计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190403
57(6): 1227-1238, 2020

基于强化学习 DQN 的智能体信任增强

亓法欣[1] 童向荣[1] 于 雷[1,2]
[1](烟台大学计算机与控制工程学院 山东烟台 264005)
[2](纽约州立大学宾汉姆顿分校计算机科学系 纽约州宾汉姆顿市 13902)
(qifaxin123@163.com)

Agent Trust Boost via Reinforcement Learning DQN

Qi Faxin[1], Tong Xiangrong[1], and Yu Lei[1,2]
[1](*School of Computer and Control Engineering, Yantai University, Yantai, Shandong* 264005)
[2](*Department of Computer Science, State University of New York at Binghamton, Binghamton, NY* 13902)

Abstract Trust recommendation system is an important application of recommendation system based on social network. It combines the trust relationship between users to recommend items to users. However, previous studies generally assume that the trust value between users is fixed, so it is unable to respond to the dynamic changes of user trust and preferences in a timely manner, thus affecting the recommendation effect. In fact, after receiving the recommendation, there is a difference between actual evaluation and expected evaluation which is correlated with trust value. The user's trust in the recommender will increase when the actual evaluation is higher than expected evaluation, and vice versa. Based on the dynamics of trust and the changing process of trust between users, this paper proposes a trust boost method through reinforcement learning. Least mean square algorithm is used to learn the dynamic impact of evaluation difference on user's trust. In addition, a reinforcement learning method deep q-learning (DQN) is studied to simulate the process of learning user's preferences and boosting trust value. Finally, a polynomial level algorithm is proposed to calculate the trust value and recommendation, which can motivate the recommender to learn the user's preference and keep the user's trust in the recommender at a high level. Experiments indicate that our method applied to recommendation systems could respond to the changes quickly on user's preferences. Compared with other methods, our method has better accuracy on recommendation.

Key words multi-agent systems; reinforcement learning; trust; deep q-learning (DQN); least mean square (LMS)

摘 要 信任推荐系统是以社交网络为基础的一种重要推荐系统应用,其结合用户之间的信任关系对用户进行项目推荐.但之前的研究一般假定用户之间的信任值固定,无法对用户信任及偏好的动态变化做出及时响应,进而影响推荐效果.实际上,用户接受推荐后,当实际评价高于心理预期时,体验用户对推荐者的信任将增加,反之则下降.针对此问题,并且重点考虑用户间信任变化过程及信任的动态性,提出了一种结合强化学习的用户信任增强方法.因此,使用最小均方误差算法研究评价差值对用户信任的动态

收稿日期:2019-06-11;**修回日期**:2019-12-20
基金项目:国家自然科学基金项目(61572418)
This work was supported by the National Natural Science Foundation of China (61572418).
通信作者:童向荣(xr_tong@163.com)

影响,利用强化学习方法 deep q-learning(DQN)模拟推荐者在推荐过程中学习用户偏好进而提升信任值的过程,并且提出了一个多项式级别的算法来计算信任值和推荐,可激励推荐者学习用户的偏好,并使用户对推荐者的信任始终保持在较高程度.实验表明,方法可快速响应用户偏好的动态变化,当其应用于推荐系统时,相较于其他方法,可为用户提供更及时、更准确的推荐结果.

关键词 多智能体系统;强化学习;信任;深度 q 学习;最小均方误差方法

中图法分类号 TP18

多智能体(agent)系统是一种分布式计算技术,是多个自主个体组成的群体系统,目标是通过个体间相互信息的通信,进行交互作用.利用多智能体系统对现实问题进行研究已经相当普遍,在社交网络背景下的信任研究是其中的典型研究内容.随着网络的发展,利用社交网络进行推荐已经非常普遍.许多研究都将社交关系网络中的用户信任值作为基础,通过用户的过往交互记录以及用户间的互动来推测用户的偏好和评级,并向用户进行相关项目的推荐.近年来,许多学者都给出了社交网络中信任计算及推荐的方法,这些方法建立在不同研究基础上,也有不同的研究目的.总体来说,大多数方法都聚焦于信任的传递及信任推荐系统,将信任视为静态不变的参数.而实际上,信任作为一种主观状态,可随用户交互经验、时间等因素的动态变化而发生变化.利用静态信任进行计算会使推荐结果渐渐偏离现实状态.

现有的动态信任研究大多针对信任相关因素的变化以及信任变化后的状态,未充分考虑影响信任动态变化的因素及动态变化过程.实际上,信任动态性将在较大程度上影响推荐结果,动态变化过程会实时地反映到推荐系统中,影响推荐系统的系数,进而实时影响推荐结果,使之得到完全不同的推荐结果.因此,将信任来源的动态性和动态变化一起考虑来改进推荐系统的性能可以得到更加准确、及时的推荐结果,使得推荐系统的实时性得到更大的提高,从而改善推荐系统的性能.

现实生活中,当 A 出于某种目的希望提升 B 对自己的信任时会主动增加与 B 的交流次数,这种交流往往是从 B 的兴趣爱好开始的.如果 B 喜爱看电影,A 会经常向 B 推荐他可能喜欢的电影.当 B 对 A 的推荐电影做出正向评价时,说明 A 的推荐符合 B 在电影方面的偏好,同时 B 将更加相信 A 在电影方面的欣赏水平,此时 B 对 A 的信任将增加;反之,则说明 B 怀疑 A 的欣赏水平,B 对 A 的信任值将降低.随着 A 向 B 推荐电影的次数增加,A 将越来越了解 B 在电影方面的偏好,并可以更精准地推荐 B 喜爱的电影,同时,B 将十分信任 A.该过程实质上是一种学习其偏好并"投其所好"的过程.

本文的方法模拟了上述过程:推荐者为增强用户信任,向用户进行项目推荐,用户接受推荐后,将对项目做出实际评价.实际评价与用户接受项目时的心理预期存在一定差异,该差异决定了用户对项目的满意程度:若实际评价高于心理预期,则用户向推荐者返回正向反馈;反之,用户返回负向反馈.正向反馈表明用户对推荐者的认可,用户对推荐者的信任将增加;负向反馈表明用户怀疑推荐者的推荐水平,导致用户对推荐者信任下降.本文利用强化学习方法实现用户的信任增强,并将其应用到推荐系统中,提高推荐结果的实时性和准确性.实验结果验证了所提出的基于强化学习的深度 q -学习(deep q-learning, DQN)的信任增强算法可以更为准确、及时地展现信任的动态变化,并得到更为可信的推荐结果.由于 DQN 方法有稳定性强、可处理大量数据的特点,所提出的方法可以很好地扩展到推荐系统使用.

本文主要贡献有 2 个方面:

1) 提出的方法结合了强化学习方法深度 q -学习(DQN),对信任变化过程进行学习以增强用户信任.以体验评价和预期评价之间的差值为依据,对用户偏好进行学习,可以得到更为完全的信息,进而提高推荐的个性化水平和准确性.

2) 提出的方法综合考虑了用户的兴趣度、直接信任、间接信任,并对这些因素进行了选择性的筛选,使计算结果更加符合实际.

