文章编号：1003-0077(2018)09-0103-10

微博网络用户的活跃性判定方法

仲兆满[1,2]，戴红伟[1]，管 燕[1]

(1. 淮海工学院 计算机工程学院，江苏 连云港 222005；
2. 江苏金鸽网络科技有限公司 大数据事业部，江苏 连云港 222005)

摘 要：推荐系统的冷启动问题是近期的研究热点，而用户的活跃性判定是冷启动问题的基础。已有方法在判定用户的活跃性时，单纯地考虑了用户发表信息量，对社交媒体的社交关系及行为等特征利用不够。该文面向微博网络，提出了系统的用户活跃性判定方法，创新性主要体现在：(1)提出了微博网络影响用户活跃性的四类指标，包括用户背景、社交关系、发表内容质量及社交行为，避免了仅仅使用用户发表信息数量判定用户是否活跃的粗糙方式；(2)提出了用户活跃性判定流程，提出了基于四类指标的用户与用户集的差异度计算模型。以新浪微博为例，选取了学术研究、企业管理、教育、文化、军事五个领域的 900 个用户作为测试集，使用准确率 P、召回率 R 及 F 值为评价指标，进行了实验分析和比较。结果显示，该文所提用户活跃性判定方法的准确率 P、召回率 R、F 值比传统的判定方法分别提高了 21%、13%和 16%，将该文所提方法用于用户推荐，得到的 P、R 和 F 值比最新的方法分别提高了 5%、2%和 3%，验证了所提方法的有效性。

关键词：微博推荐系统；用户活跃性判定；用户背景；用户社交关系；用户发表内容质量；用户社交行为

中图分类号：TP391 **文献标识码**：A

User Activeness Determination in Microblog

ZHONG Zhaoman[1,2], DAI Hongwei[1], GUAN Yan[1]

(1. School of Computer, Huaihai Institute of Technology, Lianyungang, Jiangsu 222005, China;
2. Department of Big Data, Jiangsu Jinge Network Technology Co. Ltd., Lianyungang, Jiangsu 222005, China)

Abstract: To determining the user activeness, the existing methods mainly centered on the amount of information users posted, without proper utilizing the users' social relationship and behavior on microblog. This paper proposes a systematic method of determining the user activeness on microblog. In this method, four indexes are introduced to determinate users' activeness on microblog, including users' profile, social relationship, information quality and social behavior. And we also present the flow of determining the user activeness, and computation model for the diversity between a user and the whole user set. From Sina microblog, we select 900 users as the test set from the domain of academic research, business management, education, culture and military. Precision, Recall and F-value were used as evaluation index for experimental analysis and comparison among methods. The results show that our method improves the precision, recall and F-value of the user activeness determination by 21%, 13% and 16%, respectively. Applying the proposed method to user recommendation, the precision, recall and F-value are increased by 5%, 2% and 3%, respectively.

Key words: recommendation system on Microblog; users' activeness determination; users' profile; users' social relation; users' post quality; users' social behavior

收稿日期：2017-10-09 **定稿日期**：2017-12-06

基金项目：国家自然科学基金(61403156)；江苏省六大人才高峰基金资助(XXRJ-013)；江苏高校品牌专业建设工程资助(PPZY2015A038)；连云港市 521 高层次人才基金资助

0 引言

起初，微博主要用于人们社交的需求，通过“关注”(follow)在微博网络上很容易形成类似于现实社会的交往圈子。但目前，微博已成为面向大众的舆论平台，已成为网民获取信息的重要途径之一，越来越多机构及公众人物都通过微博来发布或传播信息。

面向微博网络的推荐系统，总体上可以分为两类：(1)微博信息推荐，根据用户的兴趣取向，从微博网络海量的信息中挖掘出用户感兴趣的内容推荐给用户，避免了用户在微博网络上漫无目的地查询信息，比如高明等人[1]基于LDA主题模型推断微博的主题分布和用户的兴趣取向，提出了微博网络上用户感兴趣微博的实时推荐方法，Chen等人[2]综合微博的主题分布因子、用户在社交网络中的影响力特征、微博的内容特征以及微博的受欢迎程度等特征为用户提供个性化搜索结果；(2)用户推荐，根据用户的背景、兴趣爱好、关注领域，为用户推荐志同道合的朋友，比如文献[3-5]都从不同侧面研究了微博网络上相似用户的推荐方法。

