。基于该部分数据,依次进行环境模拟,策略训练和路径规划步骤,基于马尔科夫决策过程框架模拟学生的学习场景,提取并离散化学生在每个学习阶段的能力向量作为状态,从历史的学习行为数据中统计状态转移概率,并结合学习效果评估模块训练得到的学习资源的固有属性,训练生成强化学习过程中智能体与环境交互反馈的即时奖赏,由此将复杂的在线学习场景形式化为数学层面的马尔科夫决策过程框架,采用强化学习算法反复试错训练最优的学习策略,以上部分由于其计算时间成本的考虑为定期离线更新,最后基于训练好的学习策略,依据目标学生的当前能

力状态，为其规划最优的学习路径，为使推荐系统能够快速响应，该部分实时更新，为学生快速并持续地进行学习资源的推荐和学习路径的规划，随后将目标学生新产生的与学习资源的交互数据存入数据库。

[0061] 本发明基于强化学习，马尔科夫决策过程是对完全可观测的环境进行描述，是对强化学习问题进行的一种数学层面的抽象化和理想化，它使得复杂的环境能够转化成为机器理解的语言和文字，以便于能够使用强化学习的算法对现实环境下的复杂问题进行解答。因而需要对马尔科夫决策过程中的各个关键要素进行数学上的形式化定义，根据学生的学习行为数据，对学生在学习过程中的环境进行模拟步骤流程示意图如图2所示，学习效果评估模型训练得到的学生每个时刻的能力值作为输入，依据正态分布离散化能力值，作为状态S；基于已划分的状态，以及大量的学习行为数据，统计计算状态转移概率T；依据计算公式，即可计算即时奖赏R；基于即时奖赏，采用强化学习算法训练得到策略，即每个状态下可采取的最优动作，可用于为目标学生做推荐，输入目标学生当前的能力状态，为其规划最优的学习路径。基于上述流程，即可将在线教育中复杂的学习环境形式化为马尔科夫决策过程，可表示为一个五元组<S,A,T,R,γ>。

[0062] 本发明中涉及的策略训练步骤，流程示意如图3所示，具体步骤如下：

[0063] (1)存储环境模拟步骤中的得到的马尔科夫决策过程的五元组<S,A,T,R,γ>；

[0064] (2)从能力集合S中随机选择一个初始的能力状态S_1；

[0065] (3)基于ε-greedy策略在S_1能力状态下选择了资源A_1进行学习，学习了A_1之后，根据环境可观测下一个能力状态S_2，同时得到了即时奖赏R_2(完成行为策略)，此时选择当前能力状态下最大的Q值用以更新Q函数(完成目标策略)：

[0066] $Q_{k+1}(S_1,A_1)=(1-\alpha)Q_k(S_1,A_1)+\alpha[R_2+\gamma \max_a Q_k(S_2,A_2)]$

[0067] (4)不断循环(23)，直至学习能力达到要求，即到达终止状态，循环(22)，重新选择初始的能力状态；

[0068] (5)以字典的形式存储每个能力状态下的最优策略。

[0069] 本发明提出的基于强化学习的自适应学习路径规划方法，从目标学生当前的能力状态出发，为其规划最优的学习路径，使学生能力能够得到最高效的提升，对于推荐的学习路径，本发明对比现有技术，对于推荐的学习路径进行了实验评估，实验部分将分为两个方面，推荐路径的有效性实验，以及推荐路径的合理性实验。

[0070] 1.合理性实验

[0071] 推荐路径的合理性实验主要用于验证，推荐路径中的学习资源对于目标学生而言是否合理，从路径的长度考虑，是否能够使学生最快得获得能力的提升，即比较相同起始能力以及相同最终能力的路径，推荐路径与实际路径相比是否路径长度更短。如图4所示，本发明为每个能力状态的学生推荐一条路径，对于每一条路径，从大量的学生原始交互数据中，挑选出与推荐路径相同初始能力以及相同最终能力的非推荐路径，比较路径长度的差异，为比较不同能力层次的学生推荐路径与非推荐路径的长度的差异，本发明根据学生起始的能力向量进行聚类，将学生分为5类，从I到V，综合能力由低至高，统计每个类别下所有与推荐路径相同起止能力的非推荐学习路径长度，并在如下不同的推荐算法中比较相应推荐及非推荐路径长度的均值大小，其中UCF和ICF为协同过滤推荐算法，PI，VI，Sarsa以及Q_learning为基于强化学习的学习路径规划算法。对于实验指标，本发明直观的采用推荐路