1 相关工作

在信任的相关研究中,一些学者已经取得了一些成果.如 Jiang 等人[1]提出的邻域感知的信任网络提取方法,目的为解决信任网络中的信任传播失败问题.该方法考虑到用户在在线社交网络中的领域

感知影响力,采用有向多重图对异构信任网络中用户间的多重信任关系进行建模,随后设计了一个领域感知信任度量来度量用户之间的信任程度.Yan 等人[2]提出了一种改进后的基于邻域和矩阵分解的社会推荐算法,旨在解决关系网络中的大规模、噪声和稀疏性问题.该方法开发了一种新的关系网络拟合算法来控制关系的传播和收缩,为每个用户和项目生成一个单独的关系网络.然后将矩阵因子分解与社会正则化和邻域模型相结合,利用关系网络生成建议.一些学者在研究过程中对信任的动态性有所考虑,提出了一些关于动态信任的方法,如 Ghavipour 等人[3]考虑了信任传递过程中用户信任值的改变,提出了基于学习自动机的启发式算法 DLATrust,并使用改进后的协同过滤聚合策略来推断信任的价值.在此基础上,Ghavipour 等人[4]又提出了利用分布式学习自动机的随机信任传播的动态算法 DyTrust,两者目的均为学习发现社交网络中用户之间的可靠路径.游静等人[5]提出了一种考虑信任可靠度的分布式动态管理模型,使用可靠度对信任进行评估来降低不可靠数据的影响,并在交互结束后修正可靠度.此外,许多学者针对自适应声誉和信任相关性质等方面进行了相应的研究[5-11].本节将对前人所做的工作和 DQN 方法进行简要介绍.

1.1 DyTrust

DyTrust 算法是利用学习算法进行动态信任计算的方法之一.DyTrust 考虑了信任传播过程中节点信任值的动态变化,利用分布式学习自动机获取信任传播过程中信任的动态变化,对信任变化做出反应并根据信任的变化来动态更新可靠的信任路径.

该方法作为一种动态信任传播算法,可以更准确地推断出信任路径.但是该方法仅利用了信任的动态性特征,并未对其动态变化过程进行研究.本文的方法对信任动态变化过程进行研究,并详细阐述了该过程.

1.2 q 学习与 DQN

DQN[12]是 q 学习算法[13]的发展,也是将深度学习与强化学习结合起来而实现学习的一种新兴算法.

q 学习算法通过单一神经网络进行值函数估计与现实累积经验计算,与 q 学习相比,DQN 使用 2 个相同结构的神经网络分别计算值函数估计(Q 网络)与现实(target-Q 网络).Q 网络估计每个动作的值(Q_eval),并根据策略选择最终动作,环境根据动作返回奖励值;target-Q 网络利用奖励值进行现实估计(Q_target).

相较于 Q 网络,target-Q 网络的权重更新较慢,即往往每经过多轮更新一次 target-Q 网络.该方法保证 DQN 可避免时间连续性的影响,从而得到更优结果.

同时,DQN 方法利用经验回放对 Q 网络进行训练.DQN 进行神经网络参数训练时,利用贝尔曼方程思想计算 *LossFunction* 并更新 Q 网络权重参数:

$$LossFunction = Q\_target - Q\_eval,$$

target-Q 网络的计算方式由 Markov 决策得到.

本文中的信任增强算法结合 DQN 算法进行计算,实际上针对单个用户的信任增强过程使用 q 学习算法也可以取得相近的结果.现实中使用 q 学习方法时,状态量过多且需人工设计特征,且结果质量与特征设计质量关系紧密,导致 q 学习方法无法应用于推荐系统对大量用户进行项目推荐;同时,q 学习方法需使用矩阵存储 Q 值,当针对用户过多时,会造成数据量过大,导致存储空间需求急剧增加.推荐系统中用户群体数目庞大,推荐项目类别复杂,对 q 学习方法的数据存储来说是一场灾难.

因此,考虑到本文提出的方法应用于推荐系统时的相关问题以及 DQN 相较于 q 学习算法的先进性,本文使用 DQN 算法进行信任增强,并推广至推荐系统.

2 问题描述与基本定义

本节详细介绍了问题的基本描述、用户信息集、推荐者信息集和 DQN 信息集等.

2.1 问题描述

如图 1 所示,用户 A 为提升用户 B 对自己的信任,向 B 推荐与他感兴趣的内容相关的项目.当 B 接受 A 的推荐后,如果 B 对 A 推荐的项目的评价高于其心理预期值,B 对 A 的信任值将增加;反之,B 对 A 的信任值将降低.

2.2 基本定义

定义用户集 $U=\{u_1,u_2,\cdots,u_k\}$,项目集 $I=\{i_1,i_2,\cdots,i_k\}$,兴趣度集合 $I=\{I_1,I_2,\cdots,I_k\}$.其中 $I_1,I_2,\cdots,I_m$ 分别对应项目集 $I_1,I_2,\cdots,I_m$.

定义 1. 用户信息集$\{\boldsymbol{T},\boldsymbol{S},sp\}$.

本文对社交网络中每个用户建立用户信息集.其中,$\boldsymbol{T}$ 为用户信任矩阵,$\boldsymbol{S}$ 表示用户评价矩阵,包括用户过往评价及用户对推荐项的预期评价,sp 表示用户对推荐项的实际评价.

定义 2. 推荐者信息集$\{\boldsymbol{T},\boldsymbol{S},I,I\}$.

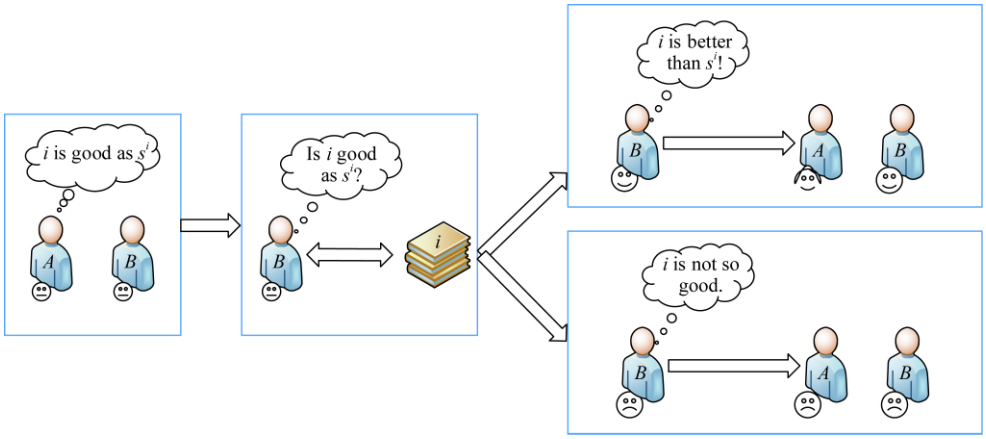

Fig. 1 Relationship between recommendation and trust

图 1 建议-信任影响关系示意图

用户进行推荐后,转化为推荐者,为之建立推荐者信息集.T,S 为推荐者对应的用户信息集中的信任矩阵及评价矩阵,I,I 分别为项目和兴趣度集合.