社交媒体上虽然有大量用户，但已有研究表明，用户的活跃度符合幂律分布，即只有少量用户是活跃的，大多数用户是非活跃的(冷启动的)。Zeng等人[6]公开的研究成果认为，大约20%的用户是活跃的。文献[7]在研究用户的标签时，通过对新浪微博的1.4亿用户统计，发现标签数小于5的用户占用户总数的93.8%，从用户标签的角度来讲，活跃用户不到10%。因为从微博网络上获取冷启动用户的相关信息很少，无论是给冷启动用户推荐微博信息，还是推荐相似用户，都成为非常困难的工作。

有些文献提出了为冷启动用户寻找替代用户的信息挖掘方法。比如，Akcora等[8]在计算社交网络用户的相似度时综合了用户的背景信息和网络结构，在Facebook平台统计发现，64%的用户缺少背景信息的描述，提出了从用户朋友已有的数据中，自动推理出用户的一些可能的背景信息；Lin等[9]使用了Twitter上的社交信息(关注者-followers)帮助解决APP推荐的冷启动问题。

可见，在微博推荐系统中，如果用户是活跃的，可以直接基于用户的历史信息进行微博信息或者用户的推荐，而对于冷启动用户，可以使用替代用户的方法进行推荐。用户活跃性的判定是推荐系统的首要工作。然而，关于用户活跃性判定的研究文献很少，少有的几篇文献中对冷启动用户的判定方式非常粗糙，普遍认为如果用户的评论信息量少于一定的阈值就认为该用户是冷启动的。比如，文献[10-11]认为用户评论的信息量少于5条的是局部冷启动用户，文献[12]认为用户发表的信息量少于20条的为局部冷启动用户。这些方法，没有考虑用户发表信息的质量，没有考虑微博等社交媒体用户具有的社交关系特征。

针对微博网络的用户活跃性判定问题，本文进行了系统的研究，创新性主要体现在：(1)面向微博网络，提出了衡量用户活跃性的四类指标：背景、社交关系、发表内容质量及社交行为，不再仅仅局限于用户评论数量的多少，这样做能更好地体现微博类社交媒体的特性；(2)在综合的考虑微博用户各类活跃性因素的基础上，提出了用户活跃性判定的整套流程，以及用户与活跃用户集/冷启动用户集的差异度计算模型。

本文后续内容安排如下：第1节介绍了已有的相关研究工作；第2节详细地阐述了本文所提方法的原理和流程，包括相关定义，微博用户活跃性判定的四类指标，用户与用户集的差异度计算模型；第3节使用了准确率、召回率及F值从用户活跃性判定的效果、不同指标的权重等方面进行了实验的比较与分析，以验证本文所提方法的有效性；第4节对本文进行了总结，探讨了该方法的优缺点，以及未来的研究方向。

1 相关工作

推荐系统的方法总体分为两类[13]：(1)基于内容的推荐，根据用户u_i的历史信息，如评价、分享、收藏过的文档等，构造用户u_i偏好模型，将属性相似度高的项目向用户u_i做出推荐。可以看出：基于内容的推荐技术从项目角度出发，寻找相似项目；(2)协同过滤推荐，构建用户—项目评价矩阵，计算用户间的相似度，将与用户u_i相似度高的用户评分高的项目向用户u_i做出推荐。可以看出：协同过滤推荐技术从用户角度出发，寻找相似用户。

但是，如果一个新的项目在评分矩阵中很少有用户为它评价，或者一个新用户在评分矩阵中很少对项目进行过评价，则无法使用推荐算法实现有效的推荐，这就是推荐系统中经典的冷启动问题[14]。

