基于跨平台的在线社交网络用户推荐研究

彭舰[1]，王屯屯[1]，陈瑜[1]，刘唐[2]，徐文政[1]
（1. 四川大学计算机学院，四川 成都 610065；2. 四川师范大学基础教学学院，四川 成都 610068）

摘 要：在社交网络用户推荐研究领域，通过提取用户的行为模式对其进行好友推荐。但是用户的行为是多样性的，在不同的社交平台，用户可能有不同的行为模型。提出跨平台用户推荐模型，同时对用户相关的所有社交网络平台进行建模，最后将用户在所有平台的行为模式进行融合。基于真实的新浪微博数据集和知乎数据集，通过一系列对比实验证明，跨平台用户推荐模型可以更加全面准确地刻画用户行为，更好地进行用户推荐。

关键词：跨平台；用户推荐；在线社交网络；数据挖掘
中图分类号：TP311
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018044

User recommendation based on cross-platform online social networks

PENG Jian[1], WANG Tuntun[1], CHEN Yu[1], LIU Tang[2], XU Wenzheng[1]
1. Computer Science School, Sichuan University, Chengdu 610065, China
2. College of Fundamental Education, Sichuan Normal University, Chengdu 610068, China

Abstract: In the field of online social networks on user recommendation, researchers extract users' behaviors as much as possible to model the users. However, users may have different likes and dislikes in different social networks. To tackle this problem, a cross-platform user recommendation model was proposed, users would be modeled all-sided. In this study, the Sina micro blog and the Zhihu were investigated in the proposed model, the experimental results show that the proposed model is competitive. Based on the proposed model and the experimental results, it can be known that modeling users in cross-platform online social networks can describe the user more comprehensively and leads to a better recommendation.

Key words: cross-platform, user recommendation, online social networks, data mining

1 引言

随着互联网应用的发展，在线社交网络已吸引和聚集了大量用户，如 Facebook、Twitter、新浪微博和知乎等知名社交网络。在这些社交网络平台上，每天的活跃用户数量和产生的社交信息量巨大。仅仅在新浪微博平台，每天的活跃用户可达 6 000 万个，平均产生数亿条微博信息[1]，新浪微博已成为国内用户数量众多的重要社交网络平台。如何对数量众多的新浪微博用户进行有效的个性化推荐已成为社交网络研究中一个非常重要的课题。

针对新浪微博的用户推荐问题，目前已有较多的研究[2,3]。研究者主要通过提取用户在新浪微博平台的行为特征对其进行建模，但是用户的行为是多样性的，在不同的社交平台会有不同的喜好。单独地利用一个平台的信息为用户进行建模，可能不足以全面反映用户的兴趣爱好。如果能同时参照和结合用户在其他社交网络平台中的信息，可以更加全面准确地了解用户的兴趣爱好以及行为特征。因此，在文档中进行用户个性化推荐研究的同时，融

收稿日期：2017-06-26；**修回日期**：2018-01-10
基金项目：国家自然科学基金资助项目（No.U1333113, No.61602330）；四川省科技支撑计划基金资助项目（No.2014GZ0111）；四川省教育厅科研基金资助项目（No.18ZA0404）
Foundation Items: The National Natural Science Foundation of China (No. U1333113, No. 61602330), Science and Technology Support Plan Foundation of Sichuan Province (No. 2014GZ0111), The Scientific Research Fund of Sichuan Provincial Education Department (No.18ZA0404)

合其他平台的信息来进行跨平台的用户推荐，从而提升推荐的效率和准确度。例如，用户张三在新浪微博平台（目标平台）注册账号，并且经常发布或转发足球相关内容，传统的单平台用户推荐模型只是根据用户在新浪微博的行为特征对其进行建模，得出的结论是他的兴趣爱好主要是足球，于是为他推荐足球相关的用户。此外，张三在知乎平台（辅助平台）经常讨论软件开发等相关问题，而且活跃度比在新浪微博高很多。如果能同时结合张三在 2 个平台上的行为，得到的结论是张三喜欢足球，但是更喜欢软件开发。为张三推荐好友时，不仅要推荐足球相关的用户，也要推荐软件开发相关的用户，而且软件开发相关的用户在推荐列表中的位置更加靠前。可能由于张三在微博平台未接触到软件开发相关的用户，导致其在微博平台没有表现出在软件开发方面相关的兴趣爱好和用户行为。

综上所述，如果只是单独地依据用户在微博平台的用户行为特征为用户进行建模，很可能会导致片面地理解用户兴趣爱好。本文所提跨平台推荐模型 URCP，不仅考虑到用户在新浪微博（目标平台）的兴趣爱好，更结合了其在知乎平台（辅助平台）上的行为特征，最终将该用户在所有相关平台的兴趣爱好结合起来进行微博好友推荐。

在进行跨平台用户推荐时，首先遇到的问题是数据采集。每个单独的平台可以通过爬虫技术或调用 API 获取相应数据，但是如何将同一个用户在不同平台的账号信息对应起来是很困难的；其次，每个社交平台具有一定的差异性。例如，知乎平台没有转发功能。如何采用统一的方法为用户进行建模也是很大的挑战；最后一个问题是用户冷启动问题，即无法获取新注册用户的行为特征。

本文的主要贡献如下。

1) 提出了一种基于跨平台的在线社交网络用户推荐模型 URCP，通过融合用户在多个社交网络平台的信息，可以更加全面地刻画用户行为，更加准确地进行好友推荐。传统的推荐算法只能对用户在目标平台的行为进行建模，不足以反映用户全部的兴趣爱好。本文所提跨平台推荐模型，不仅考虑到用户在目标平台的兴趣爱好，同时将用户在其他辅助平台的行为特征融入整个模型中，对用户进行更加全面的建模，进而可以更加准确地对用户进行好友推荐。

2) 提出了一种新的跨社交网络用户采集方法，为跨平台研究提供了数据支持，而且具有很好的扩展性。国外提出的跨平台数据采集方法主要利用账号关联工具或 Google+API。但是在国内，大部分社交平台都没有账号关联工具，而且也没有与 Google 账号进行关联，所以需要使用一个适合国内社交网络发展情况的跨平台数据采集方法。

3) 基于跨平台的用户推荐，对于新注册用户，将辅助平台的数据迁移到目标平台，可以很好地解决冷启动问题。传统的解决用户冷启动的方法是采用基于批判式会话的方式来逐渐引导用户，但是基于会话的方式会使推荐周期较长。本文所提冷启动解决方案主要借助于用户在其他平台的行为特征进行兴趣爱好的迁移。

2 相关工作

社交网络用户推荐已成为一个研究热点。由于现有的推荐系统大都建立在单一社交平台的基础上，利用用户的链接信息（例如，关注其他人与被其他人关注）或内容信息（用户个人简历和用户已经发布的内容）来进行推荐，导致其推荐效果仍有较大的提升空间。此外，在单一的平台上进行推荐容易导致数据过于稀疏和数据的过度拟合[4]。