推荐过程中,推荐者从项目集中选择项目进行推荐,用户对符合偏好的项目有高满意度,满意度将动态影响用户间信任.

定义 3. DQN 信息集$\{n,a,\pi,r\}$.

1) 推荐者状态 n.用户于时刻 τ 发出广播,推荐者根据选择策略做出动作,与该动作对应的推荐者状态为 n_τ.推荐者动作选择结束后,状态更新为 $n_{\tau+1}$并等待用户下一次广播.

2) 推荐者动作 a.推荐者从项目集选择最终向用户推荐的项目,推荐该项目即为推荐者动作 a.

3) 动作选择策略 π.选择策略决定推荐者最终选择的推荐项目.本文选择策略与 DQN 中策略相同,为 ε-greedy policy.

4) 动作奖励 r.用户对推荐者提供的推荐项将有相应的满意度.满意度对信任的影响幅度记为奖励 r,该值影响推荐者在下一时刻的动作选择.

本文通过用户预期评价与实际评价差值来表征用户满意度,利用最小均方误差方法(least mean square, LMS)方法计算评价差值与信任变化的动态映射关系.本过程将信任的动态变化视为 DQN 过程中给予推荐者的奖励,信任的变化将影响推荐者对推荐项目的选择行为.

3 兴趣度、信任计算与建议处理

本节介绍了用户间信任的基本定义及用户建议定义,信任计算结合了用户兴趣度及推荐用户信任,使得计算结果更加符合实际.本节给出了用户建议处理过程,并说明了预期评价的计算方法.

3.1 用户兴趣度

社交网络中,用户会对自己感兴趣的内容进行搜索、浏览.定义 $I=\{I_1,I_2,\cdots,I_k\}$表示信任网络中用户对不同项目类型兴趣度集合.I 越高,说明用户对相关内容了解度越高,其评分参考性越高;反之,表明评分参考性较低.用户进行选择性搜索和浏览时,会因兴趣度不同导致不同的浏览行为.用户兴趣度通过分析浏览记录得出:

1) 网页保存、收藏.$sf(p_k)$表示保存、收藏参数.用户进行保存、收藏行为时,$sf(p_k)=1$,否则 $sf(p_k)=0$.

2) 网页浏览.用户对网页内容感兴趣时,相应的网页浏览时间与访问次数均会增加.设置用户浏览时间比率表示单位页面大小的用户浏览时间,即浏览时间 $time(p_k)$与页面大小 $e(p_k)$之比,时间比率越大,表示用户对该网页内容越感兴趣.页面 p_k 被访问次数 $f(p_k)$与页面浏览时间 $time(p_k)$构成页面浏览参数 $b(p_k)$,即:

$$b(p_k)=\frac{f(p_k)\times\left(\dfrac{time(p_k)}{e(p_k)}\right)}{\max\limits_{\forall p_k\in P}\left(\dfrac{time(p_k)}{e(p_k)}\right)\times\max(f(p_k))},\quad(1)$$

其中,P 是所有用户浏览页面的集合.

3) 点击网页提供的超链接.超链接点击参数 $c(p_k)$通过点击的超链接数 $nc(p_k)$和页面 p_k 提供的超链接总数 $ls(p_k)$计算为

$$c(p_k)=\frac{nc(p_k)}{ls(p_k)}.\quad(2)$$

4) 分享、转发网页内容.分享参数 $trans(p_k)$衡

量用户的分享行为,若用户对网页 p_k 进行分享、转发操作,$trans(p_k)=1$,否则 $trans(p_k)=0$.

用户对项目内容相关网页 p_k 的兴趣度 $I_m(p_k)$ 可计算为

$$I_m(p_k)=\frac{1}{4}(sf(p_k)+b(p_k)+trans(p_k)+c(p_k)). \tag{3}$$

用户对项目 m 相关内容的兴趣度 I_m 可通过用户的网页兴趣度 $I_m(p_k)$ 综合计算得出:

$$I_m=\frac{1}{|P|}\sum_{p_k\in P}I_m(p_k). \tag{4}$$

3.2 信任计算及数据结构

用户间的信任可通过直接信任和推荐信任得到.有交互经验的用户为直接用户,无交互经验但存在信任路径的用户为间接用户.

有交互经验的直接用户之间产生直接信任,无交互经验但存在信任路径的间接用户之间产生推荐信任.

$t_{j,b}$ 表示用户 j 对用户 b 的直接信任:

$t_{j,b}$ 存储在矩阵 $\boldsymbol{T}$ 中,$t_{j,b}\in[0,1]$.

$t^{\mathrm{r}}_{j,d}$ 表示用户 j 与用户 d 的推荐信任:

$t^{\mathrm{r}}_{j,d}$ 存储在矩阵 $\boldsymbol{T}$ 中,$t^{\mathrm{r}}_{j,d}\in[0,1]$.

其中,信任值的范围为[0,1],信任值为 0 表示完全不信任,信任值为 1 表示完全信任.

设定用户间的信任路径可通过至多 6 个中间用户得出.用户对项目相关内容兴趣度越高,其评价的可信度越高.$x_1,x_2,\cdots,x_g$ 表示用户 j 与用户 d 之间的中间用户,用户 j 与用户 d 之间引入兴趣度 I 后的间接信任值 t^{r} 为

$$t^{\mathrm{r}}_{j,d}=\frac{t_{j,x_1}+t_{x_1,x_2}+\cdots+t_{x_g,d}}{|X^d_j|}\times I_d,\quad x_1,x_2,\cdots,x_g\in X^d_j, \tag{5}$$

其中,X^d_j 是用户 j 与用户 d 之间的所有用户的集合.

3.3 用户评价建议处理

1) 直接信任预期评估.根据直接用户 b 做出的评价建议 s^i_b,用户 j 对项目 i 做出的来自直接信任的预期评价 $s^i_{j,b}$ 计算为

$$s^i_{j,b}=t_{j,b}\times s^i_b. \tag{6}$$

2) 间接信任预期评估.根据间接用户 d 的评价建议 s^i_d,用户 j 对项目 i 做出的来自间接信任的预期评价 $rs^i_{j,d}$ 计算为

$$rs^i_{j,d}=t^{\mathrm{r}}_{j,d}\times s^i_d. \tag{7}$$

其中,rs^i 与 s^i 均存储在矩阵 $\boldsymbol{S}_i$ 中,$\boldsymbol{S}_i$ 为初始用户 j 通过不同用户得到的对项目 i 的预期评价存储矩阵,用户评价分值范围为[0,10].

现实中,用户 j 可能接收到来自多个与其存在信任关系的用户的建议,因此,需要对用户 j 收到的所有建议评价信息进行处理,以得到用户 j 最终的评价预期.使用 e 表示向用户 j 进行推荐的直接信任用户,f 表示向用户进行推荐的间接信任用户,利用式(8)处理 $\boldsymbol{S}_i$ 中的评分信息,可以得到用户 j 对项目的最终预期评价 s^i_j:

$$s^i_j=\frac{\sum_e t_e\times s^i_e+\sum_f t^{\mathrm{r}}_f\times rs^i_f}{\sum_e t_e+\sum_f t^{\mathrm{r}}_f}. \tag{8}$$

用户 j 接收到多项推荐时,将计算得到每个项目的预期评价,接受预期评价最高的项.