径的长度L_{rec},以及非推荐路径的平均长度L_{no_rec}。

[0072] $L_{rec}=l_{rec}$

[0073] $$L_{no_{rec}}=\mathbb{E}[\sum_i l_{no_rec_i}]$$

[0074] 1)UCF:基于用户的协同过滤算法,计算学生能力的相似性,推荐与目标学生能力相似学生的学习路径。

[0075] 2)ICF:基于物品的协同过滤算法,计算学习资源属性的相似性,搜索与目标学生历史学习资源的相似学习资源,将与此学习资源有交互行为的学生,其他的学习资源推荐给目标学生。

[0076] 3)PI:基于策略迭代的路径规划算法,基于动态规划的强化学习算法。

[0077] 4)VI:基于值迭代的路径规划算法,基于动态规划的强化学习算法。

[0078] 5)Sarsa:基于Sarsa的路径规划算法,时序差分同步策略强化学习算法。

[0079] 6)Q_learning:基于Q_learning的路径规划算法,时序差分异步策略强化学习算法,为本发明采用的策略训练方法。

[0080] 合理性实验的结果如图5所示,对比不同起始能力状态下,推荐算法在起始能力较低时的表现较好,而起始能力已经在一个较高状态时,推荐的效果与非推荐效果相差不大,表明能力值较高的学生本身已具备较强的学习能力,且可选择的资源空间较小。

[0081] 基于强化学习的推荐算法在相同的初始能力层次下,推荐的路径长度整体短于UCF和ICF算法的推荐路径,原因在于基于协同过滤的路径规划算法仅考虑了学生或学习资源的相似性,为目标学生推荐相似学生的路径或相似的学习资源,并未考虑到学生在学习过程中能力提升的需求。其中ICF更多地为学生推荐了相似的学习资源,虽有反复巩固知识的作用,减少了知识点的遗忘,也可带来能力值的提升,但反复学习类似的学习资源导致学习路径的冗余,从而学习效率降低。相比之下,UCF在路径长度上带来了相对更合理的推荐表现,但由于其搜索现有学生中已存在的学习路径,未对其他的学习路径进行探索,而相似的学生并不一定具有最优的学习路径,从而导致推荐的学习路径并不能使目标学生达到最大能力的提升,如UCF在第II类中的推荐路径长度为12,但其最终综合能力仅能达到最高能力的72%。

[0082] 比较四种基于强化学习的学习路径规划算法,在相同的初始能力下,均能够达到最高的能力状态。其中基于策略迭代的算法PI与基于值迭代的算法VI推荐效果基本一致,由于其在迭代过程中本质是一致的,均为寻找最优状态值函数,区别在策略迭代基于状态值不断评估策略改进策略,而值迭代则是直接寻找最优状态值函数,再根据状态值计算策略,但由于策略迭代进行了双层迭代,其迭代效率远低于值迭代。

[0083] Sarsa和Q_learning算法与基于动态规划的强化学习算法相比,同等初始状态能力下,推荐的学习路径长度相对更短,尤其在第I类和第II类中推荐表现更优,原因在于基于时序差分的强化学习算法是无模型的学习算法,无需依赖样本数据的环境状态转移概率,而通过不断试错的方式来自主学习环境,在学习的同时也丰富了数据的多样性。

[0084] 同样是时序差分算法,Q_learning算法与Sarsa算法相比,在较低的初始能力状态下,Q_learning的推荐学习路径更短,在高能力的初始能力状态下则表现相似,主要区别在于Sarsa在更新环境和值函数时采用同步策略,使用同一种策略更新状态和动作,以选择的

动作更新值函数，而Q_learning采用异步策略，在更新值函数时自主选择当前值函数最大时的动作值，在探索和利用方面得到了更好的平衡，因而更易得到全局最优路径，而Sarsa的更新方式则倾向于更安全的局部最优路径。