协同过滤（CF, collaborative filtering）是使用较广泛的技术，根据用户产生的评分信息来预测用户的偏好[5]。然而，实践中，CF 系统容易受到不公平评分的影响。文献[6]提出，协同过滤的方法不适用于用户推荐，并且当考虑到人与人之间的关系时，需要考虑的因素比较多。在图像挖掘领域，人们经常把社交网络用户推荐问题当作图形中的链路预测问题来解决。文献[7]将该问题定义为给定一个社交网络某个时间点的图形快照，目的就是找到在未来某个时间点之前，图形将会增加的边。但这种方法并不能够很好地反映真实生活中人们进行朋友选择的用户偏好[8]。

潜在狄利克雷分配（LDA, latent Dirichlet allocation）模型的提出使越来越多的人使用 LDA 进行语义分析和用户推荐。文献[9]利用 LDA 模型对用户进行建模，提出 top-k 推荐算法，向用户推荐 k 个关注用户以及用户可能感兴趣的文章。文献[1]针对用户的兴趣总是在发生变化这一现象，利用 LDA 对用户的内容进行主题生成来挖掘出用户潜在的兴趣。

文献[4]指出，利用其他平台的数据可以缓解数据稀疏问题，并且提高用户模型的预测性能。文献[10]提出以标签为基础的用户简历，并且提出了一系列跨系统用户建模的方法。文献[11]利用 Twitter 平台的数

据生成最新的话题，并向 YouTube 平台的用户推荐相关的视频。文献[12]提出利用源平台的数据丰富目标平台的数据来进行视频的推荐，以此解决目标平台的数据稀疏问题和冷启动问题。文献[13]利用在线 LDA（OSLDA, online streaming latent Dirichlet allocation）模型实时地生成主题向量，并通过迁移学习算法来实现多媒体的应用。文献[14]指出，源平台向目标平台的数据转移主要是通过迁移学习完成的，而转移学习主要依赖于对齐用户或对齐数据。

本文提出基于对齐用户的跨平台方法对用户进行推荐。在进行跨平台推荐系统的研究中，如何有效获取不同平台的数据是一个重要的环节。文献[10~12]利用 Friendfeed、About.me 和 Google+API 等工具进行跨平台的数据采集。但是，由于很多社交网络用户没有使用账号管理工具或 Google+，导致他们提出的数据采集方法无法对某些用户或社交平台进行数据采集。文献[15~18]提出了各种专门模型用于找到多个平台中对应多个账号的同一个用户。这些方法主要是利用自然语言处理技术或时空关系来进行判断。例如，在 2 个不同的平台（A 和 B）上，发现平台 A 上某个用户的名称与平台 B 上某个用户的名称很相似，就可以在一定程度上认为这是同一个用户。很有可能是 2 个不同的用户，由于看了同一部电影就起了相似的名称。因此，现实情况的复杂性导致这些方法的效果不是很理想。

在社交网络用户推荐中，除了数据稀疏问题外，另外一个比较重要的问题是用户冷启动问题。传统的解决方法是，不断与用户进行会话，在每一次会话中，用户对推荐对象的一个特征进行批判，然后根据用户的批判特征逐渐引导用户找到其期望的推荐对象[19]，这种方法最大的问题在于会话周期太长。虽然有很多研究者提出一些改进的方法来减少用户会话次数，但是效率还是比较低[20~22]。本文所提基于跨平台的解决方法，不用通过用户会话解决用户冷启动问题，以快速方便地进行用户推荐。

3　数据处理及用户行为建模

在线社交网络中，用户产生的行为数据极其庞大。不同的社交网络平台之间的差异也较大。因此，在研究跨平台的社交网络用户推荐时，如何处理好单个用户在多个平台的行为数据成为一个关键问题。本节将重点介绍如何获取用户在多个平台的行为数据，并对其进行数据清洗，最后利用这些数据对用户进行建模。本文所提模型总体框架如图 1 所示。

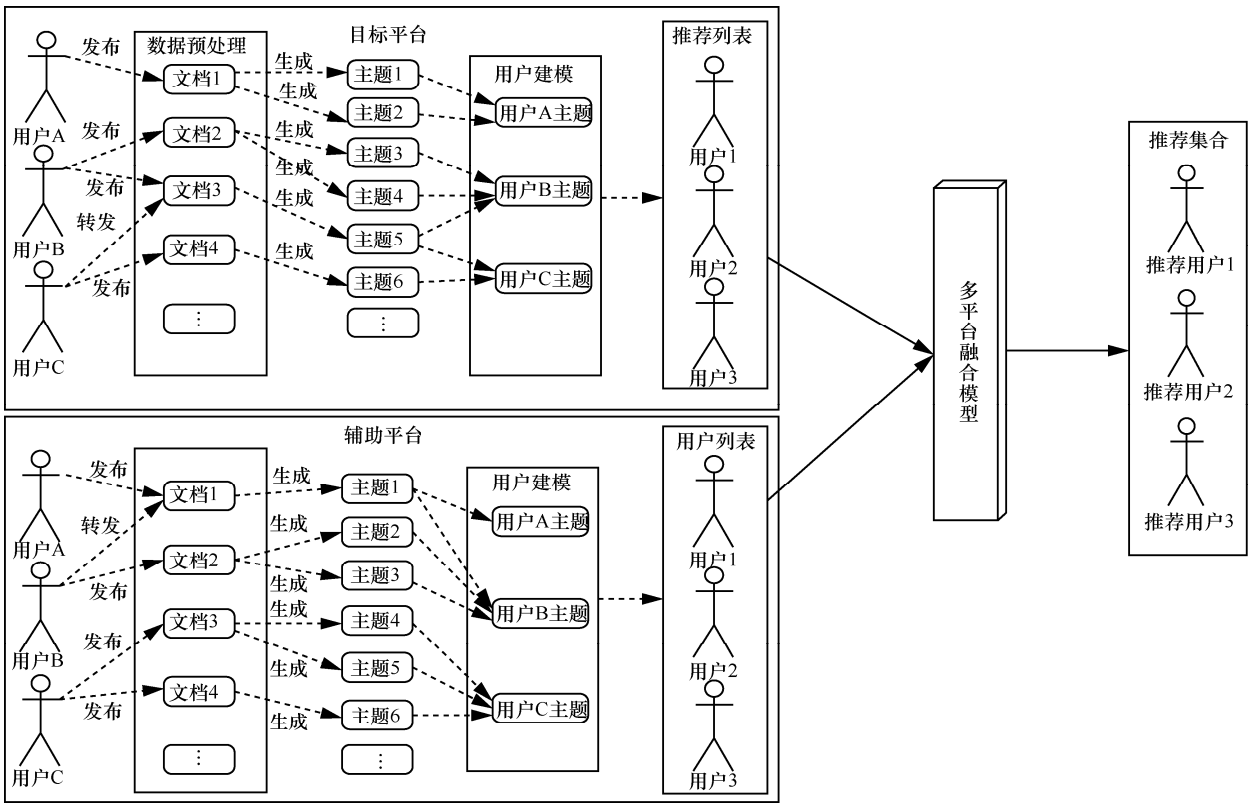


图 1　模型总体框架

在图 1 中，上部分为目标平台，下部分为辅助平台的其中一个。在目标平台中，用户 A 发布了“文档 1”，“文档 1”生成了“主题 1”和“主题 2”，这 2 个主题构成了用户 A 的主题。用户 B 发布了“文档 2”和“文档 3”，其中，“文档 2”生成了“主题 3”和“主题 4”，“文档 3”生成了“主题 5”，最终“主题 3”和“主题 4”以及“主题 5”构成了用户 B 的主题。用户 C 转发了“文档 3”，发布了“文档 4”，其中，“文档 4”生成了“主题 6”，最终“主题 5”和“主题 6”构成了用户 C 的主题。通过各个用户的主题向量，为目标用户推荐“用户 1”“用户 2”“用户 3”等。辅助平台与目标平台结构类似。最终将所有平台的推荐列表融合起来，为目标用户生成最终的推荐列表。