为了解决用户冷启动问题，一些研究者提出了

基于用户之间信任关系的推荐思想[15-16]。这种推荐思想考虑了网络用户之间的关系，根据用户的直接或间接信任用户预测其对项目的评分。

Ocepek 等人[17]将用户的冷启动分为两种情况：(1)用户没有任何评论信息，称为绝对冷启动(absolute cold start)；(2)用户有很少的评论信息，称为局部冷启动(partial cold start)，并分为五种情况，有 1 条、2 条、3 条、4 条和 5 条评论信息的分别记为 CS1、CS2、CS3、CS4 和 CS5。可见，Ocepek 等人是将小于等于五条评论信息的用户作为局部冷启动用户。

于洪等[18]针对完全新项目，即不存在任何一个用户曾经对该项目评价过，在充分考虑用户、标签、项目属性、时间等信息的基础上，获得个性化的预测评分值，用于解决新项目冷启动的问题，并提出了积极用户(喜欢去关注并评价新事物的用户)和消极用户(比较喜欢去关注已经被很多用户评价过的事物)的概念，进一步用时间权重进行区分。

Pereira 等[12]提出了一种基于同步聚类和学习技术的混合推荐方法(SCOAL)，针对绝对冷启动用户和局部冷启动用户进行了实验分析，核心问题是将某个冷启动用户划分到合适的类里去，选取的用户特征仅仅是用户的评论信息条数，将发表信息少于 20 条的视为局部冷启动用户。

已有判定用户活跃性的指标单一、方法粗糙，普遍认为发表信息量等于 0 的为绝对冷启动用户，小于 α 条的直接认定为局部冷启动用户，除此之外的用户都认定为活跃用户，文献[10-11，17，19]都将 α 设为 5，而文献[12] 将 α 设为 20。

2　微博用户的活跃性判定方法

2.1　相关定义

定义 1　微博网络[3]，形式化描述为一个六元组：$\mathrm{MBN}=\{U,\mathrm{MBlog},E_{\mathrm{UMB}},E_{\mathrm{UU}},F_{\mathrm{UMB}},C_{\mathrm{UMB}}\}$，其中，$U$ 为微博上的注册用户集；MBlog 为用户发表的微博集(含原创、转发或者评论的各类微博)；$E_{\mathrm{UMB}}=\{e=(u_i,\mathrm{mblog}_j)\,u_i\in U,\mathrm{mblog}_j\in\mathrm{MBlog}\}$ 为用户与其所发表微博的连接边集；$E_{\mathrm{UU}}=\{u_i\rightarrow u_j)\mid u_i \text{ follows } u_j\}$ 为用户通过关注而形成的连接关系集，通过 follow 关系容易得到用户的粉丝关系集；$F_{\mathrm{UMB}}=\{(u_i,\mathrm{mblog}_j)\mid u_i\in U,u_i \text{ forwarded } \mathrm{mblog}_j\}$ 是用户与其所转发的微博的关系集；$C_{\mathrm{UMB}}=\{(u_i,\mathrm{mblog}_j)\mid u_i\in U,u_i \text{ commented on } \mathrm{mblog}_j\}$ 是用户与其所评论的微博的关系集。

定义 2　微博用户[3]，形式化描述为一个六元组：$u_i=\{u_i\_\mathrm{Name},u_i\_\mathrm{Bg},u_i\_\mathrm{MBlog},u_i\_\mathrm{Follower},u_i\_\mathrm{Followee},u_i\_\mathrm{Visitor}\}$，其中，$u_i$_Name 为微博的用户名，是微博网络中用户的唯一标识符；u_i_Bg 为微博网络上的用户背景，不同微博网络背景有所差异；u_i_MBlog 为用户在微博网络上发表的微博集；u_i_Follower 为用户的关注集；u_i_Followee 为用户的粉丝集；u_i_Visitor 为用户的访客集，访客类用户指没有与用户 u_i 构建关注和粉丝关系，但与 u_i 进行了微博互动，包括发表微博时的“@”、转发或者评论行为。