[0085] 而由此带来的问题是Q_leanring的收敛速度较Sarsa更慢，但考虑到本发明的研究内容，可以离线训练策略，采用训练好的策略为学生在线实时推荐学习路径，因而Q_learning是本发明的一个更好的选择。

[0086] 2.有效性实验

[0087] 推荐有效性实验，如图6所示，利用学生已有的历史交互数据，分析真实学习路径与推荐路径的匹配程度与学生在真实学习场景下能力提升的分布，即分析基础相同的学生，完成相同数量的学习资源后，与推荐路径匹配越多，是否能力值提高越多。

[0088] 本发明为每个能力状态的学生推荐一条最优路径，对于每一条路径，从大量的学生原始交互数据中，挑选出与推荐路径相同初始能力的真实学习路径，并以推荐路径的长度截断，比较分析实际路径与推荐路径的匹配程度，以及最终能力值相比起始能力值的提升，即比较相同起始能力状态及相同路径长度下，分析其与推荐路径的匹配程度和能力提升的分布情况。

[0089] 匹配度Match，表示相同起始能力状态下，推荐路径与截断后非推荐路径的匹配程度：

[0090] $$Match = \frac{||Path_{rec} \cap Path_{no_rec}||}{||Path_{rec}||}$$

[0091] 其中，$||Path_{rec} \cap Path_{no_rec}||$表示推荐路径与非推荐路径最长连续公共子串的长度，$||Path_{rec}||$表示推荐路径的长度。

[0092] 图7为基于Q_learning的路径规划算法实验数据，行表示相同匹配程度下，不同初始能力对应的能力增益；列表示相同初始能力下，不同匹配程度对应的能力增益。其中’-’表示在学生的历史交互行为数据中未找到与推荐路径完全匹配的实际路径。由数据可以看出，在相同匹配程度下，初始能力较低时，能力提升越大，如图。当匹配程度为40％及以上时，相同起始能力状态下，能力增益随匹配程度增加而提高，如图7所示，即实际路径与推荐路径匹配程度越高，越有利于学生能力的提升，充分验证了推荐的路径对于学生能力提升的有效性。

[0093] 且对于第I，II类初始能力状态下，在实际交互行为数据中，无法找到与推荐路径完全匹配的真实路径，表示基于Q_learning的推荐算法基于已有的数据探索了新的全局最优路径。

[0094] 上面所述的仅是体现本发明基于强化学习的自适应学习路径规划方法实施例。本发明并不限于上述实施例。本发明的说明书是用于进行说明，不限制权利要求的范围。对于本领域的技术人员，很显然可以有很多的替换、改进和变化。凡采用等同替换或等效变换形成的技术方案，均落在本发明要求的保护范围内。