每个用户在不同的平台，可以发布自己的原创文章或转发其他人的文章。通过提取用户的文档信息，可以挖掘出其潜在的兴趣爱好。在每个平台上，都会生成用户的主题空间，利用用户的主题分布，向目标用户推荐与其兴趣爱好最接近的用户。在不同的平台，用户的主题分布不同，导致同一个用户在不同的平台产生的推荐列表不一致。通过一个融合模型，对所有的推荐用户列表进行重排序，从而得到目标用户最终的推荐列表。下面将详细介绍框架各个部分。

3.1 数据采集以及用户冷启动

在跨平台研究中，目前主流的数据采集方法是利用账号工具和调用谷歌提供的 API。但是，由于一些社交网络平台的各种限制，这些方法难以有效获取到用户的完整数据。例如，新浪微博并没有提供账号管理工具与其他平台进行关联，也未对外提供 API 帮助开发者获取新浪微博用户关联的其他平台账号。为解决此问题，本文提出了一种新的跨平台数据采集方法。首先，通过网络爬取获取账号匹配关系；然后，通过各个平台对外提供的 API 进行数据获取。例如，新浪微博提供了跨平台登录功能，在登录某些社交平台时，注册用户可以使用微博账号进行登录，如知乎、豆瓣、人人网和优酷等社交平台均允许采用微博账号登录。因此，基于该功能可获得微博用户在其他平台的数据信息。本文为改善新浪微博用户的推荐效果，选用知乎作为新浪微博的辅助平台。由于知乎是一个话题性讨论及典型的问答社区，在该平台上更容易分析出用户感兴趣的话题和用户的喜好。

如果一个用户同时拥有新浪微博账号以及知乎账号，并且该用户曾经利用新浪微博账号登录知乎，知乎平台会在该用户首页进行特殊标识，表明该用户的知乎账号已经与其新浪微博账号进行关联。为获取用户的新浪微博和知乎的账号对应关系，本文需要对知乎的用户首页进行网络爬取。通过解析网页，判断是否已关联新浪微博平台。如果用户已经关联，通过数据解析即可获取用户在新浪微博和知乎上的账号对应关系。此外，由于新浪微博提供了公开的 API，可以使开发者容易获取到新浪微博用户的大部分信息。因此，利用新浪微博平台和知乎平台对外提供的这些 API，通过传入不同的用户 ID 即可便捷地获取到用户对应的各种信息，从而有效解决了难以获取不同平台信息的问题。具体实现过程为大量爬取知乎用户的个人首页，保存网页内容到本地。离线对网页内容进行解析，发现知乎名为“张亮”的用户，关联了新浪微博平台，并给出了其在新浪微博平台的个人首页地址。对该网址进行解析，可以得到该用户新浪微博和知乎平台的账号对应关系为{izlmichael，张亮}。最后将用户名作为参数传入对应平台的 API，即可获取该用户在不同平台的信息。

在用户冷启动问题上，本文提出了较为简单和高效的解决方案。如果用户需要在目标平台上进行推荐，但是该用户对于目标平台是新注册用户，传统的推荐算法无法对其进行推荐。基于会话的冷启动解决方案需要与用户进行交互，而且比较耗时。本文所提跨平台解决方案，可以将该用户在其他平台上的信息复制到目标平台，这样就可以较为快速地解决用户冷启动问题。例如，当需要对新浪微博上的一个用户进行推荐时，发现该用户是新注册用户，无法确定其兴趣爱好。但是该用户的新浪微博账号已经关联了知乎账号，而且该用户在知乎上发布过很多文章。可以将该用户在知乎上的兴趣爱好迁移到新浪微博平台，这样就可以获取其兴趣爱好，并利用推荐算法对其进行推荐。如果用户在知乎上的兴趣爱好与其在微博上的一致，可以解决用户冷启动问题；如果用户在 2 个平台的兴趣爱好不一致，兴趣爱好的迁移，可以发现用户在新浪微博平台无法表现出的兴趣爱好。本文认为，用户的兴趣爱好由该用户在所有平台的兴趣爱好共同组合而成，任意单个平台的兴趣爱好都不足以完全表示用户的兴趣爱好。

3.2　数据预处理

在线社交网络平台的结构较为复杂，如何选取合适的用户特征进行建模是用户推荐系统中的另一个重要环节。考虑到用户的兴趣会随着时间发生变化，如果能充分利用用户最近的行为即可有效解决该问题。在本文模型中，利用用户在各个平台最近发布的文章或内容来分析该用户最新的兴趣。同时，本文通过对新浪微博用户观察分析，发现水军有个普遍存在的特征：关注了很多其他用户，但是很少有其他用户关注自己。因此，本文通过某个用户的关注数量和被关注数量的比值来识别和剔除社交网络中水军的相关数据。在实验分析中，发现若某用户的关注人数远远大于其粉丝数量，则该用户是水军的概率也较大。文献[23]指出，可以通过一个用户的所有关注用户和被关注用户之间的关注关系间的距离来判断该用户是否为水军。本文利用该方法对实验对象进行逐一判断，并将水军直接剔除。由于微博平台具有访问次数限制，无法直接通过接口验证用户间的关注关系。为解决此问题，本文爬取各个用户的关注列表，并通过分析关注列表来分析用户之间的关注关系。例如，若用户 A 在用户 B 的关注列表中，则认为 B 关注了 A，即 B 是 A 的粉丝。

兴趣爱好建模是针对每个用户的。本文采用最常用的概率主题模型 LDA 模型来对微博用户进行兴趣度建模。利用 LDA 进行用户建模时，需要对文档进行分词和过滤等数据预处理操作。本文采用的分词算法来源于网上的开源代码（jieba）。关于单词的过滤，比较流行的做法是按照词性进行过滤，例如，只保留动词和名词。但是这样会导致用户信息的进一步缺失。为了尽可能地保留用户信息，本文按照单词的出现次数进行过滤。

