Journal of Computer Applications
计算机应用,2021,41(7):1871 - 1877
ISSN 1001-9081
CODEN JYIIDU
2021-07-10
http://www.joca.cn

文章编号:1001-9081(2021)07-1871-07
DOI:10.11772/j.issn.1001-9081.2020111745

基于知识图谱和重启随机游走的跨平台用户推荐方法

余敦辉[1,2],张蕗怡[1],张笑笑[1*],毛　亮[1]
(1.湖北大学 计算机与信息工程学院,武汉 430062;
2.湖北省教育信息化工程技术研究中心(湖北大学),武汉 430062)
(*通信作者电子邮箱zxxhubu@163.com)

摘　要:针对单一社交网络平台中推荐相似用户结果单一,对用户兴趣和行为信息了解不够全面的问题,提出了基于知识图谱和重启随机游走的跨平台用户推荐方法(URCP-KR)。首先,在分割、匹配出的目标平台图谱和辅助平台图谱的相似子图中,利用改进的多层循环神经网络(RNN)预测出候选用户实体,再综合利用拓扑结构特征相似度和用户画像相似度筛选出相似用户;然后,将辅助平台图谱中的相似用户的关系信息补全到目标平台图谱;最后,计算目标平台图谱中的用户游走到社区内每个用户的概率,从而得到用户之间的兴趣相似度来实现用户推荐。实验结果表明,与协同过滤(CF)算法、基于跨平台的在线社交网络用户推荐算法(URCP)和基于多开发者社区的用户推荐算法(UR-MC)相比,URCP-KP在推荐精确率及推荐多样性等方面均有所提高,推荐精确率最高可达95.31%,推荐覆盖率最高可达88.42%。

关键词:知识图谱;实体链接;关系补全;重启随机游走;用户推荐
中图分类号:TP391　　**文献标志码**:A

User recommendation method of cross-platform based on knowledge graph and restart random walk

YU Dunhui[1,2], ZHANG Luyi[1], ZHANG Xiaoxiao[1*], MAO Liang[1]
(1. *School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei* 430062, *China*;
2. *Hubei Provincial Education Informationization Engineering and Technology Center (Hubei University), Wuhan Hubei* 430062, *China*)

Abstract: Aiming at the problems of the single result of recommending similar users and insufficient understanding of user interests and behavior information for single social network platforms, a User Recommendation method of Cross-Platform based on Knowledge graph and Restart random walk (URCP-KR) was proposed. First, in the similar subgraphs segmented and matched by the target platform graph and the auxiliary platform graph, an improved multi-layer Recurrent Neural Network (RNN) was used to predict the candidate user entities. And the similar users were selected by comprehensive use of the similarity of topological structure features and user portrait similarity. Then, the relationship information of similar users in the auxiliary platform graph was used to complete the target platform graph. Finally, the probabilities of the users in the target platform graph walking to each user in the community were calculated, so that the interest similarity between users was obtained to realize the user recommendation. Experimental results show that the proposed method has higher recommendation precision and diversity than Collaborative Filtering (CF) algorithm, User Recommendation algorithm based on Cross-Platform online social network (URCP) and User Recommendation algorithm based on Multi-developer Community (UR-MC) with the recommendation precision up to 95.31% and the recommendation coverage up to 88.42%.

Key words: knowledge graph; entity linking; relationship completion; restart random walk; user recommendation

0　引言

随着互联网及大数据时代的发展,信息呈爆炸式增长,面对海量数据,用户如何高效获取自己需要的信息愈发困难,因此推荐系统成为各大社交平台争先关注和应用的焦点。现有的推荐系统大多建立在单一社交平台的基础上,仅仅利用单一平台中用户的兴趣、发布的内容和历史记录等进行用户推荐,导致对用户兴趣和行为信息了解不够全面,推荐结果单一,容易出现数据过度拟合现象[1]以及用户冷启动问题[2]。例如:用户在平台1上喜欢关注体育类型的博主,在平台2上喜欢关注科技类型的用户,并且对科技类型信息的关注活跃度高于对体育类型的关注,但是平台1一直推送关于体育类型的信息和博主给用户,未能发掘用户对科技类型信息和博主的兴趣,缺乏对用户兴趣信息的全面了解。

但是,融合多个社交网络中的数据,存在用户信息获取困难、关联效率较低的问题,因此,如何在跨平台社交网络中进

收稿日期:2020-11-09;**修回日期**:2021-01-30;**录用日期**:2021-02-03。　**基金项目**:国家重点研发计划项目(2018YFB1003801);国家自然科学基金面上项目(61977021);湖北省技术创新专项(重大项目)(2018ACA13)。

作者简介:余敦辉(1974—),男,湖北武汉人,教授,博士,CCF会员,主要研究方向:知识图谱、大数据、服务计算、众包数据管理;　张蕗怡(2000—),女,湖北武汉人,主要研究方向:知识图谱;　张笑笑(1995—),女,山东滨州人,硕士研究生,CCF会员,主要研究方向:知识图谱、众包数据管理;　毛亮(1998—),男,湖北武汉人,硕士研究生,主要研究方向:知识图谱。