4 DQN 信任增强过程

本节详细介绍了 DQN 信任增强过程(trust boost via deep q-learning, DQN-TB),说明了该过程的方法和流程,并给出了相应的伪代码.图 2 给出了 DQN-TB 过程的流程图框架.

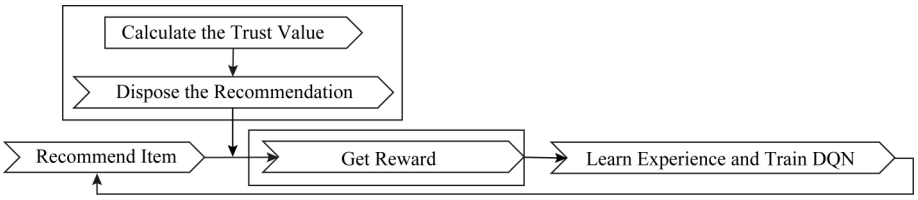

Fig. 2 Flow chart of DQN-TB

图 2 DQN-TB 过程流程图

需要说明,DQN-TB 过程中可以存在 1 对 1 关系,即只有用户 u_1 期望提高 u_2 对自己的信任值 t_{u_1,u_2};也可以存在多对 1 的关系,即用户 $u_1,u_2,\cdots$ 均期望提高用户 u_3 对自己的信任值.

4.1 模型框架

如图 3 所示,项目集中每个项目分别对应不同动作,DQN-TB 方法将用户视为环境主体,在每个状态通过记忆池中的数据进行训练,并从项目集中

选择项目作为最终动作向用户推荐项目，同时获得用户返回的奖励并将其与状态、动作存入记忆池中进行下一次网络训练更新.

本过程使用 Q 网络预期回报，并根据图 2 的过程流程及图 3 的框架图，更新网络：

Step1. 推荐项目

用户在时间 τ 发出广播，推荐者根据用户广播中的项目要求，从对应项目集中选择项目进行推荐，所选项对应动作 a_τ，推荐者状态记为 n_τ.

Step2. 信任更新

用户收到并接受推荐项目后，预期评价与实际评价的差值将影响用户对推荐者的信任.将用户视为 DQN-TB 过程的环境，信任值 t 随着用户对推荐项目的满意度进行更新，用户的信任变化幅度 Δt 将作为 DQN-TB 过程的奖励值.

Δt 与动作 a_τ、推荐者状态 n_τ 和未来状态 $n_{\tau+1}$ 存储在记忆池中，并作为网络输入.

Step3. 网络训练学习

DQN-TB 过程使用记忆池中的数据与 Q 网络进行动作预期选择，通过 target-Q 网络模拟用户现实，并根据 Q 网络与 target-Q 网络的差值更新 Q 网络.

Step4. 重复 Step1～Step3.

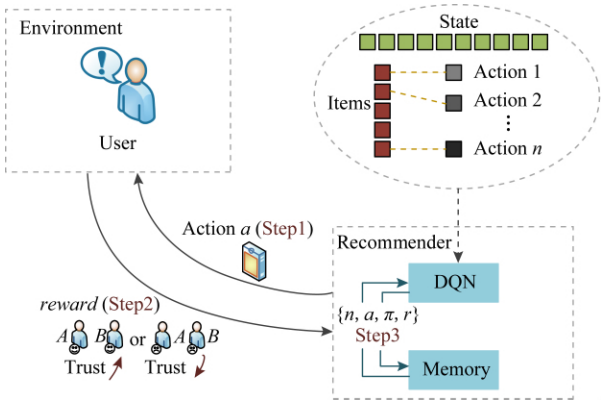

Fig. 3 Framework steps of DQN-TB

图 3 DQN-TB 过程步骤框架

考虑到现实情况，用户间的推荐过程一旦建立，将不会无条件停止.用户对推荐者信任值过低时，推荐者提供的意见不会被用户采纳.因此，若推荐者始终得到负向奖励或推荐失败次数过多，将会终止循环.为使推荐者可拥有更多机会进行推荐学习，同时考虑到实际情况，信任值过低时进行推荐不合现实，通过实验验证，发现当设定 $t<0.2$ 时终止推荐，会得到较好的结果.并且，为防止信任值溢出，规定当 t 更新结束后，$t>1$ 时，取 $t=1$.

4.2 DQN-TB 设计

考虑信任的动态性以及推荐项目的具体过程，本方法使用 Q 网络来估计推荐者选择某项目进行推荐(即动作 a_τ)的回报.将信任变化程度作为奖励值后，项目选择回报可模型化为

$$Q(n_\tau, a_\tau; \omega) \approx Q^\pi(n_\tau, a_\tau), \tag{9}$$

其中，ω 为 Q 网络权重参数，π 为所选策略.考虑到推荐者与用户可能无交互经验，规定推荐策略：

1) 初次推荐.所有备选项目选择概率相等，随机选择推荐项.

2) 后续推荐.使用 ε-greedy policy，选择概率根据推荐结果动态变化，最终收敛.

根据马尔可夫性，随机状态中下一时刻的状态只与当前状态有关.因此 DQN-TB 过程通过 Q 网络计算动作概率并作出动作选择后，会得到奖励值 Δt 及下一步的状态 $n_{\tau+1}$.同时，实际回报由 target-Q 网络模拟计算：

$$Q(n_\tau, a_\tau) = \Delta t_\tau + \lambda \max_{a_{\tau+1}} Q(n_{\tau+1}, a_{\tau+1}; \bar{\omega}), \tag{10}$$

其中，λ 是折现系数，用来平衡即时回报与未来回报；Δt_τ 表示动作 a_τ 获得的奖励.$\bar{\omega}$ 为 target-Q 网络的权重参数.target-Q 网络模拟计算实际汇报后，利用 Bellman 方程思想计算目标函数：

$$L(\omega) = E[(\Delta t_\tau + \lambda \max_{a_{\tau+1}} Q(n_{\tau+1}, a_{\tau+1}; \bar{\omega}) - Q(n_\tau, a_\tau; \omega))^2]. \tag{11}$$

同时，对目标函数 $L(\omega)$ 使用随机梯度下降，即

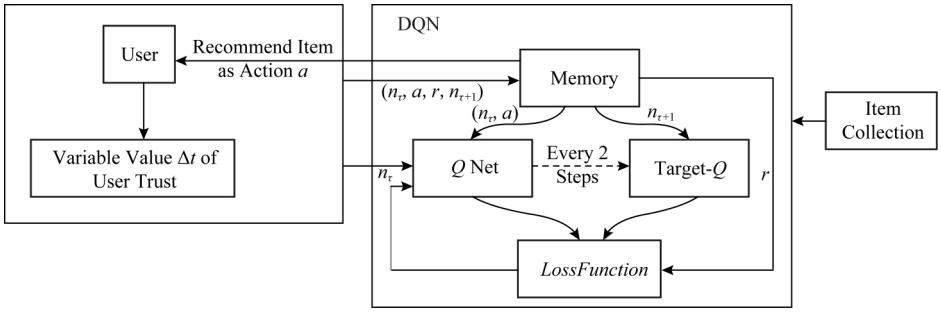

Fig. 4 Data transmission process of DQN-TB

图 4 DQN-TB 过程数据传输流程图

可更新 Q 网络参数 ω.DQN-TB 过程中,Q 网络与 target-Q 网络结构相同,target-Q 网络的权重参数与 Q 网络权重参数相同,每 2 次迭代同步 1 次.