3.3　单平台用户行为建模

在对微博数据进行分析时，发现用户 A 关注用户 B 的主要动机是由于用户 A 和用户 B 具有相同的兴趣爱好和共同关注的话题。基于此，本文通过共同的话题关系来对用户进行建模，并采用 LDA 主题模型进行用户的话题建模。LDA 主题模型是话题建模中较流行的方法，可用于识别大规模文档集中潜藏的主题信息。然而，由于微博内容具有典型的短文本特性，直接使用经典的 LDA 模型用于微博用户建模的效果并不理想。因此，本文提出一种改进的基于 LDA 的微博用户模型。在改进模型中，首先将所有用户的微博基于用户进行划分，并将每个用户发布的微博进行合并，以此来作为该用户的信息来源。同时，将标准 LDA 模型的“文档—主题—词”3 层结构转变为“用户—主题—词”的用户模型，并利用该模型进行用户建模。在使用 LDA 模型对用户进行建模时，模型的好坏与以下 3 个参数联系非常紧密：伸张系数α和β以及主题数量 *topic_number*。

传统的用户推荐中，一般通过评分矩阵获得用户向量。但是这种方法容易受到不公平评分的影响。此外，实践中，用户的打分也不能完全反映出该用户的真实兴趣爱好。而 LDA 模型通过用户进行概率主题分析，可以获取其潜在的兴趣爱好。为便于理解本文模型，表 1 给出了文章中出现的符号及其描述。

表 1　　**符号及其描述**

符号	描述
$\boldsymbol{D}$	所有用户的全部文档集合
$\boldsymbol{T}$	所有文档的主题集合
d	一个文档
t	单个文档的主题单词
N	单个文档的主题数量
$\boldsymbol{P}_{\text{doc\_topic}}$	单个文档的概率主题分布
Sim	2 个用户间的相似度
e	社交网络平台的数量
R	单个平台的主题数量
$\boldsymbol{VOC}$	所有文档组成的单词集合
$\boldsymbol{\theta}$	所有平台的权值向量
$H(x)$	预测函数
$J(x)$	损失函数
$Precision$	模型的准确率
$Recall$	模型的召回率
F_1	模型的 F_1 值
MAP	模型的平均准确率
D	2 个用户间的距离
cos	2 个向量间的相似性
L	学习步长

定义 1　文档集合 $\boldsymbol{D}$。令 $\boldsymbol{D}=(d_1,d_2,\cdots,d_i,\cdots,d_u)$，其中，$d_i$ 表示第 i 个用户最近发布的 k 篇文章，$\boldsymbol{d}_i=(d_{i1},d_{i2},\cdots,d_{ij},\cdots,d_{ik})$，则 $\boldsymbol{D}$ 是文档的集合。

定义 2　文档的主题（topic）。令 $\boldsymbol{T}=(t_1,t_2,\cdots,t_i,\cdots,t_z)$，其中，$t_i$ 表示生成的第 i 个主题，则 $\boldsymbol{T}$ 是文档主题。利用 LDA 模型对用户文档进行概率主题建模，得到其主题 $\boldsymbol{T}$。

定义 3 文档中各主题的概率。令 $\boldsymbol{P}_{\text{doc\_topic}}=(P_{t1}, P_{t2},\cdots, P_{ti},\cdots,P_{tk})$，其中，$P_{ti}$ 表示文档 d 对应主题集合 $\boldsymbol{T}$ 中第 i 个 topic 的概率，且 $P_{ti}=\frac{n_{ti}}{N_d}$，$n_{ti}$ 表示文档 d 中对应第 i 个 topic 的词的数目，N_d 表示文档 d 中所有单词的数量，则 $\boldsymbol{P}_{\text{doc\_topic}}$ 指文档中各主题的概率。

定义 4 每个主题由各个单词生成的概率组成。令 $\boldsymbol{P}_{\text{topic\_word}}=(P_{w1},P_{w2},\cdots,P_{wi},\cdots,P_{wm})$，其中，$P_{wi}$ 表示 t 生成单词集合中第 i 个单词的概率，且 $P_{wi}=\frac{n_{wi}}{N_t}$。$n_{wi}$ 表示 t 对应到 VOC 中第 i 个单词的数目，N_t 表示所有对应到 t 的单词总数。

为获得每个用户的主题分布，本文中通过 3 个步骤来实现：1) 从一个用户的微博中抽取一个主题；2) 从抽取到的主题中抽取一个单词；3) 重复前面 2 个步骤，查找出微博中的所有单词。将上述主题中单词出现的次数转换为向量空间模型。

得到每个用户的主题向量后，通过计算代表用户兴趣爱好的主题向量间的距离，计算 2 个用户的相似度。这里的距离采用的是相对熵（KL, kullback leibler divergence），相对熵越小，说明 2 个用户的兴趣越相似。

在计算 2 个用户间的相似度得分时，本文根据式(1)来计算 2 个不同用户在不同平台的行为相似度。2 个用户间的相似度得分定义为

$$\begin{aligned}Sim(u_i,u_j)&=Score(u_i,u_j)\\&=\sum_{k=1}^{e}Score_k(u_i,u_j)\\&=\sum_{k=1}^{e}\frac{\theta_k}{D_k(u_i,u_j)}\\&=\sum_{k=1}^{e}\frac{2\theta_k}{D_k(u_i\,|\,u_j)+D_k(u_j\,|\,u_i)}\\&=\sum_{k=1}^{e}\frac{2\theta_k}{\sum_{t=1}^{R_k}P_{u_i,t}\ln\left(\frac{P_{u_i,t}}{P_{u_j,t}}\right)+\sum_{t=1}^{R_k}P_{u_j,t}\ln\left(\frac{P_{u_j,t}}{P_{u_i,t}}\right)}\end{aligned}\tag{1}$$

其中，e 代表子网数量，R_R 代表主题数量，$P_{u_i,t}=\frac{n_t}{N_{u_i}}$。这里用一个分值来表示 2 个用户间的相似度，这个分值是由用户在各个平台的分数累加得到。用户在每个平台的分数与用户间的距离成反比。本文认为“u_i 和 u_j 间的距离”与“u_j 与 u_i 间的距离”不相等，这是由新浪微博中用户关注关系的实际情况决定的。出于计算简单的考虑，这里将 2 个距离取均值作为最终的距离。

若 2 个用户的相似度得分越高，说明 2 个用户越相似，则需要优先进行推荐。

4 多平台用户特征融合

对于同一个用户，在不同的社交网络平台中利用相似度分别进行推荐时，通常推荐结果未必能保持一致性。为解决如何得到一个合理的、统一的推荐集合的问题，本文将逻辑回归的思想引入所提模型中，将用户在不同平台的推荐结果进行线性拟合。通过挖掘用户信息中内在支配推荐排序的信息，很好地解决了传统使用经验参数的问题，使模型能够自适应不同的应用场景，从而可使用任意的辅助平台进行跨平台推荐。

对于任意的 2 个用户，他们之间融合后的行为相似度由他们在各个平台的行为相似度进行加权线性组合得到。

$$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_k\cdots,\boldsymbol{\theta}_m)\tag{2}$$