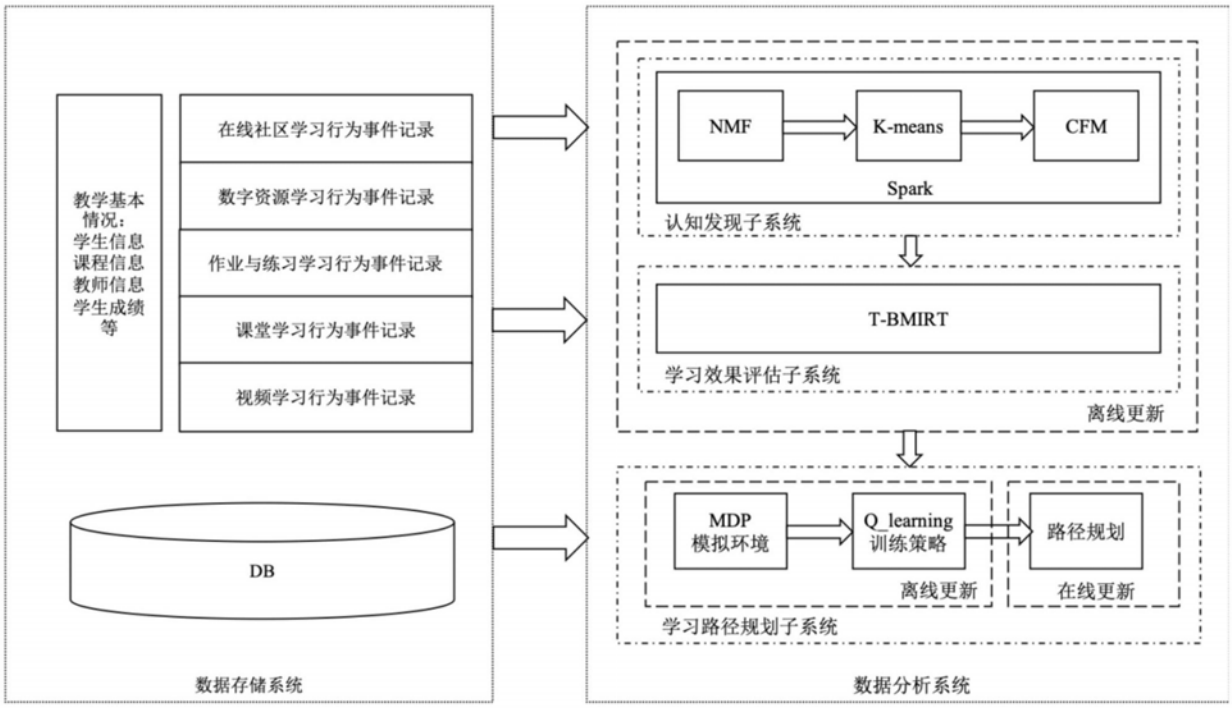

教学基本情况：学生信息 课程信息 教师信息 学生成绩等
在线社区学习行为事件记录
数字资源学习行为事件记录
作业与练习学习行为事件记录
课堂学习行为事件记录
视频学习行为事件记录
DB
数据存储系统
NMF
K-means
CFM
Spark
认知发现子系统
T-BMIRT
学习效果评估子系统
离线更新
MDP 模拟环境
Q_learning 训练策略
离线更新
路径规划
在线更新
学习路径规划子系统
数据分析系统