图 4 给出了 DQN-TB 过程的数据传输流程过程.Q 网络与 target-Q 网络提取记忆池中数据进行计算,根据计算出的 Q_eval 值从项目集中选取项目作为动作进行推荐.用户接受项目后,实际体验会使用户信任发生改变.用户信任更新后,信任变化值返回 DQN-TB,Q 网络的权重根据 $LossFunction$ 进行更新,并进行下一轮迭代.DQN-TB 过程中,Q 网络每 2 步将网络权重传输至 target-Q 网络.

4.3 奖励参数 Δt 设置及信任更新

静态信任由于数值固定,无法准确表示用户在未来的信任关系,这一问题导致许多推荐算法不能响应用户关系及用户偏好的改变,使推荐结果的准确性降低.而随着经验累积,动态信任中的推荐者可及时响应用户的偏好改变,从而使推荐结果愈加精准.

已有部分学者将 DQN 方法应用于推荐系统,这使推荐过程保持长久的动态性.Zheng 等人[14]提出了一种应用于新闻推荐的深度强化学习框架.该方法根据用户特征及行为反馈计算动态奖励值,使推荐系统能够捕捉用户偏好的改变,从而获得长久的奖励,并保持用户对推荐系统的兴趣.

本文提出的方法受到文献[14]的启发,考虑到信任的动态变化特性及长期推荐过程的经验学习,将信任动态变化幅度 Δt 作为奖励值,采用 DQN 进行过程建模.

信任动态变化幅度 Δt 可通过评分差值与信任变化的相关关系灵活刻画.由 3.3 节可知,推荐者 u_2 向用户 u_1 推荐项目 i 后,u_1 将根据 u_2 的评分与信任得出预期评价 $s^i_{u_1,u_2}$,并根据最终的预期评价总和选择接受项.u_1 对项目 i 做出实际评价 $sp^i_{u_1}$ 后,令 $e^i_{u_2}=sp^i_{u_1}-s^i_{u_1,u_2}$,$e^i_{u_2}$ 可衡量推荐项 i 与 u_1 的偏好差距.偏好差距表示用户对推荐者推荐内容的满意度,它将影响用户在未来对推荐者建议的参考程度,因此,$e^i_{u_2}$ 与 t_{u_1,u_2} 存在正相关关系.

以 u_1,u_2 分别为用户和推荐者为例,使用 LMS 算法对信任变化及更新过程进行模拟:

由于实际信任更新过程中评价误差及信任均基于单个用户(即 u_1 和 u_2),误差成本函数定义为

$$E(t)=\frac{1}{2}(sp^i_{u_1}-s^i_{u_1,u_2})^2. \tag{12}$$

更新梯度 g 定义为

$$g(u_2)=\Delta E(t)=\frac{1}{\partial t}\partial\left(\frac{1}{2}(sp^i_{u_1}-s^i_{u_1,u_2})^2\right)=-\eta e^i_{u_2}s^i_{u_1,u_2}, \tag{13}$$

其中,η 为学习率参数,$e^i_{u_2}\in[-10,10]$.

为保证数值计算合理性,防止信任更新值溢出,更新梯度 g 被约束为

$$g(u_2)\leftarrow 0.01g(u_2)=-0.01\eta e^i_{u_2}s^i_{u_1,u_2}. \tag{14}$$

通过计算用户 u_1 的实际评分与预期评分的差值,可对用户 u_1 与推荐者 u_2 的信任进行更新.若 u_1 与 u_2 的信任关系为直接信任,两者信任更新表示为

$$t_{u_1,u_2}\leftarrow t_{u_1,u_2}-g(u_2)=t_{u_1,u_2}+0.01\eta e^i_{u_2}s^i_{u_1,u_2}. \tag{15}$$

同样地,若 u_1 与 u_2 的信任关系为间接信任,更新表示为

$$t^r_{u_1,u_2}\leftarrow t^r_{u_1,u_2}-g(u_2)=t^r_{u_1,u_2}+0.01\eta e^i_{u_2}s^i_{u_1,u_2}. \tag{16}$$

当信任更新完成后,评分矩阵 $\boldsymbol{S}$ 中用户 u_1 对项目 i 的评价 $s^i_{u_1}$ 将更新为实际评价 $sp^i_{u_1}$:

$$s^i_{u_1}\leftarrow sp^i_{u_1}. \tag{17}$$

推荐者 u_2 获得的奖励值为 $0.01g(u_2)=\Delta t$.

DQN-TB 过程中,用户 u_1 接受用户 u_2 的推荐并做出实际评价后,该奖励值将作为参数输入网络中进行下一步计算.

4.4 Markov 决策过程参数

表 1 给出了 DQN-TB 过程的 Markov 决策过程相关定义.

Table 1 Parameters of Markov Decision Process

表 1 Markov 决策过程参数

Parameter	Numerical Definition
State Space	∞
Action Space	$\lvert I\rvert$
Immediate Reward	Δt
γ	0.9

用户推荐过程不会无条件停止,因此,用户状态数将随着推荐过程不断增加.推荐过程中的动作为推荐项目,因此可选动作与项目集中的项目数量相关.通过查阅相关文献和参考资料,本文设定 $\gamma=0.9$.

4.5 算法伪代码

算法 1. DQN-TB 算法.

① 初始化记忆池 D 的容量 N;

② 初始化 Q 网络的权重 ω;

③ 初始化 target-Q 网络权重 $\bar{\omega}=\omega$;

④ for ($episode=1$) do

⑤ 初始化序列 $n_1=\{x_1\}$,序列预处理 $\phi_1=\phi(n_1)$;

⑥ for ($\tau=1$) do

⑦ 初次推荐使用随机概率 ε 选择动作 a_τ;

⑧ 后续推荐 $a_\tau=\arg\max_a Q(\phi(n_t),a;\omega)$;

⑨ if (accept) do

⑩ 得到动作 a_τ 的奖励值 Δt,载入 $x_{\tau+1}$;

⑪ 令 $n_{\tau+1}=n_\tau,a_\tau,x_{\tau+1}$,预处理 $\phi_{\tau+1}=\phi(n_{\tau+1})$;

⑫ 将 $(\phi_\tau,a_\tau,r_\tau,\phi_{\tau+1})$ 存储至 D;

⑬ D 中获取样本 $(\phi_j,a_j,\Delta t_j,\phi_{j+1})$;

⑭ Set

$$y_i=\begin{cases}\Delta t_j,\\ \text{如果 } episode \text{ 在 } j+1 \text{ 步停止};\\ \Delta t_j+\lambda\ \max_{a'} Q(\phi_j,a_j;\omega),\\ \text{otherwise};\end{cases}$$

⑮ 对 $(y_i-Q(\phi_j,a_j;\omega))^2$ 进行梯度下降更新 ω;

⑯ 每 2 步重置 $\bar{\omega}=\omega$;

⑰ else back to ⑦;

⑱ end if

⑲ end for

⑳ end for

4.6 计算复杂度分析

DQN-TB 使用了随机梯度下降方法进行参数更新,因此,可知 DQN-TB 算法的计算复杂度为 $T(n)=(C+n)\times n\times n\times n\approx T(n^4)=O(n^4)$.可知算法复杂度为多项式级别.