其中，$\boldsymbol{\theta}_k$ 表示用户在第 k 个平台上行为的参数，m 表示平台个数。

为了得到权值向量，本文需要先定义预测函数。

$$h_\theta(x)=g(\boldsymbol{\theta}^{\mathrm{T}}x)=\frac{1}{1+\mathrm{e}^{(-\boldsymbol{\theta}^{\mathrm{T}}x)}}\tag{3}$$

其中，$h_\theta(x)$ 的值表示结果候选用户为目标用户已关注好友的概率，x 表示各平台用户推荐排序结果。

在预测函数基础上，可以进一步定义在一次用户推荐过程中的误差函数。

$$Cost(h_\theta(x),y)=\begin{cases}-\lg(h_\theta(x)), y=1\\-\lg(1-h_\theta(x)), y=0\end{cases}\tag{4}$$

在单次用户推荐误差函数式(4)的基础上，可以得到在 m 次推荐中总的误差函数。

$$J_{(\theta)}=\frac{1}{m}\sum_{i=1}^{m}Cost\left(h_\theta(x),y_i\right)\tag{5}$$

权值更新过程如下

$$\boldsymbol{\theta}_{j\_\text{new}}=\boldsymbol{\theta}_j-L\frac{\partial}{\partial_{\boldsymbol{\theta}_j}}J(\boldsymbol{\theta})\tag{6}$$

其中，$\boldsymbol{\theta}_j$ 表示这一轮迭代结束时的权值向量，$\boldsymbol{\theta}_{j\_\text{new}}$ 表示下一轮迭代结束后的权值向量。L 表示搜索步长。

对总误差函数式(5)进行求导，并代入式(6)得到最终的权值更新过程。

$$\boldsymbol{\theta}_{j\_\text{new}} = \boldsymbol{\theta}_j - L\frac{1}{m}\sum_{i=1}^{m}\left(h_q\left(x_i\right)-y_i\right)x_i^j \tag{7}$$

按照式(7)，不断进行迭代，当参数不再发生变化时，认为已经达到收敛条件并结束迭代过程。将模型收敛时得到的参数向量作为本文最终的训练结果。

5　用户推荐

通过对微博数据的分析发现，绝大多数微博用户通过兴趣爱好来关注其他用户，即微博用户和其好友之间拥有相似的话题分布。基于此观察，在候选用户中，优先推荐那些与目标用户话题分布相近的用户。$\boldsymbol{U}$ 为用户集合，u_i 为目标用户，$\boldsymbol{U_i}$ 为候选用户，且 $\boldsymbol{U_i}$=$\boldsymbol{U}$–u_i。

不同的社交网络平台上，对候选用户集中的每一个用户分别与目标用户按照相似度进行降序排列。这样排在前面的用户，与目标用户更相似，需要优先推荐。由于在每个平台上的排序不同，需要对这些用户进行重新排序。推荐过程实现伪代码如算法 1 所示。

算法 1　跨平台用户推荐

输入　用户 u_i 的候选用户集合 $\boldsymbol{U_i}$；平台集合 $\boldsymbol{PL}$；平台权值向量 $\boldsymbol{\theta}$；用户 u_i 粉丝集合 $\boldsymbol{FA}$；用户 u_i 关注用户集合 $\boldsymbol{FO}$；用户概率主题分布向量 $\boldsymbol{P}$；各个平台主题数量向量 $\boldsymbol{R}$；推荐用户数量 K

输出　用户 u_i 推荐列表 $\boldsymbol{T_i}$

1) for each $u\in\boldsymbol{U_i}$；/*$\boldsymbol{U_i}$ 中的每个用户 u*/

2)　for each $q\in\boldsymbol{PL}$；/*$\boldsymbol{PL}$ 中的每个平台 q*/

3)　if $\boldsymbol{P}_{uq}=0$；/*用户在某平台信息为空*/

4)　for each $f\in\boldsymbol{FA}$；/*把粉丝信息赋值给该用户*/

5)　if $\boldsymbol{P}_{fq}$!=0；

6)　$\boldsymbol{P}_{uq}=\boldsymbol{P}_{fq}$

7)　else

8)　for each $f\in\boldsymbol{FO}$；/*把关注者的信息赋值给该用户*/

9)　if $\boldsymbol{P}_{fq}$!= 0

10)　$\boldsymbol{P}_{uq}=\boldsymbol{P}_{fq}$

11) 根据式(1)计算各个用户间的相似度

12) $\boldsymbol{T}_i$=(u_1,u_2,⋯,u_j,⋯,u_k) /*按照与目标用户 u_i 的相似度得分降序排序并取前 K 个用户*/

13) return $\boldsymbol{T}_i$

对推荐集合 $\boldsymbol{T}_i$ 的每个用户 u_j 分别判断该用户是否为目标用户的好友，如果是好友关系则认为此次推荐是成功的。

6　实验

为验证推荐模型的有效性，本文选取国内用户活跃度最大的新浪微博作为目标测试平台，并将国内最大的话题讨论平台知乎作为辅助平台。实验中，通过对 2 个不同平台的融合来向微博用户进行好友推荐。同时，将本文提出的推荐模型 URCP 与文献中的 PYMK[5]、*K*-means[6]、TWILITE[9]等算法进行了一系列的对比实验。在文献[5]中，作者通过用户在 MySpace 平台的信息为用户进行打分，将特征最接近的用户作为推荐列表。在文献[6]中，作者利用用户的文本内容对其进行建模，最后利用 *K*-means 对用户进行推荐。在文献[9]中，作者提取用户在 Twitter 平台的主体分布，推荐前 K 个用户作为推荐列表。本文所提模型 URCP 不仅对目标平台进行建模，同时对辅助平台进行建模，而且最后通过模型融合算法将用户的行为模式融合起来作为用户最终的行为模型，这样可以更加全面地对用户进行分析。

6.1　评价指标

本文通过式(8)和式(9)来计算用户 u_i 的推荐准确率和推荐召回率。

$$precision\left(u_i\right)=\frac{\sum_{j=1}^{k}g\left(u_i,u_j\right)}{K} \tag{8}$$

如果 u_i 关注了 u_j，则 $g(u_i,u_j)$=1；否则，$g(u_i,u_j)$=0。

$$recall\left(u_i\right)=\frac{\sum_{j=1}^{k}g\left(u_i,u_j\right)}{N} \tag{9}$$

其中，N 表示候选用户 u_i 中目标用户的好友数量。

单独使用准确率或召回率无法对一个推荐模型进行全面的评价，这里采用一个统一的评价指标 F_1 值。定义 F_1 值为

$$F_1\left(u_i\right)=\frac{2precision(u_i)recall(u_i)}{precision(u_i)+recall(u_i)} \tag{10}$$