图1

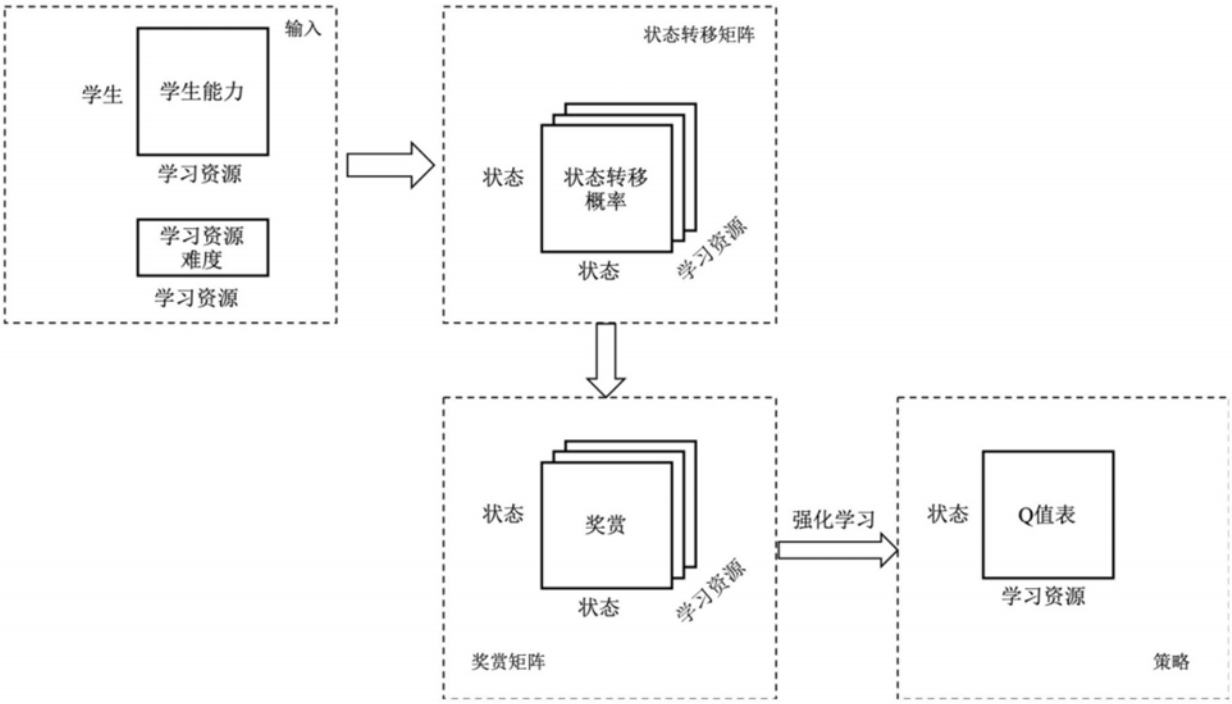

输入
学生
学生能力
学习资源
学习资源 难度
学习资源
状态转移矩阵
状态
状态转移概率
状态
学习资源
状态
奖赏
状态
学习资源