行用户推荐已成为一个研究热点。

目前众多学者围绕社交网络中的用户推荐展开研究，其中一部分是研究单一平台中的用户推荐：文献[3]中的协同过滤(Collaborative Filtering, CF)是使用较广泛的推荐技术，根据用户对物品的评分信息来预测用户的偏好；然而，协同过滤系统容易受到不公平评分和稀疏数据的影响。文献[4]中通过基于用户相似性的重启随机游走推荐社交网络中的事件给用户，但忽略了跨平台用户之间的关联。文献[5]中提出了基于用户行为特征挖掘的个性化推荐算法，采用显著数据分块检测方法进行社交网络用户特征的行为信息融合处理；但是仅融合单一平台的用户信息，推荐的多样性仍受到限制。单一平台用户信息有限，推荐结果单一，因此也有很多研究是围绕跨平台和跨社区的用户推荐展开的：文献[6]中提出利用源平台的数据丰富目标平台的数据来进行推荐，以此解决目标平台的数据稀疏问题和冷启动问题；但是源平台中的数据往往有限。文献[7]中提出跨平台用户推荐方法，通过爬取用户在不同平台的关联账号匹配关系，根据平台的应用程序编程接口(Application Programming Interface, API)获取用户数据；然而无法获取没有关联账号的用户数据，且很多平台没有提供API，用户匹配较为困难。文献[8]中分析了用户与用户之间的互动，建立了跨社区的开发者网络，利用重启随机游走实现对开发者的用户推荐；但是该方法只考虑了用户的文本和标签相似性，仅在特定社区效果较好，扩展性较差。文献[9]中提出了结合用户社区和评分矩阵联合社区的推荐模型，在面向联合社区的矩阵中结合用户社区进行矩阵分解；但该模型仅考虑了用户静态兴趣偏好与社交关系，忽略了用户的兴趣和关系的动态变化。

综上，在推荐系统的研究中，面临推荐单一、跨平台用户匹配效率低的问题，因此，本文提出基于知识图谱和重启随机游走的跨平台用户推荐方法(User Recommendation method of Cross-Platform based on Knowledge graph and Restart random walk，URCP-KR)，该方法包括基于知识图谱的跨平台用户关系补全算法(Cross-Platform user Relationship Completion algorithm based on Knowledge Graph, RCCP-KG)和基于重启随机游走的用户推荐算法(User Recommendation algorithm based on Restart Random Walk，UR-RRW)。首先，将目标平台图谱和辅助平台图谱分割匹配得到相似子图，利用多层循环神经网络(Recurrent Neural Network, RNN)得到候选用户实体，通过用户筛选匹配两个不同平台中可能相同的用户；然后，通过关系补全将辅助平台中的用户信息补全到目标平台图谱，完善用户之间的关系；最后，利用重启随机游走算法，获得具有相似兴趣的用户推荐列表，从而解决跨平台用户匹配、推荐单一的问题，提高用户推荐的准确率及多样性。

本文的主要工作如下：

1)提出了一种新的跨平台数据融合算法，基于知识图谱的跨平台用户关系补全算法RCCP-KG，把辅助平台中的相同用户信息补全到目标平台图谱中，实现跨平台数据融合，从而更加全面地刻画用户行为，同时发掘不同平台间的潜在用户关系，更加准确地进行相似用户推荐。

2)提出了一种基于重启随机游走的用户推荐算法UR-RRW，在传统随机游走的基础上融入用户兴趣信息，提升了推荐的多样性。

1 整体方案

1.1 相关定义

定义1 社交网络知识图谱(Social Network Knowledge Graph, SN-KG)。本文通过社交网络有向图构建SN-KG，将用户作为知识图谱中的实体，用户间的关注关系作为关系，用户属性作为实体属性，用户属性值作为实体属性值。

定义2 目标平台图谱 $G=(V,E)$，其中：$V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_i\}$是目标平台图谱中用户节点的集合，$\boldsymbol{v}_i$表示用户节点向量；$E=\{\boldsymbol{e}_{1,2},\boldsymbol{e}_{2,3},\cdots,\boldsymbol{e}_{i,j}\}$是目标平台图谱中用户节点之间边的集合，$\boldsymbol{e}_{i,j}$表示用户节点$\boldsymbol{v}_i$和$\boldsymbol{v}_j$之间的边。

定义3 辅助平台图谱 $G'=(V',E')$，其中：$V'=\{\boldsymbol{v}_1',\boldsymbol{v}_2',\cdots,\boldsymbol{v}_i'\}$是辅助平台图谱中用户节点的集合，$\boldsymbol{v}_i'$表示用户节点；$E'=\{\boldsymbol{e}_{1,2}',\boldsymbol{e}_{2,3}',\cdots,\boldsymbol{e}_{i,j}'\}$是辅助平台图谱中用户节点之间边的集合，$\boldsymbol{e}_{i,j}'$表示用户节点$\boldsymbol{v}_i'$和$\boldsymbol{v}_j'$之间的边。

定义4 实体关系三元组表示为(头实体，关系，尾实体)(h,r,t)，边r的类型统一为“关注”关系，即用户h关注用户t。若干三元组的集合$H=\{(h,r,t)\,|\,h\subseteq V,r\subseteq E,t\subseteq V\}$构成一个知识图谱，如图1所示，社交网络知识图谱可以表示用户之间的关注关系。实体属性三元组表示为(实体，属性，属性值)$(\boldsymbol{v}_i,a_i,Va)$，其中用户属性集合为$A_i=\{a_1,a_2,\cdots,a_i\}$。

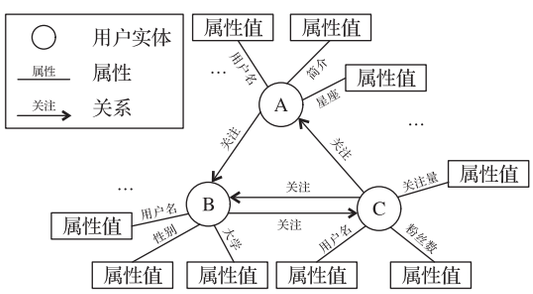


图1 社交网络知识图谱示意图

Fig. 1 Schematic diagram of social network knowledge graph

定义5 目标平台图谱子图集合$SG=\{g_1,g_2,\cdots,g_i\}$，是通过对目标平台图谱进行分割得到的子图集合，其中g_i表示目标平台图谱的子图。

定义6 辅助平台图谱子图集合$SG'=\{g_1',g_2',\cdots,g_j'\}$，是通过对辅助平台图谱进行分割得到的子图集合，其中g_j'表示辅助平台图谱的子图。