其中，$precision(u_i)$表示用户 u_i 的准确率，$recall(u_i)$表示用户 u_i 的召回率。

用户 u_i 平均准确率（MAP, mean average precision）的计算式为

$$MAP=\frac{1}{r}\sum_{i=1}^{r}\frac{i}{\text{第}i\text{个好友所在的位置}} \tag{11}$$

其中，r 表示候选用户集合中目标用户的好友数量。

除了准确性外，本文还采用了覆盖率（coverage）作为评价指标，用来测评一个推荐系统挖掘长尾用户的能力，定义为

$$coverage=\frac{|\bigcup_{u\in U}\boldsymbol{R}(u)|}{|\boldsymbol{U}|} \tag{12}$$

其中，$\boldsymbol{U}$ 表示系统中所有用户的集合，$R(u)$表示为用户 u 推荐一个长度为 N 的候选用户集合。覆盖率越高，代表该推荐系统越好，有更多的人会被推荐，能够更加有效地缓解推荐系统的马太效应。

6.2 数据集

实验数据集由 4 个部分组成：1) 用户间的关注数据，由 625 万对关注关系组成；2) 用户的微博数据，包含 50 万条用户微博组成；3) 用户的知乎数据，由 5 万条知乎回答组成；4) 账号匹配数据，由 2 万条用户的账号信息组成。在每组实验中，随机选择 $\frac{3}{4}$ 用于训练模型，$\frac{1}{4}$ 用于模型测试。

如果某个用户在知乎平台没有回答过问题，说明该用户在知乎平台不是很活跃，即知乎平台对该用户的兴趣集合影响很小，对于该用户在所有平台的权值向量中，知乎平台对应的值设置为 0。

6.3 参数设置

本文将新浪微博作为目标平台，即向微博用户进行好友推荐。知乎平台作为辅助平台，提高用户的推荐效果。对于每个用户，本文使用 LDA 主题模型来获取其主题分布，采样方法选择 Gibbs 抽样方法。关于 LDA 中比较主要的参数有α、β和主题数量 $topic\_number$。其中，$\alpha=\frac{50}{topic\_number}$，$\beta$一般取经验值 0.01，即$\beta$=0.01。本文根据在测试集合上的推荐效果来确定主题数量。在测试集合的数据上，本文对用户进行好友推荐。根据式(10)得到如图 2 所示的实验结果。

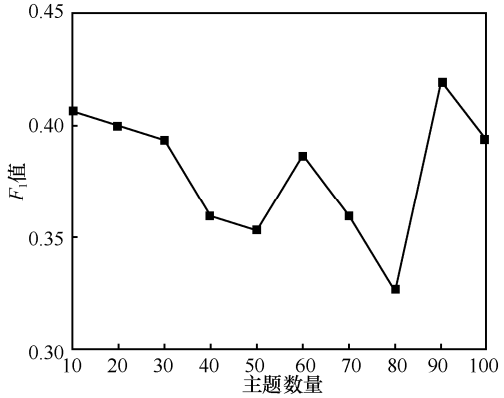


图 2 模型的 F_1 值随主题数量变化

在用户以及用户发布的所有文档确定的情况下，对文档进行主题提取，主题数量不同，得到的用户主题向量不同，最终会影响到后序的用户推荐。由图 2 可知，当主题数量为 90 时，对文档提取出的主题向量是最合适的，因而推荐结果最好。故选取 $topic\_number$=90、$\alpha=\frac{5}{9}$、β=0.01。

在模型训练阶段，本文将训练集数据代入融合模型进行参数训练时需先确定模型的初始向量$\boldsymbol{\theta}$以及搜索步长 L。如果无法选取合适的初始向量，会使收敛到最优值的时间较长，甚至无法收敛到最优值。此外，步长 L 选取过大，会导致算法无法达到最优值；步长 L 选取过小，会导致收敛速度非常慢，实验结果如表 2 所示。

表 2 学习步长与初始向量

L	$\boldsymbol{\theta}\_init$	$\boldsymbol{\theta}\_final$
0.01	(0,0)	(−0.018 571 521, −0.011 475 758)
0.01	(1,1)	(−0.018 571 521, −0.011 475 758)
0.01	(0,1)	(−0.018 571 521, −0.011 475 758)
0.01	(1,0)	(−0.018 571 521, −0.011 475 758)
0.1	(0,0)	(−0.129 991 725, −0.108 686 021)
0.1	(1,1)	(−0.129 991 725, −0.108 686 021)
0.1	(1,0)	(0.060 068 123, 0.077 906 093)
0.1	(0,1)	(0.060 068 123, 0.077 906 093)

通过表 2 可知，当步长 L=0.01 时，无论初始向量取何值，收敛后的向量基本保持不变，为(−0.018 571 521, −0.011 475 758)。因此，在实验中，选取步长 L=0.01。将上述收敛后的向量归一化，得到最终的参数向量$\boldsymbol{\theta}=(0.618,0.382)^{\mathrm{T}}$。归一化过程为

$$0.618\approx\frac{0.018\,571\,521}{0.018\,571\,521+0.011\,475\,758}$$
$$0.382\approx\frac{0.011\,475\,758}{0.018\,571\,521+0.011\,475\,758}\quad(13)$$

6.4　用户推荐结果

将 6.3 节得到的参数代入模型，对用户进行推荐。本文主要采用离线实验方法，将数据分为训练集和测试集，其中，训练集用来对模型进行训练，测试集用来对模型的性能进行测评。本文主要从推荐准确性和覆盖率对模型进行评估，准确性主要包括准确率、召回率、F_1 值和 *MAP* 值，具体定义见 6.1 节。经过实验发现，本文提出的 URCP 模型在准确性和覆盖率方面均优于对比模型，并分别针对 URCP 模型准确性和覆盖率较高的原因进行了详细解释。

根据式(8)，得到实验结果如图 3 所示。

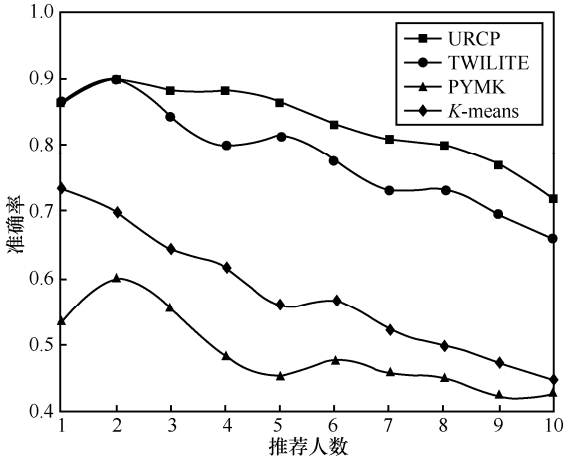


图 3　模型的准确率

由图 3 可知，本文提出的 URCP 模型在准确率方面要优于其他对比模型。随着推荐人数 K 的增大，分母不断增大，分子的增加速度低于分母，导致模型的准确率不断下降。同时也发现，当推荐的用户越少，模型的推荐效果越好。PYMK 模型的准确率随着推荐人数 K 的增加，起伏波动较大，其他 3 个模型基本呈现平缓下降趋势。当候选用户集合长度为 2 时，URCP 模型的准确率最好。