奖赏矩阵
强化学习
状态
Q值表
学习资源
策略

图2

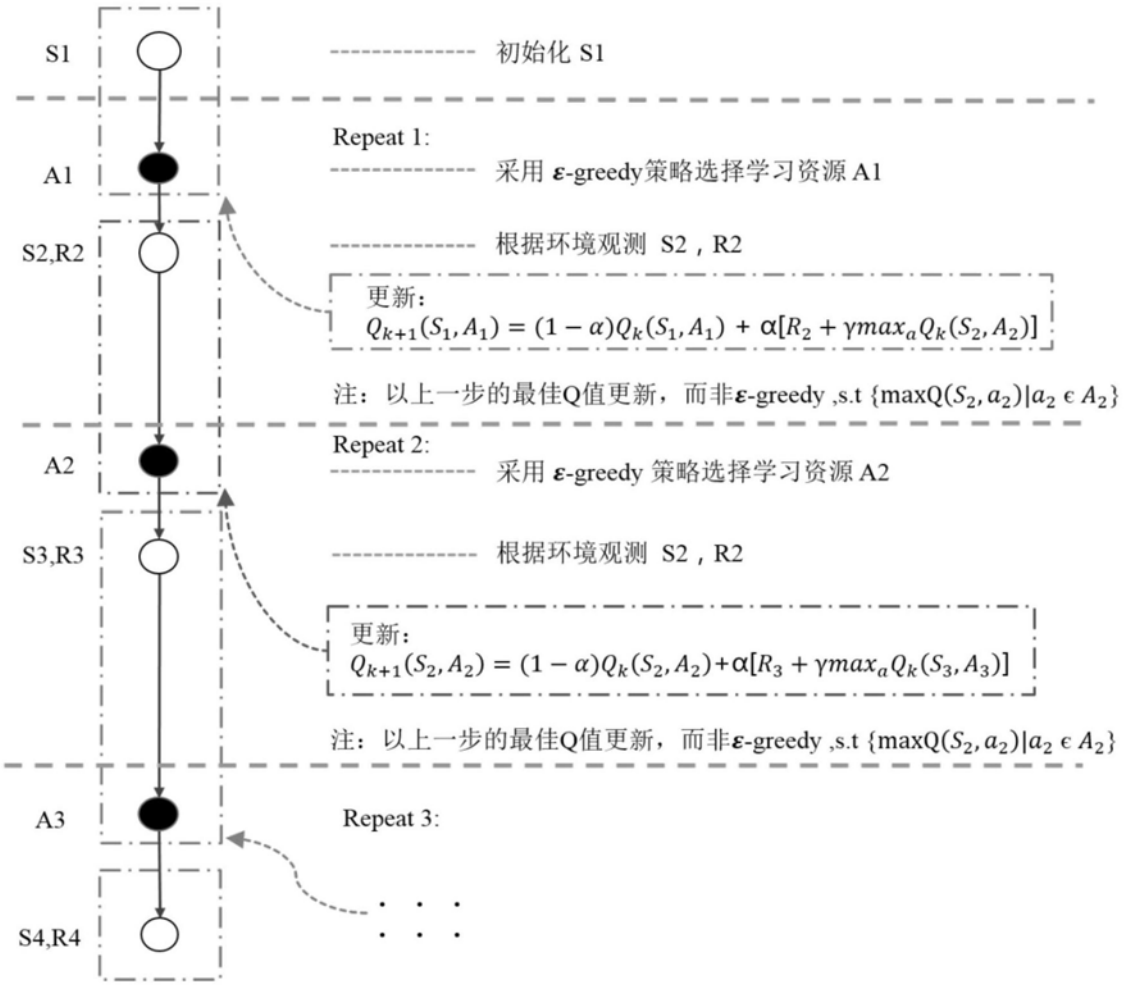
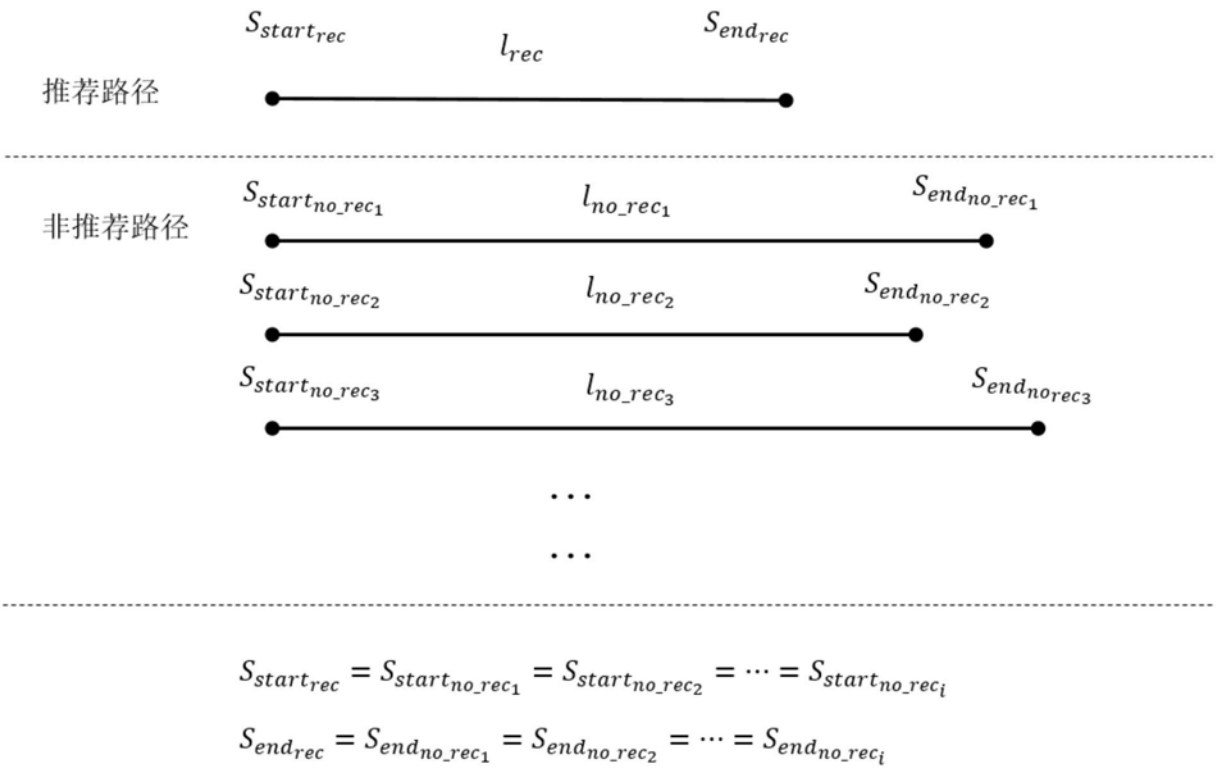