定义7 用户活跃时间关键词序列$(\boldsymbol{v}_i,K_n)$，其中用户$\boldsymbol{v}_i$在活跃时间t内的发表的博文内容关键词$\boldsymbol{k}_n$按照时间先后顺序构成$K_n=\{\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_n\}$。

1.2 整体框架

本文设计和实现了基于知识图谱和重启随机游走的跨平台用户推荐方法，其中包括RCCP-KG算法和UR-RRW两个算法。具体实现过程如下：

1)利用社区检测实现目标平台图谱和辅助平台图谱的图分割和子图匹配，得到相似子图；

2)通过输入用户活跃时间关键词序列训练RNN模型得到候选用户实体集合；

3)通过网络拓扑结构特征相似度和用户画像相似度对用户实体进行筛选，得到跨平台相同用户，实现实体链接；

4)将相同用户在辅助平台图谱的关系补全到目标平台图谱；

5)通过重启随机游走在目标平台图谱的社区中找到兴趣相似程度高的用户,形成用户推荐列表,实现跨平台用户推荐。

方法整体框架如图2所示。

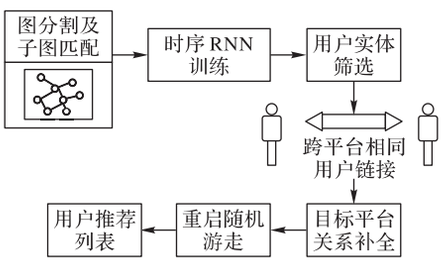


图2　本文方法整体框架

Fig. 2　Overall framework of the proposed method

2　基于知识图谱的跨平台用户关系补全算法

用户在不同平台所关注的内容和兴趣点可能不同,但同一用户的个人信息、好友以及影响力等具有极大的相似性。因此,识别两个平台中的相同用户问题,转换为匹配两个知识图谱中相同用户节点的问题。本章提出 RCCP-KG 算法。首先,利用社区检测算法将目标平台图谱和辅助平台图谱分割为不同的子图,匹配跨平台中相似的子图;然后在相似子图中利用多层循环神经网络预测候选实体,再通过图谱的拓扑结构、用户画像相似度筛选出相似度最高的用户节点对;最后将相同用户在辅助平台图谱中的关系补全到目标平台图谱,实现跨平台间的用户关系补全。

2.1　基于社区检测的图谱分割及子图匹配

由于社交网络图谱用户数量庞大,直接匹配跨平台间的相似用户工作量较大,且误差较高,因此,本文首先将图谱分割为相似子图,然后在子图中匹配相似用户。

Louvain 是一种基于模块度的社区发现算法[10],通过模块度来衡量一个社区的紧密程度。关联关注紧密的用户可以分割在同一子图中,能够更好地体现用户群体的特性,便于子图相似性匹配。

首先使用 Louvain 算法将目标平台图谱和辅助平台图谱分割为不同的子图集合 $SG=\{g_1,g_2,\cdots,g_i\}$ 和 $SG'=\{g_1',g_2',\cdots,g_j'\}$;然后,利用 Pearson 矩阵相似度计算方法[11]匹配目标平台图谱子图集合和辅助平台图谱子图集合中相似子图对 (g_i,g_j');接下来在相似子图对中匹配相同节点对。

2.2　基于多层循环神经网络的用户实体预测

本节在相似子图对 (g_i,g_j') 中,利用改进的多层循环神经网络预测出候选用户实体,再综合利用拓扑结构特征相似度和用户画像相似度筛选出相似用户。

首先,利用目标平台图谱中的用户实体和博文关键词序列训练多层循环神经网络;然后,将辅助平台图谱中的实体输入多层循环神经网络,经过用户相似度计算,得到在辅助平台图谱中与输入实体匹配的多个候选用户实体;最后,利用网络拓扑结构特征相似度和用户画像相似度对候选用户实体进行筛选,选出相似度最高的一个用户实体,与输入实体 $\boldsymbol{v}_i$ 组成相似用户节点对 $\boldsymbol{v}_i=\boldsymbol{v}_j'$。

2.2.1　基于多层循环神经网络的候选用户实体预测

用户活跃在不同平台的时间可能相同也可能是错开的,但是发布的内容可能是相同的。目前很多社交网络平台允许用户分享相同的内容到其他的社交平台中,如知乎允许用户将发布的内容同时分享到微博。例如,18:00,A 用户在知乎上发布了一条动态,同时将此动态分享到微博。因此,用户在不同平台发布内容和时间是判断是否为同一用户的重要内容。并且,用户活跃时间关键词序列可以近似地看作是一个由词组成的句子。

RNN 是一种神经序列模型,已经在语言建模和机器翻译等许多自然语言处理任务上取得了优良的表现。知识图谱中的三元组(实体,关系,实体)由句子提取,本质上是具有序列性的。Guo 等[12]在 RNN 的基础上提出了一种用于知识图谱补全的深度序列模型(Deep Sequential model for Knowledge Graph completion, DSKG),此模型具有序列特性,给定三元组中的头实体和关系,能够预测尾实体。DSKG 使用不同的 RNN 单元处理实体 v 和关系 r。根据模型 DSKG,本文选取用户的博文关键词取代关系进行训练。

因此,本文设计了一种基于多层 RNN 的实体预测模型(Entity Prediction model based on Multilayer RNN, EP-MRNN),如图3所示,输入层为用户 $\boldsymbol{v}_i$ 活跃时间关键词序列 $(\boldsymbol{v}_i,K_n)$,然后通过多层 RNN 训练得到用户 $\boldsymbol{v}_i$ 隐藏状态 $\boldsymbol{h}_i$,中间层输入是辅助平台图谱用户,输出层是目标平台用户与辅助平台用户的相似度概率。