5 实 验

本节将对 DQN-TB 过程中推荐信任与直接信任的转化比例、奖励参数计算中的信任更新学习率进行说明,同时说明 DQN-TB 过程的信任增强效果,并对 DQN-TB 应用于推荐系统后的性能给出了相应的对比验证,包括推荐成功率与感知用户偏好的动态变化.

5.1 基本介绍

本文使用仿真实验验证模型性能,来模拟推荐方向单个用户进行推荐,用户对推荐方的信任随推荐而变化的过程.实验环境基于 OpenAI Gym,其中,奖励参数值 *reward* 随着 DQN-TB 的每一轮推荐,根据 LMS 方法动态更新,并传输至 DQN-TB.实验数据使用从豆瓣采集的用户影评数据及电影项目类别,包括 10 个用户对 11 个电影项目类别中不同电影的评价数据,所有用户的观影总数为 300 部,影评数据规模为 510 条.实验从所有用户中随机选择用户作为目标用户,并进行推荐.

DQN-TB 过程目的为提高单个用户信任值,本实验中 Q 网络结构示意如图 5 所示,Q 网络从记忆池中提取数据输入到网络中,通过隐藏层计算 Q 值,并根据相应动作选择策略来选择最终动作.DQN-TB 过程中的状态为用户当前信任值,动作为 DQN-TB 可向用户推荐的项目.

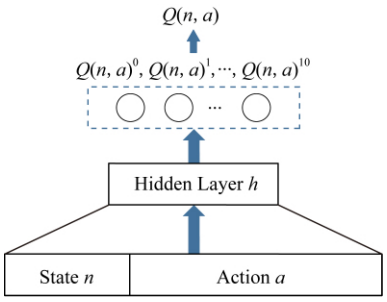

Fig. 5 Q network structure

图 5 Q 网络结构图

5.2 推荐信任与直接信任转化

用户项目推荐过程中,当通过推荐信任进行推荐后,用户间交互更新为直接信任,此时,为更符合现实情境,通过推荐信任计算出的信任值需进行一定折扣才可转化为直接信任值,并进行后续计算.推荐信任折扣因子由 μ 表示.

为确定 μ 的具体数值,本节使用 4 组小数据对选择不同折扣因子导致的结果变化进行分析.4 组数据分别对应高推荐信任与高推荐评价、高推荐信任与低推荐评价、低推荐信任与高推荐评价、低推荐信任与低推荐评价,同时,4 组数据中其他项均相同.分析使用的数据集由表 2 给出.推荐信任折扣因子 μ 采用不同数值时对结果影响如图 6 中整体评分项表示,对比评分项为仅使用 Direct Trust1 至 Direct Trust4 计算得出的预期评价.

由图 6 可知,当 μ 较低时,整体预期评价值低于对比评分;当 μ 较高时,推荐信任用户的评价结果对总结果起正向激励作用.该对比实验使用数据虽不能代表全部现实情况,但依旧可以反映推荐信任折扣因子 μ 对结果的影响.考虑到现实因素,当用户第一次进行直接信任推荐时,依旧会对评价主体用户有相应评分影响.为使用户评分由信任值影响,并尽量少的受到 μ 的干扰,本文设定 $\mu=0.8$.

Table 2 Trust Value and Score Value

表 2 信任及评分值表

Value Types	Direct Trust1	Direct Trust2	Direct Trust3	Direct Trust4	Recommend Trust1	Recommend Trust2	Recommend Trust3	Recommend Trust4
Trust Value	0.5	0.9	0.6	0.8	0.7	0.7	0.5	0.5
Score Value	6.5	7.0	7.0	6.5	9.0	5.5	9.0	5.5

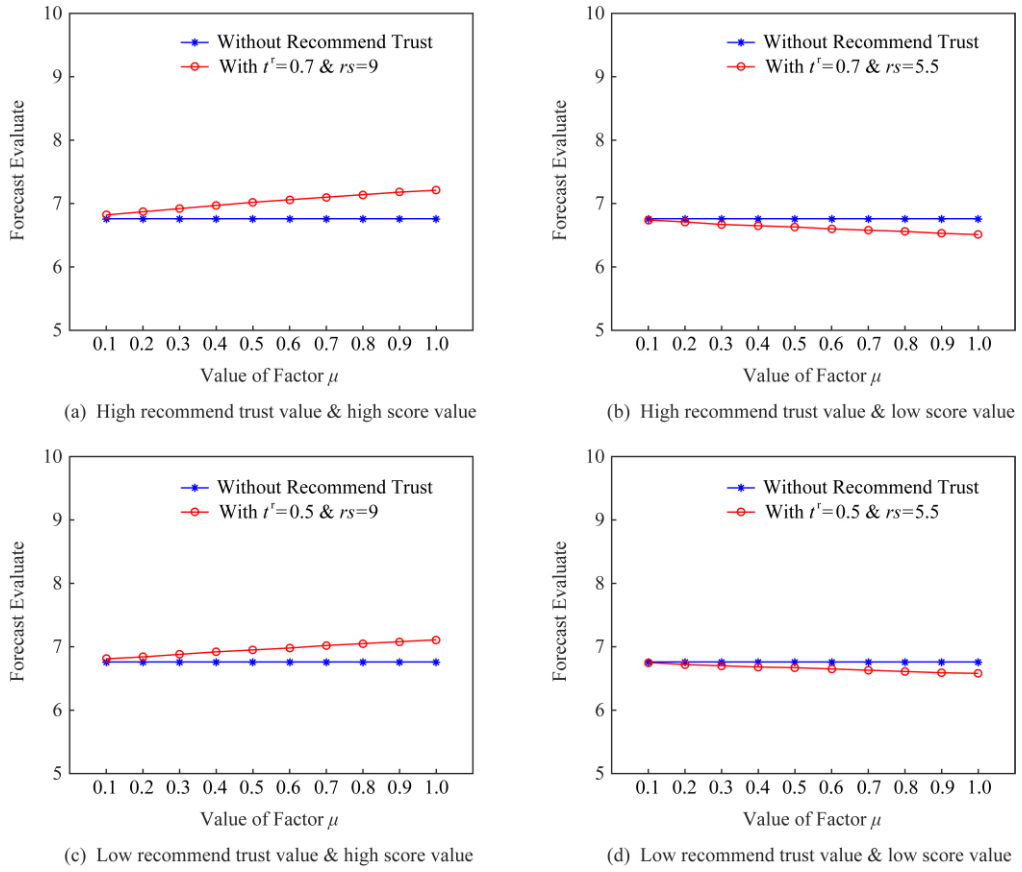

Fig. 6 Performance comparison on different μ

图 6 不同 μ 下预期评价比较

5.3 信任更新学习率

由 4.3 节可知,奖励参数 $reward$ 需要通过信任更新得到.信任更新的学习率 η 对更新结果有直接的影响.η 过小会使更新步长过小,收敛速度过慢;η 过大时收敛速度会提高,但可能因为步长过大而导致无法收敛.因此,本节将针对不同学习率对信任更新幅度的影响进行讨论.