根据式(9)，分析了模型的召回率，实验结果如图 4 所示。

由图 4 可知，本文提出的 URCP 模型在召回率方面要优于其他对比模型。随着 K 的增大，推荐用户增多，有更多的好友被推荐，导致召回率不断增大。从图 4 中可知，URCP 模型和 TWILITE 模型的召回率要明显优于 PYMK 模型和 *K*-means 模型。

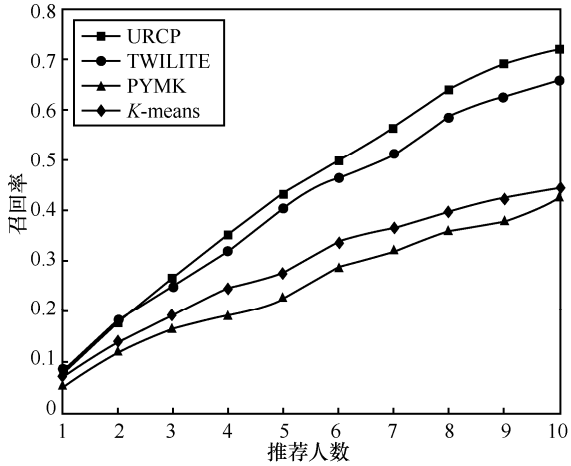


图 4　模型的召回率

在实验中，依据式(10)来计算 F_1 值，图 5 给出了实验结果。

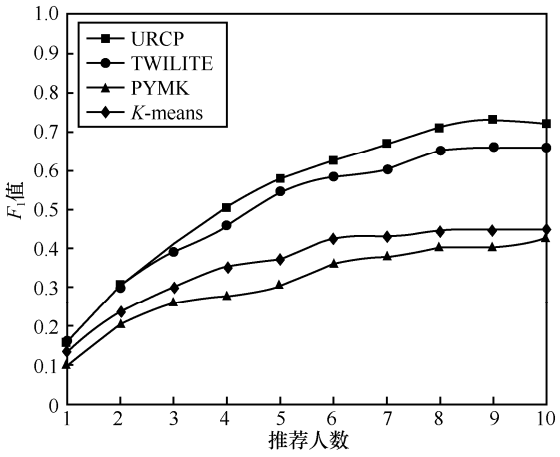


图 5　模型的 F_1 值

由图 5 可知，本文 URCP 模型在 F_1 值方面优于其他模型。综合准确率和召回率的测试结果，发现随着 K 的增大，虽然准确率在不断减小，但是召回率和 F_1 值在不断增大。当推荐人数大于 6 时，准确率的下降速度与召回率的增长速度基本持平，使 F_1 值增长速度变缓。当推荐人数大于 9 时，由于准确率的下降速度大于召回率的增长速度，导致 URCP 模型的 F_1 值开始呈现下降趋势。

依据式(11)，测试了对比模型的 *MAP*，测试结果如图 6 所示。

由图 6 可知，本文提出的 URCP 模型在平均准确率方面要优于对比模型。*MAP* 值越大，说明模型能够把目标用户最喜欢的用户放到推荐列表的靠前位置，能够更加准确地刻画用户的兴趣爱好，推荐效果更好。URCP 模型的 *MAP* 值明显高于其他 3

个模型，说明该模型为目标用户推荐的候选用户更能满足其兴趣爱好。

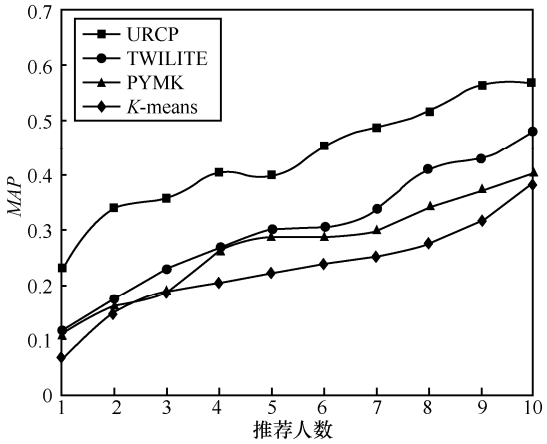


图 6 模型的 *MAP* 值

图 3~图 6 的测试结果显示，相比既有模型，本文 URCP 模型在推荐的准确性方面具有显著的优越性。对于 TWILITE 模型，只是在新浪微博平台对用户进行主题提取，并推荐兴趣爱好最接近的前 *K* 个用户。利用这种模型得到的用户兴趣爱好是片面的，对用户的兴趣爱好描述的不够全面。例如，用户在新浪微博平台发布足球相关的博文，但是其对军事也很感兴趣，却很少表现出相关的行为动作，导致该模型无法有效地刻画出用户在军事方面的兴趣爱好，因此推荐效果不是很理想；对于 *K*-means 模型，利用用户发布的文档之间的相似性对用户进行好友推荐，效果不如通过提取文档主题来进行推荐。现有的自然语言处理技术，还无法很好地对文档直接进行描述，例如对同义词的处理。文档 A 中出现了“推荐系统”一词，文档 B 中出现了“推荐引擎”一词，在计算 2 个文档间的距离时，会认为这是 2 个完全不同的单词，导致文档差异性较大。本文提出的模型将文档进行主题提取后，能够更好地反映用户兴趣爱好；对于 PYMK 模型，利用协同过滤的思想进行用户推荐，由于系统中有很多用户更加倾向于关注大量其他用户，自己却很少发布博文，于是基于协同过滤的模型会将这些出度特别大但并不经常发布微博的用户推荐给目标用户。其次由于社交网络的数据稀疏性，用户之间无法通过关注关系很好地联系起来，于是很多候选用户与目标用户兴趣爱好相似却不会被推荐，导致推荐准确性不高；本文 URCP 模型，分别提取用户在各个平台发布的文章，利用隐语义模型提取文章的主题分布，用文档的主题分布表示用户的兴趣爱好，可以更好地对用户行为进行描述，并利用回归模型将用户在所有平台的兴趣爱好进行综合考虑，可以更加全面地对用户的兴趣爱好进行刻画，因此，推荐效果比较好。

一个好的推荐系统，不仅要对用户进行准确的推荐，还需要尽可能地保证每一个用户都有机会被推荐给其他用户，防止关注度越高的用户越容易被关注，关注度较低的用户更加不被关注。依据式(12)，本文对模型的覆盖率进行了测评，测试结果如图 7 所示。