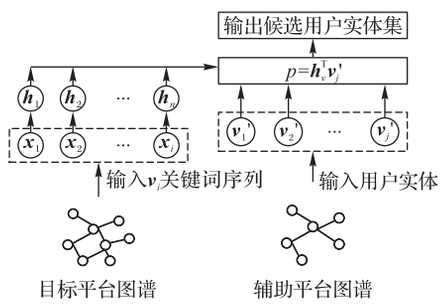


图3　基于多层 RNN 的实体预测模型

Fig. 3　Entity prediction model based on multi-layer RNN

如图4所示,给出了两层 RNN,其中 c_1、c_2、c_3、c_4 全都是各不相同的 RNN 单元。使用 c_1、c_2 处理实体 $\boldsymbol{v}_i$,使用 c_3、c_4 处理关键词 k_i。

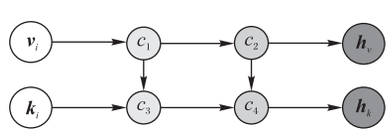


图4　两层 RNN 示意图

Fig. 4　Schematic diagram of two-layer RNN

c 表示一个 RNN 单元,它将之前的隐藏状态和当前元素作为输入,预测下一个隐藏状态,计算如下:

$$\boldsymbol{h}_t^i=\begin{cases}f(\boldsymbol{W}_h^i\boldsymbol{h}_{t-1}^i+\boldsymbol{W}_x^i\boldsymbol{x}_t+\boldsymbol{b}^i), & i=0\\ f(\boldsymbol{W}_h^i\boldsymbol{h}_{t-1}^i+\boldsymbol{W}_x^i\boldsymbol{h}_t^{i-1}+\boldsymbol{b}^i), & i>0\end{cases}\tag{1}$$

$$\begin{cases}\boldsymbol{W}_h^i=\begin{cases}V_h^i, & \boldsymbol{x}_t\in V\\ K_h^i, & \boldsymbol{x}_t\in K\end{cases}\\ \boldsymbol{W}_x^i=\begin{cases}V_x^i, & \boldsymbol{x}_t\in V\\ K_x^i, & \boldsymbol{x}_t\in K\end{cases}\\ \boldsymbol{b}^i=\begin{cases}b_V^i, & \boldsymbol{x}_t\in V\\ b_K^i, & \boldsymbol{x}_t\in K\end{cases}\end{cases}\tag{2}$$

其中:$\boldsymbol{W}_h^i$、$\boldsymbol{W}_x^i$、$\boldsymbol{b}^i$ 是第 i 个 RNN 单元的参数;$\boldsymbol{h}_t^i$ 是第 i 个 RNN 单元在时间步骤 t 的隐藏状态。

得到隐藏状态之后，预测用户实体的隐藏状态$\boldsymbol{h}_v$与辅助平台用户实体概率：

$$p = \boldsymbol{h}_v^{\mathrm{T}} \boldsymbol{v}_j' \tag{3}$$

最后，通过最小化真实值和预测值间的交叉熵损失来训练模型。

$$L = -\sum_{i=1}^{|V|} p_i \log(q_i) \tag{4}$$

其中：q_i为预测概率分布，p_i为真实的概率分布。

与用户$\boldsymbol{v}_i$相似度概率值大于等于阈值δ的辅助平台用户构成候选用户实体集$C_v = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_n'\}$，进入下一步筛选。

2.2.2 基于网络拓扑结构特征相似度的用户实体筛选

根据网络拓扑结构特征相似度将用户$\boldsymbol{v}_i$与其候选用户实体集合$C_v = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_n'\}$中的用户进行筛选。目标平台图谱节点$\boldsymbol{v}_i$的网络拓扑结构特征集合为$\boldsymbol{T}_i = \{DC_i, CL_i\}$，由度中心性[13]$DC(i)$和聚类系数[14]$CL_i$构成，计算如下：

$$DC_i = d_i / d_{\max} \tag{5}$$

其中：d_i表示节点v_i的度，$d_{\max}$表示图谱中节点的最大度。

$$CL_i = \frac{2n}{d_i(d_i - 1)} \tag{6}$$

使用余弦相似度计算目标平台图谱网络子图节点$\boldsymbol{v}_i$和辅助平台图谱中的候选节点$\boldsymbol{v}_n'$的拓扑结构特征相似度：

$$Sim_T(\boldsymbol{v}_i, \boldsymbol{v}_n') = \cos(\boldsymbol{T}_i, \boldsymbol{T}_n') = \frac{\boldsymbol{T}_i \cdot \boldsymbol{T}_n'}{\|\boldsymbol{T}_i\| \|\boldsymbol{T}_n'\|} \tag{7}$$

其中：$\boldsymbol{T}_n'$是候选用户实体集合C_v中节点为$\boldsymbol{v}_n'$的网络拓扑结构特征集合。

在$C_v = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_n'\}$中，筛选出拓扑结构特征相似度大于等于拓扑结构特征相似度阈值λ的用户，组成新的候选用户集合$\overline{C_v} = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_j'\}(j \leqslant n)$，进入下一步基于用户画像相似度的用户实体筛选。

2.2.3 基于用户画像相似度的用户实体筛选

用户在社交网络中的公开信息可以用于描述一个人的特征。使用用户画像可以快速、有效地描述社交网络中用户的个人信息及特征，如性别、年龄、职业、所在地、毕业院校等客观信息。联系电话和电子邮箱可以标识唯一的用户，因此，如果不同平台用户节点的联系电话或者电子邮箱相同，则认为是同一用户[15]。

首先，使用Word2vec[16]方法把用户实体的画像和属性A_i信息转化为画像特征向量$\boldsymbol{F}$，进而通过用户画像的相似性度量用户节点的相似性。计算目标平台图谱用户与候选用户集合$\overline{C_v} = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_j'\}$中用户的用户画像相似度$Sim_F(\boldsymbol{v}_i, \boldsymbol{v}_j')$：

$$Sim_F(\boldsymbol{v}_i, \boldsymbol{v}_j') = \cos(\boldsymbol{F}, \boldsymbol{F}') = \frac{\boldsymbol{F} \cdot \boldsymbol{F}'}{\|\boldsymbol{F}\| \|\boldsymbol{F}'\|} \tag{8}$$