S1
初始化 S1
Repeat 1:
A1
采用 ε-greedy策略选择学习资源 A1
S2,R2
根据环境观测 S2，R2
更新：
$Q_{k+1}(S_1,A_1)=(1-\alpha)Q_k(S_1,A_1)+\alpha[R_2+\gamma max_a Q_k(S_2,A_2)]$
注：以上一步的最佳Q值更新，而非ε-greedy ,s.t $\{\max Q(S_2,a_2)|a_2 \in A_2\}$
Repeat 2:
A2
采用 ε-greedy 策略选择学习资源 A2
S3,R3
根据环境观测 S2，R2
更新：
$Q_{k+1}(S_2,A_2)=(1-\alpha)Q_k(S_2,A_2)+\alpha[R_3+\gamma max_a Q_k(S_3,A_3)]$
注：以上一步的最佳Q值更新，而非ε-greedy ,s.t $\{\max Q(S_2,a_2)|a_2 \in A_2\}$
A3
Repeat 3:
S4,R4

图3



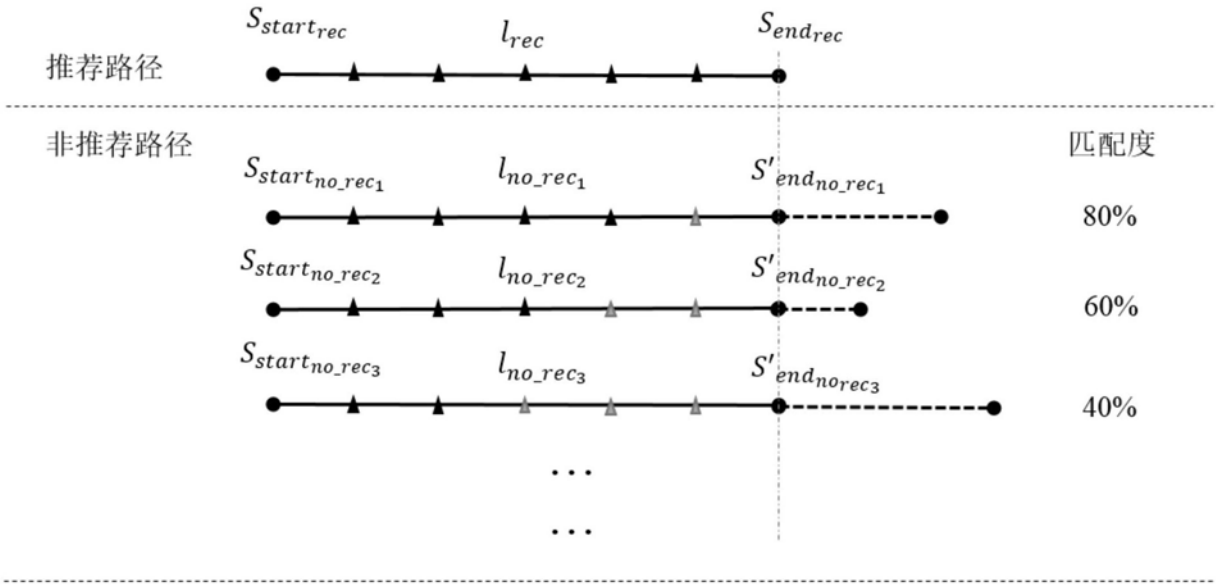
$$S_{start_{rec}} = S_{start_{no_rec_1}} = S_{start_{no_rec_2}} = \cdots = S_{start_{no_rec_i}}$$

$$S_{end_{rec}} = S_{end_{no_rec_1}} = S_{end_{no_rec_2}} = \cdots = S_{end_{no_rec_i}}$$

图4

模型	I		II		III		IV		V	
	L_{rec}	L_{no_rec}	L_{rec}	L_{no_rec}	L_{rec}	L_{no_rec}	L_{rec}	L_{no_rec}	L_{rec}	L_{no_rec}
UCF	14	15.2	12	13.9	8	8.3	3	3.3	1	1
ICF	46	16.9	34	12.4	26	10.2	10	4.3	3	1
PI	14	17.2	11	13.9	6	10.9	3	5.4	1	1
VI	14	17.2	10	13.9	6	10.9	3	5.4	1	1
Sarsa	13	17.9	11	14.7	5	11.1	3	5.4	1	1
Q_learning	9	18.2	8	14.8	5	11.1	3	5.4	1	1

图5



$$S_{start_{rec}} = S_{start_{no_rec_1}} = S_{start_{no_rec_2}} = \cdots = S_{start_{no_rec_i}}$$

$$l_{rec} = l_{no_rec_1} = l_{no_rec_2} = \cdots = l_{no_rec_i}$$

图6

初始能力 / 匹配度	I	II	III	IV	V
20%	18.39%	19.10%	6.65%	8.26%	4.06%
40%	37.39%	18.16%	10.94%	4.46%	1.84%
60%	44.05%	22.47%	11.85%	7.62%	1.04%
80%	47.91%	33.27%	18.11%	8.08%	4.19%
100%	-	-	15.26%	10.90%	8.88%

图7