依据定义 2，可以容易地获取用户 u_i 的关注数量$|u_i_\mathrm{Follower}|$及粉丝数量$|u_i_\mathrm{Fans}|$。

定义 3　用户背景，在微博网络上，用户具有的自身信息的描述及系统自动赋予的级别，包括简介、学习工作经历、兴趣标签、微博等级等，称为用户的背景。

定义 4　用户社交关系，在微博网络上，用户通过关注(follow)关系构建了紧密的社交圈子，在这种社交圈子中用户拥有的粉丝及关注称为用户的社交关系。

定义 5　用户发表信息质量，在微博网络上，用户发表的微博信息(包括原创和评论的内容)、发表信息的受众称为用户发表信息质量。

定义 6　用户社交行为，在微博网络上，用户转发、点赞、收藏等行为称为用户社交行为。

定义 7　冷启动用户，在微博网络上，用户在背景、社交关系、发表信息质量、社交行为等诸多方面都不活跃的用户称为冷启动用户。

定义 7 与已有的冷启动用户定义的不同体现在，在衡量冷启动用户时，用户发表信息质量只是特征之一。不同的社交行为都是用户在微博网络上的活跃性体现，包括转发、点赞、收藏等社交行为。用户背景的完善程度、用户的微博等级也能反映用户的活跃程度。此外，用户的活跃性还受社交关系的影响，在微博网络上用户的社交关系(关注/粉丝)能够反映用户的活跃程度，比如用户 u_i 经常更新关注对象，或者经常有其他用户关注用户 u_i。定义 7 给出的微博用户活跃性判定的指标更为全面、更加准确，比如用户 u_i 发表微博信息很少，但其社交行为比较频繁，那么用户 u_i 的活跃性也比较高。

又如，用户 u_1 发表了 5 条信息，已有的方法将

u_1 判定为冷启动用户，用户 u_2 发表了 6 条信息，已有的方法将 u_2 判定为活跃用户。但用户 u_1 发表的 5 条信息可能质量较高，引起了大量的阅读、转发行为，而用户 u_2 发表的 6 条信息，都是简短的评论，质量较低。用户 u_1 发表信息产生的影响力远远大于用户 u_2，用户 u_1 相比用户 u_2 更为活跃。

2.2 微博用户活跃性判定方法流程

任意给定一个用户 u_i，判断其是属于活跃用户，还是冷启动用户，该问题的判定流程如下：

(1) 输入用户 u_i；

(2) 计算用户 u_i 发表信息的质量 IQ_{u_i}，如果 $IQ_{u_i} < \alpha$，则 u_i 是冷启动用户，转步骤(6)，如果 $IQ_{u_i} > \beta$，则 u_i 是活跃用户，转步骤(6)，否则，将 u_i 标识为边缘用户 mu_i（即用户发表信息质量没有明显的特征），转步骤(3)；

(3) 对于边缘用户 mu_i，借助其他用户集进行判定。在微博网络上任意采集 n 个用户，依据用户发表信息的质量，将 n 个用户划分为两个集合：活跃用户集（简记为 AU）和冷启动用户集（简记为 IAU）；

(4) 分别提取边缘用户 mu_i 及用户集 AU、IAU 与活跃性相关的四类指标，计算用户 mu_i 与 AU 的差异度 diversity(mu_i,AU)，用户 mu_i 与 IAU 的差异度为 diversity(mu_i,IAU)；

(5) 如果 diversity(mu_i,AU)>diversity(mu_i,IAU)，则 mu_i 是活跃用户，否则 mu_i 是冷启动用户；

(6) 输出判定结果，即用户 u_i 是属于活跃用户，还是冷启动用户。

步骤(2)判定用户 u_i 的活跃性可以描述为式(1)。

$$A(u_i)=\begin{cases}\text{inactive} & IQ_{u_i}\leqslant\alpha\\ \text{marginal} & \alpha<IQ_{u_i}<\beta\\ \text{active} & IQ_{u_i}\geqslant\beta\end{cases}\quad(1)$$