其中：$\boldsymbol{F}$目标平台图谱节点的用户画像特征向量，$\boldsymbol{F}'$为候选用户集合$\overline{C_v}$中的用户画像特征向量。

然后，在候选用户集合$\overline{C_v} = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \cdots, \boldsymbol{v}_j'\}$中筛选出相似度最高的用户，将其确定为$\boldsymbol{v}_i$的相同用户，记为相同用户节点对$\boldsymbol{v}_i = \boldsymbol{v}_j'$，实现跨平台用户实体链接。

2.3 基于辅助平台图谱的目标平台图谱关系补全

首先，通过广度优先遍历[17]，获取目标平台图谱用户节点$\boldsymbol{v}_i$所有的关系集合$R_i = \{r_{i1}, r_{i2}, \cdots, r_{ij}\}$，其中关系为一阶关系，即记录可到达的下一个邻接用户节点的路径，$r_{ij} = \boldsymbol{v}_i \rightarrow \boldsymbol{v}_j$。同理，在辅助平台图谱中获取相似子图中用户节点$\boldsymbol{v}_i$的所有路径集合$R_i'$。然后，根据上一节识别出的所有相同用户节点队，将辅助平台图谱路径中的相同用户替换为目标平台图谱的用户，去除相同的关系，即得到用户$\boldsymbol{v}_i$需要补全的关系集合R_i。

$$R_i = R_i' \cup R_i \tag{9}$$

同理，获取子图中所有节点需要补全的关系集合。

当目标平台图谱中缺少辅助平台图谱的相似用户节点时，建立对应的虚拟节点与之替换，当新的节点出现在目标平台图谱时，可以与之比较，建立相应关系。

最后，按照关系集合中的关系，补全目标平台图谱中节点之间的关系。

2.4 算法执行过程

算法1 RCCP-KG算法。

输入 目标平台图谱$G = (V, E)$，辅助平台图谱$G' = (V', E')$，用户博文的时间关键词序列$(\boldsymbol{v}_i, K_n)$；

输出 关系补全后的目标平台图谱$G = (V, E)$。

```
for v_i in G
    执行Louvain算法，得到G = (V, E)的子图集合
      SG = {g_1, g_2, ⋯, g_i}
end for
同理得到辅助平台图谱中的子图集合
  SG' = {g_1', g_2', ⋯, g_j'}
利用Pearson矩阵相似度计算得到匹配子图(g_i, g_j')
for v_i' in g_j'
    使用g_j'与(v_i, K_n)训练EP-MRNN
      for v_i in g_i
          c_1 ← v_i RNN单元c_1处理用户实体
          c_3 ← K_n RNN单元c_3处理用户关键词
          得到候选实体集合C_v = {v_1', v_2', ⋯, v_n'}
      end for
      for v_n' in C_v
          计算 Sim_T(v_i, v_n')
          if Sim_T(v_i, v_n') ⩾ λ
              C_v ← v_n'
          end if
      end for
    Max Sim_F(v_i, v_j') , v_n' ⊂ C_v
end for
get v_i = v_j'
v_j' ← v_i
R_i = R_i' ∪ R_i，更新G = (V, E)
return G = (V, E)
```

3 基于重启随机游走的用户推荐算法

传统的随机游走算法利用网络的拓扑结构，只考虑图的出度和入度，忽略了节点本身的特性。因此本文提出了基于重启随机游走的用户推荐算法UR-RRW，在补全后的目标平台图谱中，融合用户兴趣和用户间关系，采用重启随机游走完成用户推荐。社交网络知识图谱中的部分重启随机游走，如

图5所示。

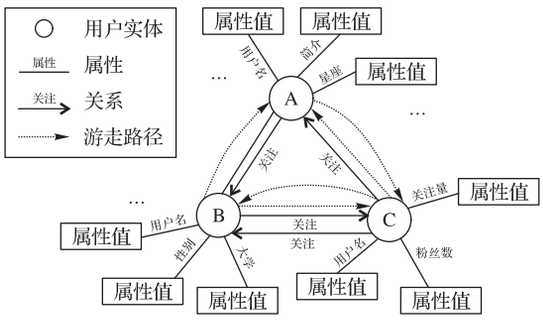


图5　重启随机游走示意图

Fig. 5　Schematic diagram of restart random walk

在补全后的目标平台图谱中,首先,确保用户节点间有较高的紧密连接,使用社区检测Louvain算法再次将用户实体进行社区划分,得到社区集合 $S=\{s_1,s_2,\cdots,s_i\}$,然后在每个社区中建立用户兴趣相似度矩阵,通过随机游走来寻找相似用户,计算用户间游走的平稳概率值,将其降序排列得到用户推荐列表。

3.1　用户兴趣相似度矩阵建立

首先,通过社区检测将用户划分成不同社区。在目标平台图谱,对于新添加的关系,给予与辅助平台图谱中相同的权重值。在原有社区的基础上,再次利用社区检测Louvain算法,重新对社区内的用户节点进行模块度计算[18]。根据模块度变化值,增加或删除社区内的节点,最终确定目标平台图谱中的社区划分,得到新的社区集合 $S=\{s_1,s_2,\cdots,s_i\}$。

然后,定义用户的兴趣相似度矩阵。用户间兴趣相似度需满足以下条件:

1)同一社区的用户间共同关注的用户的类型种类数 l_i 越多,则相似度越高;