图 7 展示了不同 η 取值对信任更新幅度的影响.为方便计算,本节计算信任更新幅度时设定 $s_j^i=5$.同时,η 取值为 0.1～0.9,可更全面显示学习率不同对信任更新幅度的改变.

信任更新时,根据建议用户与用户主体社会关系的远近,用户主体信任的更新幅度也会有所区别.本文将推荐用户分为直接用户和推荐用户.

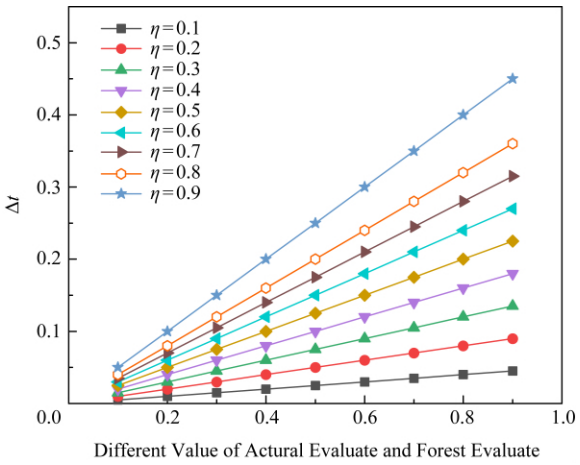

Fig. 7 Performance comparison on different η

图 7 不同 η 下信任更新幅度比较

推荐用户由于社会关系较远,不会被用户主体给予高度包容性,同时由于心理预期较低,推荐成功后用户主体的信任值将变化较大,因此推荐信任的更新步长相对较大.并且,由于信任值范围为[0,1],η 数值过高会导致信任变化过大,因此设定推荐用户信任更新学习率 $\eta=0.2$.

对于直接用户,由于社会关系近,用户主体会抱有更多包容性,直接用户比推荐用户单次信任更新步长相对小,因此本文设定直接信任用户的信任更新学习率 $\eta=0.1$.但直接信任用户信任更新存在累计作用,因此直接用户学习率设定为

$$\eta=\begin{cases}0.1, & \text{第 1 次推荐},\\ 0.1p, & \Delta t>0,\\ 0.1q, & \Delta t<0,\end{cases}$$

其中,p 为直接用户推荐得到正面反馈的次数,q 为得到负面反馈次数.

5.4 信任动态变化

根据 DQN-TB 过程,用户通过推荐结果学习到相关经验后,推荐选择将会进一步调整,以符合被推荐用户的相关兴趣偏好.图 8 给出了 DQN-TB 过程中随轮次增加的信任变化折线图,用户初始推荐信任值为 0.67.

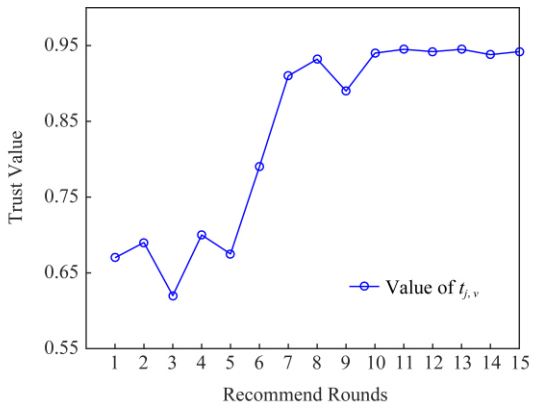

Fig. 8 The line chart of dynamic change of trust value

图 8 信任动态变化折线图

当用户间第 1 轮推荐结束后、第 2 轮推荐开始前,用户间信任将根据 5.2 节中转化率进行推荐信任—直接信任转换,使得信任值有一定程度的下降.由图 8 可知,当轮次较少时,DQN-TB 过程处在探索阶段,此时记忆池中的经验不够丰富,因此信任变化折线较为波折,由于最初用户间为推荐信任,因此当推荐者经验增加时,对被推荐用户的偏好的了解加深,此时用户间的信任值持续上升.并且由于成功经验的增多,后续推荐轮次中用户间信任值始终处于较高水平,且波动幅度很小.

DQN-TB 过程可较准确地刻画用户的信任变化状态,并取得较好的效果.用户信任的动态变化可以实时地反映用户偏好的变化以及社交关系的改变,因此 DQN-TB 的动态性研究是很有意义的,这一特性也为 DQN-TB 应用于推荐系统带来更多灵活性.

5.5 DQN-TB 应用于推荐系统

5.4 节中的实验验证了 DQN-TB 对于信任的动态变化及增强都有准确的刻画,因此该方法亦可应用于推荐系统中,为系统中的用户提供精准的推荐.本节将 DQN-TB 与 Li 等人[15] 的 CSIT 方法和 Gohari 等人[16] 提出的 CBR 方法进行了对比,并比较了三者向用户进行推荐的成功率及三者对用户偏好改变的响应灵敏度.

CSIT 方法是一种性能优越的矩阵因子分解和上下文感知推荐者法,作者同时提供了 GMM 方法进行增强并同时处理分类上下文和连续上下文.CBR 方法使用对用户意见的信任和意见的确定性来描述用户信心,并将用户信心引入信任建模,通过隐式信任模型向用户提供一系列的推荐.由于目前的推荐系统仅利用用户的社交关系以及关系网络中其他用户的偏好来进行相关推荐,无法反映用户的信任变化,且用户的偏好变化捕捉只能来源于关系网络中的其他用户,造成系统对用户偏好的变化反馈不及时、不准确.

图 9 给出了用户信任与偏好动态变化下,20 轮内不同轮次对应的 DQN-TB 方法、CSIT 方法和 CBR 方法的平均成功率对比.随着推荐轮次的增加,DQN-TB 拥有越来越高的准确率,这是由于 DQN-TB 对于动态变化响应的灵活性.同时,由于 CSIT 方法和 CBR 方法的对用户偏好感知的计算方法影响,两者

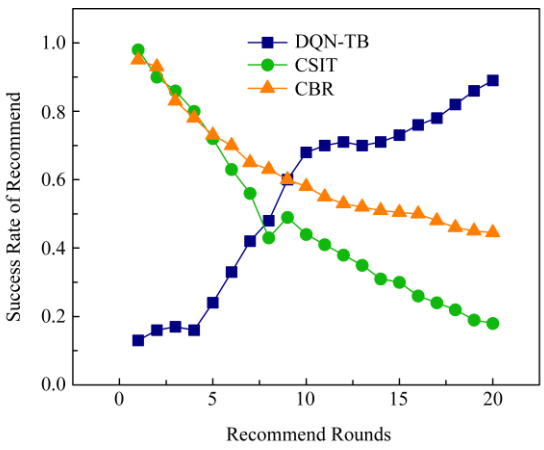

Fig. 9 Change of success rate in recommend system

图 9 推荐系统成功率变化

的准确性随条件的动态变化而逐渐下降.实际情况中,普通推荐者法准确率下降的速度随用户偏好的变化幅度而有所偏差,但亦足以说明 DQN-TB 的优越性.