其中 $A(u_i)$ 为用户的活跃性表示，inactive 表示用户 u_i 是冷启动的（非活跃的），active 表示用户 u_i 是活跃的，marginal 表示用户 u_i 是边缘的，即还没能明确判定出其属于活跃的，还是冷启动的，(α,β) 为决策阈值对。

α 值用于直接确定用户是否是冷启动的，参考文献[10-11,17,19]，定义 $\alpha=5$。β 值用于直接确定用户是否是活跃的，为了选出真正的活跃用户，应该让 β 的取值偏大（本文中 $\beta=30$）。对于模棱两可的用户，可以归为边缘用户，进一步使用其他指标判定其活跃性。

步骤(2)中的用户发表信息质量计算的方法，详见 2.3.1 小节的论述。

步骤(3)采集用户时，n 的取值尽量大，这样更能进行显著性分析。

步骤(4)提取与用户活跃性相关的四类指标，进行差异度计算，详见 2.3 节的论述。

该问题的解决思路，同样可以用于批量用户的活跃性判定问题：任意给定 k 个用户，判定 k 个用户的活跃性。如果 k 大于一定的阈值，则不需要采集其他用户，直接将 k 个用户划分为冷启动用户集和活跃用户集，否则，需在微博网络上采集其他的用户。所不同的是，由于是对多个用户的活跃性进行判定，当一个用户加入活跃用户集/冷启动用户集时，需要更新对应的用户集的相关指标。

2.3 微博用户活跃性的差异度计算模型

对于边缘用户 mu_i，仅仅根据其自身发表的信息质量难以直接将其判定为是活跃用户，还是冷启动用户，尤其对于与冷启动阈值 α 比较接近的用户，强制性地根据用户发表信息的质量判定某用户是否为冷启动用户，势必造成判定结果的偏差。

Balcan[20] 等人研究了人类的流动性（human mobility）和传染病地域扩散之间的关系。人类动力学研究启发我们，尽管每一个人的行为具有随机性和不可预测性，但是在群体结构上具有较强的规则性，这为判定用户的活跃性提供了基础。

在 2.2 节阐述的流程中，根据冷启动阈值 α、活跃用户阈值 β，可以将一些活跃性是否明显的用户直接分为两个集合，基于这两个集合，利用两个集合的群体特征，对边缘用户 mu_i 进行活跃性的判定。

对微博网络上的用户而言，我们提出的边缘用户 mu_i 与活跃用户集 AU 的差异度计算模型如图 1 所示。

图 1 所示的计算模型中，$Profile_{mu_i}$ 指用户的背景，SR_{mu_i} 指用户的社交关系，PC_{mu_i} 指用户的发表信息质量，SB_{mu_i} 指用户的社交行为，在分别提取边缘用户 mu_i、活跃用户集 AU 的四类指标的基础上，可以计算 mu_i 与 AU 的差异度。