2)同一社区的用户间对单一类型用户的关注次数 N_i 越多,则相似度越高。

为满足以上条件,定义矩阵 $\boldsymbol{M}\in\mathbf{R}^{n\times n}$,矩阵元素 m_{ij} 为 $\boldsymbol{v}_i\in V$ 与 $\boldsymbol{v}_j\in V$ 之间的相似度评分,如用户间无共同关注则 $m_{ij}=0$,其形式如式(11)所示:

$$m_{ij}=l_{ij}z_{ij} \tag{10}$$

$$l_{ij}=\frac{l_i\cap l_j}{l_i\cup l_j} \tag{11}$$

$$z_{ij}=\frac{N_{ij}}{N_i+N_j} \tag{12}$$

其中:l_{ij} 为用户 $\boldsymbol{v}_i$ 与用户 $\boldsymbol{v}_j$ 共同关注用户的类型权重,l_i 和 l_j 分别是用户 $\boldsymbol{v}_i$ 与用户 $\boldsymbol{v}_j$ 关注的用户类型种类数;z_{ij} 为用户 $\boldsymbol{v}_i$ 与用户 $\boldsymbol{v}_j$ 共同关注用户的次数权重。

$$\boldsymbol{M}=\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ m_{31} & m_{32} & \cdots & m_{3n} \\ \vdots & \vdots & & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{bmatrix}$$

对 $\boldsymbol{M}$ 进行归一化处理得到用户兴趣相似度矩阵 $\boldsymbol{M}'$。

$$m_{ij}'=\frac{m_{ij}}{\sum_{i=1}^{n}m_{ij}} \tag{13}$$

3.2　基于重启随机游走的用户推荐

本文采用基于用户兴趣相似度的重启随机游走算法,将得到的用户兴趣相似度矩阵 $\boldsymbol{M}$ 作为概率转移矩阵并输入到重启随机游走模型中,通过归一化的用户兴趣相似度矩阵计算得到平稳概率值[4]。

$$\boldsymbol{y}_{\text{Last}}=(1-\theta)(\boldsymbol{I}-\theta\boldsymbol{M})^{-1}\boldsymbol{y}_0 \tag{14}$$

其中:$\boldsymbol{y}_{\text{Last}}$ 为经过 n 步游走后所得到的平稳概率值;θ 为下一步游走到其最近邻节点的概率,θ 为可调参数,通过实验验证 θ 取0.1;$\boldsymbol{y}_0$ 为用户的列向量,是向量 $\boldsymbol{y}$ 的初始向量,向量 $\boldsymbol{y}$ 中的每一项值代表着从用户 $\boldsymbol{v}_i$ 经过 n 步游走后到达各个节点的概率值。

最后将平稳概率值 $\boldsymbol{y}_{\text{Last}}$ 进行降序排序,形成用户推荐列表 $RL=\{u_1,u_2,\cdots,u_j\}$,将用户列表中未关注的用户进行推荐。

3.3　算法执行过程

算法2　UR-RRW。

输入　目标平台图谱 $G=(V,E)$;

输出　用户推荐列表 $RL=\{u_1,u_2,\cdots,u_j\}$。

1)　在补全后的目标平台图谱执行Louvain算法
2)　get $S=\{s_1,s_2,\cdots,s_i\}$
3)　计算用户兴趣相似度矩阵 $\boldsymbol{M}'$
5)　for $u_i\in V$
6)　　构建用户列向量 $\boldsymbol{y}_0$
7)　　计算 $(1-\theta)(\boldsymbol{I}-\theta\boldsymbol{M})^{-1}\boldsymbol{y}_0$
8)　end for
9)　降序排列 $\boldsymbol{y}_{\text{Last}}$
10)　return $RL=\{u_1,u_2,\cdots,u_j\}$

4　实验与分析

本实验选取经典的协同过滤(CF)算法[3]、基于跨平台的在线社交网络用户推荐算法(User Recommendation algorithm based on Cross-Platform online social network, URCP)[7],以及使用重启随机游走的基于多开发者社区的用户推荐算法(User Recommendation algorithm based on Multi-developer Community, UR-MC)[8]作为对比方法与本文提出的URCP-KR方法相比较。当改变用户数量、用户平均关联关系数量以及用户平均博文数量时,本文选取精确率 pre(precision)、召回率 rec(recall)、F1值 F_1、覆盖率 cov(coverage)作为评价指标,具体指标介绍及公式如下:

精确率 pre 为正确推荐用户数 TP 与推荐列表用户总数 $|RL|$ 的比值:

$$pre=TP/|RL| \tag{15}$$

召回率 rec 为正确推荐用户数 TP 与目标平台中的好友总数 FN 的比值:

$$rec=TP/FN \tag{16}$$

F1值均匀地反映了推荐效果:

$$F_1=2\times pre\times rec/(pre+rec) \tag{17}$$

覆盖率 cov 为推荐用户总量 $|RL|$ 与总用户数量 $|V|$ 的比值,反映推荐的多样性:

$$cov=|RL|/|V| \tag{18}$$

4.1　实验环境

本实验在具有2.8 GHz Inter Core i7处理器和16 GB内存的机器上运行,操作系统为Windows 10,编程语言为Python。

4.2　实验数据

本实验选取微博作为目标平台,选取知乎作为辅助平台图谱。从微博和知乎采集了用户信息及其发布内容,把从已知的实名认证的用户作为起始节点爬取用户相关信息。通过深度优先搜索的方法采集节点的邻居信息,根据小世界理

论[19]，本文将深度优先搜索的深度设置为4。数据集包含用户的21万条微博数据，19万条知乎数据，共36万对用户间关注关系数据。

4.3 参数设置

本文中用户相似度概率阈值β、拓扑结构特征相似度阈值λ和重启随机游走可调参数θ通过在测试集数据上实验，根据用户推荐F1值的变化确定。由图6可知，当$\beta = 0.4$，$\lambda = 0.6$，$\theta = 0.1$时，用户推荐的F1值最优。