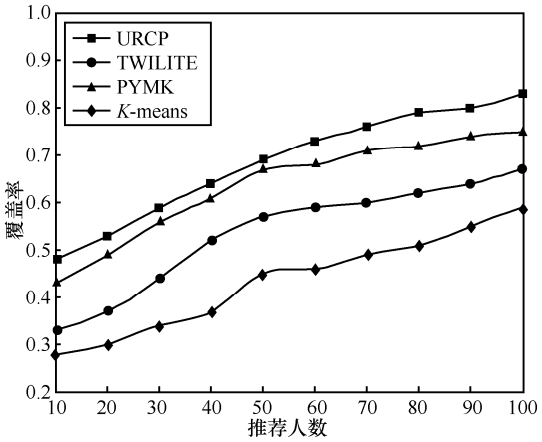


图 7 模型的覆盖率

由图 7 可知，本文提出的 URCP 模型在覆盖率方面要优于其他对比模型。基于 *K*-means 的聚类模型，会形成以大 V 用户为中心的极大簇，推荐给目标用户的候选用户，很大一部分都是距离簇心较近的用户，因此，覆盖率较低；基于协同过滤的 PYMK 模型，主要通过关注关系对用户进行推荐，由于新浪微博中数据较为稀疏，导致很多用户无法被推荐，覆盖率较低[24]；基于主题模型的 TWILITE 模型，对用户进行兴趣爱好建模，根据用户的兴趣爱好相似度进行推荐，因此，推荐列表中的用户大都集中在某个主题领域，所以覆盖率较小；本文 URCP 模型，虽然也是利用兴趣爱好相似度进行推荐，但是该模型不仅刻画了用户在新浪微博平台的行为特征，还综合考虑了用户在其他平台的兴趣。例如，该模型为用户 A 推荐了用户 B，是因为用户 A 喜欢“足球”相关的内容，虽然用户 B 在新浪微博平台并没有相关的行为特征，但是在知乎平台发布了很多“足球”相关的文章，也会被推荐给用户 A，因此，该模型具有较高的覆盖率。

7　结束语

在现有的在线社交网络用户推荐方法中，大部分是基于单平台的用户推荐。在单个平台上，无法全面地理解用户行为。此外，在单个平台上，容易发生用户冷启动现象。对于一个新加入的用户，人们无法获取其行为特征，不能有效地对其进行好友推荐。因此，如何有效进行用户推荐仍是一个具有挑战性的问题。如果能融合其他平台的数据进行跨平台推荐，则会大大提高用户推荐的准确性。然而，已有的一些利用跨平台进行用户推荐的方法，虽可以在一定程度上解决用户冷启动问题，但是其数据采集方法和用户推荐模型不具有可扩展性。基于此，本文提出了一种新的跨平台数据采集方法，具有较好的扩展性。此外，本文提出的跨平台用户推荐方法，不仅具有很好的推荐效果，而且可以较好地移植到其他跨平台研究。本文采用基于跨平台的用户推荐方法有效融合其他平台的用户信息，能够更加全面地对用户进行兴趣爱好建模，更加准确地进行用户推荐。基于真实数据集上的实验结果表明，本文提出的基于跨平台的用户推荐模型可以提高用户推荐效果。

参考文献：

[1] DUAN J, AI Y. LDA topic model for microblog recommendation[C]//International Conference on Asian Language Processing (IALP). 2015: 185-188.

[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]. Journal of Computer, 2013, 36(2): 349-359.

[3] 尚燕敏, 张鹏, 曹亚男. 融合链接拓扑结构和用户兴趣的朋友推荐方法[J]. 通信学报, 2015, 36 (2): 117-125.
SHANG Y M, ZHANG P, CAO Y N. New interest-sensitive and network-sensitive method for user recommendation[J]. Journal on Communications, 2015 , 36 (2) :117-125.

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

[5] MORICZ M, DOSBAYEV Y, BERLYANT M. PYMK: friend recommendation at myspace[C]//The 2010 ACM SIGMOD International Conference on Management of data. 2010: 999-1002.

[6] DENG Z, HE B, YU C, et al. Personalized friend recommendation in social network based on clustering method[M]//Computational Intelligence and Intelligent Systems. Springer Berlin Heidelberg, 2012: 84-91.

[7] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[8] DAS A S, DATAR M, GARG A, et al. Google news personalization: scalable online collaborative filtering[C]//The 16th international conference on World Wide Web. 2007: 271-280.

[9] KIM Y, SHIM K. TWILITE: a recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation[J]. Information Systems, 2013, 42(3): 59-77.

[10] ABEL F, ARAUJO S, GAO Q, et al. Analyzing cross-system user modeling on the social Web[C]//International Conference on Web Engineering. 2011: 28-43.

[11] DENG Z, YAN M, SANG J, et al. Twitter is faster: personalized time-aware video recommendation from Twitter to YouTube[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2013, 11(2): 1-23.

[12] DENG Z, SANG J, XU C. Personalized video recommendation based on cross-platform user modeling[C]//IEEE International Conference on Multimedia and Expo (ICME). 2013.

[13] ROY S D, MEI T, ZENG W, et al. Social transfer: cross-domain transfer learning from social streams for media applications[C]//The 20th ACM International Conference on Multimedia. 2012: 649-658.

[14] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[15] ZHOU X, LIANG X, ZHANG H, et al. Cross-platform identification of anonymous identical users in multiple social media networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(2): 411-424.

[16] KONG X, ZHANG J, YU P S. Inferring anchor links across multiple heterogeneous social networks[C]//The 22nd ACM international conference on Information & Knowledge Management. 2013: 179-188.

[17] GAGA O, LEI H, PARTHASARATHI S H K, et al. Exploiting innocuous activity for correlating users across sites[C]//The 22nd International Conference on World Wide Web. 2013: 447-458.

[18] NARAYANAN A, SHMATIKOV V. De-anonymizing social networks[C]//30th IEEE Symposium on Security and Privacy. 2009: 173-187.

[19] SALEM Y, HONG J, LIU W. CSFinder: a cold-start friend finder in large-scale social networks[C]//IEEE International Conference on Big Data (Big Data). 2015: 687-696.

[20] MANDL M, FELFERNIG A. Improving the performance of unit critiquing[C]//International Conference on User Modeling, Adaptation, and Personalization. 2012: 176-187.

[21] MCCARTHY K, SALEM Y, SMYTH B. Experience-based critiquing: reusing critiquing experiences to improve conversational recommendation[C]//International Conference on Case-Based Reasoning. 2010: 480-494.

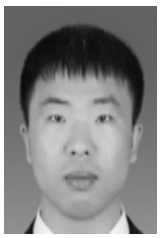
[22] SALEM Y, HONG J. History-aware critiquing-based conversational recommendation[C]//The 22nd International Conference on World Wide Web. 2013: 63-64.

[23] JONGHYUK S, SANGHO L, JONG K. Spam filtering in twitter using sender-receiver relationship[J]. Recent Advances in Intrusion Detection-international Symposium, 2011, 6961: 301-317.

[24] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[作者简介]

彭舰（1970-），男，四川成都人，博士，四川大学教授，主要研究方向为大数据、传感器计算、移动计算等。



王屯屯（1992-），男，河南安阳人，四川大学硕士生，主要研究方向为数据挖掘、推荐系统、用户行为建模等。

陈瑜（1974-），男，四川成都人，博士，四川大学讲师，主要研究方向为进化计算、机器学习等。

刘唐（1980-），男，四川乐山人，博士，四川师范大学副教授，主要研究方向为无线传感器网络、无线能量传输等。

徐文政（1985-），男，四川成都人，博士，四川大学副研究员，主要研究方向为社交网络、物联网、移动计算。