5.6 响应灵敏度

定义 M_{ANR} 表示用户偏好变化后推荐系统响应变化所需要的轮次(answer rounds, ANR),M_{SUMR} 表示推荐总轮次(sum rounds, SUMR),M_{SEN} 表示推荐系统相应用户偏好变化的灵敏度(sensitivity, SEN),则 M_{SEN} 可计算为

$$M_{SEN}=1-\frac{M_{ANR}}{M_{SUMR}}.$$

表 3 给出了 450 轮推荐中用户偏好变化 60 次后各推荐系统的响应灵敏度.隐式信任模型与上下文矩阵分解过多的依赖用户的邻居及相似用户,当偏好改变多次时,相关信息的分析将失去其准确性,将无法及时反馈用户的偏好.CBR 方法同时为用户推荐多个项,因此具有一定的覆盖性,响应灵敏度优于 CSIT 模型.表 3 的数据验证了 DQN-TB 过程对用户偏好具有较好的灵敏度,这一特性与 DQN-TB 过程中的动态奖励及经验学习有关.因此,将 DQN-TB 过程应用到推荐系统可及时感知用户偏好的改变,并相应地调整推荐项目的选择.

Table 3 Response Sensitivity of Each Recommendation System

表 3 各推荐系统响应灵敏度

Recommendation System	M_{ANR}	M_{SEN}
DQN-TB	84	0.813
CSIT	317	0.295
CBR	259	0.424

6 总结及未来展望

本文结合强化学习方法提出了一种基于动态信任的信任增强方法,该方法通过用户信任的动态变化感知用户偏好的变化,并根据推荐经验进行学习,以提供更加准确的推荐,从而使用户信任增加并保持在较高水平.实验表明:所提出的方法是高效、准确的.同时,本方法也可应用于推荐系统,并达到感知用户偏好变化、进行精准推荐的目的.

本文的方法重点考虑用户信任的动态变化,未来,将针对信任及建议计算方法进行改进以使推荐的结果更加精准、有效.

参 考 文 献

[1] Jiang Cuiqin, Liu Shixi, Lin Zhangxi, et al. Domain-aware trust network extraction for trust propagation in large-scale henterogeneous trust networks [J]. Knowledge-Based Systems, 2016, 111(C): 237-247

[2] Yan Surong, Lin K, Zheng Xiaolin, et al. An approach for building efficient and accurate social recommender systems using individual relationship networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(10): 2086-2099

[3] Ghavipour M, Meybodi M. Trust propagation algorithm based on learning automata for inferring local trust in online social networks [J]. Knowledge-Based Systems, 2018, 143: 307-316

[4] Ghavipour M, Meybodi M. A dynamic algorithm for stochastic trust propagation in online social networks: Learning automata approach [J]. Computer Communications, 2018, 123: 11-23

[5] You Jing, Shangguan Jinglun, Zhuang Lihua, et al. An autonomous dynamic trust management system with uncertainty analysis [J]. Knowledge-Based Systems, 2018, 161: 101-110

[6] Tong Xiangrong, Zhang Wei, Long Yu, et al. Subjectivity and objectivity of trust [C] //Proc of Agents and Data Mining Interaction. Berlin: Springer, 2013: 105-114

[7] Tong Xiangrong, Zhang Wei, Long Yu. Transitivity of agent subjective trust [J]. Journal of Software, 2012, 23(11): 2862-2870 (in Chinese)
(童向荣, 张伟, 龙宇. Agent 主观信任的传递性[J]. 软件学报, 2012, 23(11): 2862-2870)

[8] Liu Guanfeng, Liu Yi, Liu An, et al. Context-aware trust network extraction in large-scale trust-oriented social networks [J]. World Wide Web, 2018, 21(3): 713-738

[9] Liu Zhiquan, Ma Jianfeng, Jiang Zhongyuan, et al. FCT: A fully-distributed context-aware trust model for location based service recommendation [J]. Science China: Information Sciences, 2017, 60(8): 101-116

[10] Tong Xiangrong, Huang Houkuan, Zhang Wei. Prediction and abnormal behavior detection of Agent dynamic interaction trust [J]. Journal of Computer Research and Development, 2009, 46(8): 1364-1370 (in Chinese)
(童向荣, 黄厚宽, 张伟. Agent 动态交互信任预测与行为异常检测模型[J]. 计算机研究与发展, 2009, 46(8): 1364-1370)

[11] Wang E, Li Yueping, Ye Yunming, et al. A dynamic trust framework for opportunistic mobile social networks [J]. IEEE Transactions on Network and Service Management, 2018, 15(1): 319-329

[12] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533

[13] Watkins C. Learning from delayed rewards [D]. Cambridge, UK: King's College, 1989

[14] Zheng Guanjie, Zhang Fuzheng, Zheng Zihan, et al. DRN: A deep reinforcement learning framework for news recommendation [C] //Proc of WWW 2018. New York: ACM, 2018: 168-176

[15] Li Jun, Chen Chaochao, Chen Huiling, et al. Towards context-aware social recommendation via individual trust [J]. Knowledge-Based Systems, 2017, 127: 58-66

[16] Gohari F, Aliee S, Haghighi H. A new confidence-based recommendation approach: Combining trust and certainty [J]. Information Sciences, 2018, 422: 21-50

Qi Faxin, born in 1995. Master candidate in the School of Computer and Control Engineering at Yantai University. Student member of CCF. Her main research interests include propagation models of trust.

Tong Xiangrong, born in 1975. PhD in computer science and technology from Beijing Jiaotong University. Member of CCF. Full professor with Yantai University. His main research interests include computer science, intelligent information processing and social networks.

Yu Lei, born in 1976. Received his PhD degreee from Arizona State University. Associate professor with Tenure at State University of New York at Binghamton and an adjunct professor at Yantai University. His main research interests include machine learning, data mining and artificial intelligence.

更正声明

刊登在《计算机研究与发展》2020 年第 57 卷第 5 期上的论文“图灵测试的明与暗”第 2 节中，由于作者疏忽，将素数的定义误写为：

它的内涵表示就是“只能被 1 和自身整除的自然数”这个命题；它的外延表示就是集合{1,2,3,5,7,11,13,17,19,23,29,…}.

正确的表述应该是：

它的内涵表示就是“只能被 1 和自身整除的大于 1 的自然数”这个命题；它的外延表示就是集合{2,3,5,7,11,13,17,19,23,29,…}.

特此更正，并向广大读者致歉！

作者：于剑

2020 年 5 月 20 日