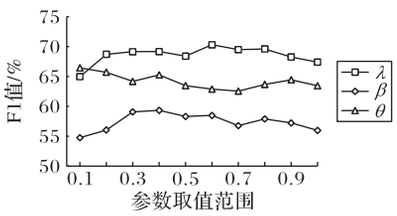


图6 F1值随λ、β和θ的变化结果

Fig. 6 F1 changing with λ, β and θ

本文选取用户数$|V|$、用户平均关联关系数rn和用户平均博文数量pb作为实验参数，如表1所示。

表1 实验参数表

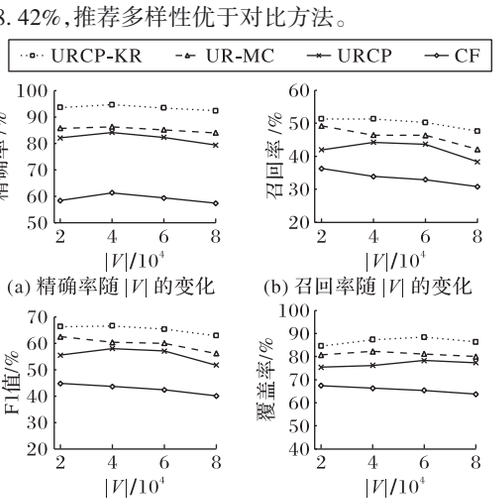
Tab. 1 Experimental parameter table

参数	取值
用户数$\|V\|$	20 000，40 000，60 000，80 000
用户平均关联关系数rn	10，20，40，60
用户平均博文数pb	5，20，80，320

4.4 实验结果分析

将数据分为训练集和测试集，其中训练集用于训练模型，测试集用于衡量模型的性能。使用十折交叉验证，训练集采用总数据集的90%数据量，测试集使用10%。

由图7可知，在用户数$|V|$较小时，本文提出的URCP-KR方法在精确率、召回率、F1值和覆盖率都要优于对比方法；随着用户数量的不断增加，由于用户数量增长幅度比用户推荐数量增加幅度大，因此四种方法精确率、召回率和F1值都有所下降，但是URCP-KR方法下降趋势缓慢；在覆盖率方面，URCP-KR方法先增后减，当用户数量为60 000时，覆盖率最高为88.42%，推荐多样性优于对比方法。



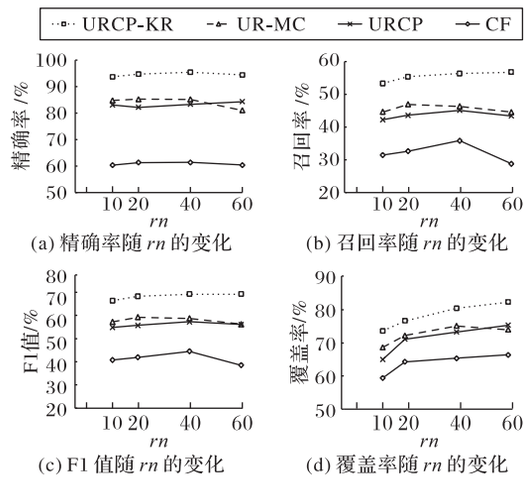
(a) 精确率随$|V|$的变化 (b) 召回率随$|V|$的变化

(c) F1值随$|V|$的变化 (d) 覆盖率随$|V|$的变化

图7 用户数$|V|$变化时各方法的度量指标对比

Fig. 7 Comparison of metrics of each method when changing user numbers $|V|$

由图8可知，随着用户平均关联关系数rn不断增加，URCP-KR方法的精确率先增加后减小，当用户平均关联关系数量为40时精确率达到最大，为95.31%；覆盖率方面逐步增加。因为URCP-KR方法将不同平台的用户关系进行融合，丰富了目标平台的用户关系。CF算法和UR-MC算法则随用户平均关联关系数量增加到一定程度而受到用户评分数量和用户标签数量的限制。



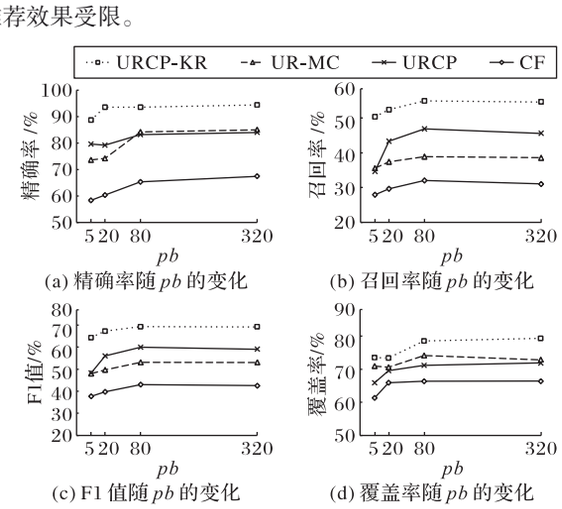
(a) 精确率随rn的变化 (b) 召回率随rn的变化

(c) F1值随rn的变化 (d) 覆盖率随rn的变化

图8 rn变化时各方法的度量指标对比

Fig. 8 Comparison of metrics of each method when changing average number of user relationships rn

由图9可知，随着用户平均博文数pb增加，URCP-KR方法的精确率和覆盖率逐步增加并趋于平稳，因为用户博文数据越多，实体链接效率越高。同样，URCP算法获取到的用户兴趣爱好主题数量增加，推荐精确率和覆盖率有较大的提高；但由于URCP算法通过API获取相同用户数量的有限性，因此推荐效果受限。



(a) 精确率随pb的变化 (b) 召回率随pb的变化

(c) F1值随pb的变化 (d) 覆盖率随pb的变化

图9 pb变化时各方法的度量指标对比

Fig. 9 Comparison of metrics of each method when changing average number of user's blogs pb

5 结语

在跨平台推荐系统环境下，本文提出了基于知识图谱和重启随机游走的跨平台用户推荐方法。首先，提出了基于知识图谱的跨平台用户关系补全算法RCCP-KG，利用改进的多层RNN预测出候选用户实体，筛选出不同平台中可能相同的

用户;然后,通过关系补全将辅助平台图谱中的用户信息补全到目标平台,完善用户之间的关系;最后,提出了基于重启随机游走的用户推荐算法 UR-RRW,利用重启随机游走算法计算同一社区中用户的相似性,形成用户推荐列表,从而解决跨平台用户推荐问题,提高用户推荐的准确率及多样性。未来将对实体及关系消歧进行研究,增强跨平台间相同用户的匹配,进一步提高跨平台用户间的推荐准确性和多样性。

参考文献 (References)

[1] ZHONG E, FAN W, WANG J, et al. ComSoc: adaptive transfer of user behaviors over composite social network[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 696-704.

[2] 陈克寒,韩盼盼,吴健. 基于用户聚类的异构社交网络推荐算法[J]. 计算机学报, 2013, 36(2): 349-359. (CHEN K H, HAN P P, WU J. User clustering based social network recommendation [J]. Chinese Journal of Computers, 2013, 36(2): 349-359.)

[3] DENG Z, HE B, YU C, et al. Personalized friend recommendation in social network based on clustering method[C]// Proceedings of the 2012 International Symposium on Intelligence Computation and Applications, CCIS 316. Berlin: Springer, 2012:84-91.

[4] 马铁民,周福才,王爽. 基于用户相似度的随机游走社交网络事件推荐算法[J]. 东北大学学报(自然科学版), 2019, 40(11): 1533-1538. (MA T M, ZHOU F C, WANG S. Social network event recommendation algorithms based on user similarity random walk [J]. Journal of Northeastern University (Natural Science), 2019, 40(11): 1533-1538.)

[5] 刘晓飞,朱斐,伏玉琛,等. 基于用户偏好特征挖掘的个性化推荐算法[J]. 计算机科学, 2020, 47(4): 50-53. (LIU X F, ZHU F, FU Y C, et al. Personalized recommendation algorithm based on user preference feature mining[J]. Computer Science, 2020, 47 (4): 50-53.)

[6] DENG Z, SANG J, XU C. Personalized video recommendation based on cross-platform user modeling[C]// Proceedings of the 2013 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2013: 1-6.

[7] 彭舰,王屯屯,陈瑜,等. 基于跨平台的在线社交网络用户推荐研究[J]. 通信学报, 2018, 39(3):147-158. (PENG J, WANG T T, CHEN Y, et al. User recommendation based on cross-platform online social networks[J]. Journal on Communications, 2018, 39 (3):147-158.)

[8] 时宇岑,印莹,赵宇海,等. 基于多开发者社区的用户推荐算法[J]. 软件学报, 2019, 30(5): 1561-1574. (SHI Y C, YIN Y, ZHAO Y H, et al. User recommendation algorithm based on multi-developer community[J]. Journal of Software, 2019, 30(5): 1561-1574.)

[9] 文凯,朱传亮,何少元. 结合用户社区和评分矩阵联合社区的推荐算法研究[J]. 小型微型计算机系统, 2019, 40(10): 2119-2124. (WEN K, ZHU C L, HE S Y. Research on recommendation algorithm combining user community and score matrix joint community[J]. Journal of Chinese Computer Systems, 2019, 40 (10): 2119-2124.)

[10] DUGUÉ N, PEREZ A. Directed Louvain: maximizing modularity in directed networks: hal-01231784[R]. Orléans: Université d'Orléans, 2015:1-14

[11] ILIEVSKI F, VOSSEN P, VAN ERP M. Hunger for contextual knowledge and a road map to intelligent entity linking[C]// Proceedings of the 2017 International Conference on Language, Data and Knowledge, LNCS 10318. Cham: Springer, 2017: 143-149.

[12] GUO L, ZHANG Q, GE W, et al. DSKG: a deep sequential model for knowledge graph completion[C]// Proceedings of the 2018 China Conference on Knowledge Graph and Semantic Computing, CCIS 957. Singapore: Springer, 2018:65-77.

[13] ŽALIK K R. Evolution algorithm for community detection in social networks using node centrality [M]// BEMBENIK R, SKONIECZNY Ł, PROTAZIUK G, et al. Intelligent Methods and Big Data in Industrial Applications. Cham: Springer, 2019: 73-87.

[14] ALEMI M, HAGHIGHI H, SHAHRIVARI S. CCFinder: using Spark to find clustering coefficient in big graphs[J]. The Journal of Supercomputing, 2017, 73(11): 4683-4710.

[15] 马江涛. 基于社交网络的知识图谱构建技术研究[D]. 郑州: 战略支援部队信息工程大学, 2018:30-38. (MA J T. Research on knowledge graph construction technology based on social network [D]. Zhengzhou: Information Engineering University, 2018:30-38.)

[16] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2020-11-12]. https://arxiv. org/pdf/1301. 3781. pdf.

[17] HENZINGER M, KRINNINGER S, NANONGKAI D. Sublinear-time maintenance of breadth-first spanning trees in partially dynamic networks[J]. ACM Transactions on Algorithms, 2017, 13(4): No. 51.

[18] MINERVINI P, D'AMATO C, FANIZZI N, et al. Leveraging the schema in latent factor models for knowledge graph completion [C]// Proceedings of the 31st Annual ACM Symposium on Applied Computing. New York: ACM, 2016: 327-332.

[19] MAIER B F, HUEPE C, BROCKMANN D. Modular hierarchical and power-law small-world networks bear structural optima for minimal first passage times and cover time[J]. Journal of Complex Networks, 2019, 7(6):865-895.

This work is partially supported by the National Key Research and Development Program of China (2018YFB1003801), the Surface Program of National Natural Science Foundation of China (61977021), the Technology Innovation Special Program of Hubei Province (Major Program) (2018ACA13).

YU Dunhui, born in 1974, Ph. D., professor. His research interests include knowledge graph, big data, services computing, crowdsourced data management.

ZHANG Luyi, born in 2000. Her research interests include knowledge graph.

ZHANG Xiaoxiao, born in 1995, M. S. candidate. Her research interests include knowledge graph, crowdsourced data management.

MAO Liang, born in 1998, M. S. candidate. His research interests include knowledge graph.