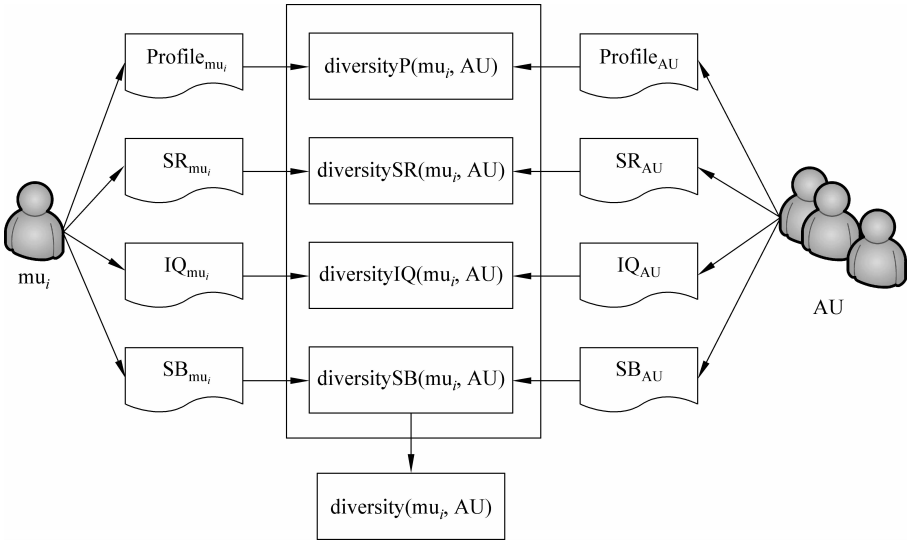

图 1　边缘用户与活跃用户集的差异度计算模型

类似于图 1 所示的计算模型，可计算边缘用户 mu_i 与冷启动用户集 IAU 的差异度。

结合 2.1 节微博用户的背景、社交关系、发表信息质量、社交行为的定义，我们给出微博网络上四类指标的具体计算方法。

2.3.1　用户背景及其差异度计算

微博网络上，微博等级是用户活跃和荣誉的见证，随着用户在微博网络上的探索和成长，等级会随之增长。不同的微博网络，关于“等级”的设置不同，比如，对新浪微博而言，其“等级”共分 24 个，能体现微博用户背景的丰富程度、发表微博信息的活跃程度等指标。用户 mu_i 的微博等级值用如式(2)方式量化。

$$\text{level}(mu_i) = \text{level}_{mu_i}/\text{maxlevel} \tag{2}$$

其中，level_{mu_i} 代表用户 mu_i 的当前等级，maxlevel 表示某微博平台上的最高等级。

在用户众多的背景信息中，用户的标签是自定义描述职业、兴趣爱好的关键词。用户在微博网络上定义自己的标签时，既可以在微博网络的标签库中选取，也可以人工输入。已有的文献普遍认为用户标签在描述用户的兴趣偏好时有重要的参考价值，比如文献[21-22]。通过标签，可以让更多人找到自己，让用户找到更多同类。用户 mu_i 的标签值用如式(3)方式量化。

$$\text{tag}(mu_i) = \text{tag}_{mu_i}/\text{maxtag} \tag{3}$$

其中，tag_{mu_i} 代表用户 mu_i 的标签个数，maxtag 表示某微博网络上最大标签个数。比如，新浪微博，每个用户最多可添加十个标签。

用户 mu_i 的 Profile_{mu_i} 计算方法如式(4)所示。

$$\text{Profile}_{mu_i} = (\text{level}(mu_i) + \text{tag}(mu_i))/2 \tag{4}$$

活跃用户集 AU 的 Proflie_{AU} 计算方法如式(5)所示。

$$\text{Profile}_{AU} = \sum_{j=1}^{|AU|} \text{Profile}_{u_j}/|AU| \tag{5}$$

用户 Profile_{mu_i} 与用户集 Profile_{AU} 的差异度计算方法如式(6)所示。

$$\begin{aligned}&\text{diversityProfile}(\text{Profile}_{mu_i}, \text{Profile}_{AU})\\ &= |\text{Profile}_{mu_i} - \text{Profile}_{AU}|\end{aligned} \tag{6}$$

2.3.2　用户社交关系及其差异度计算

用户在微博上，通过关注形成了紧密的社交圈子，社交关系是微博网络上用户互动的基础。对于任意一个用户 mu_i，可以方便地得到其关注(follower)的数量、粉丝(followee)的数量。

关注(follower)的数量可以反映用户 mu_i 对微博网络上其他用户的关注程度，被关注用户发表的信息将会直接推送给用户 mu_i，增加了用户 mu_i 获取信息的可能性。用户 mu_i 关注的用户越多，说明其获取信息的量越大，越可能在微博网络上进行社交活动(包括转发、收藏、点赞等社交行为)。而粉丝(followee)的数量可以反映用户 mu_i 在微博网络上对其他用户的影响力，粉丝数越多，说明用户 mu_i 在微博网络上越有影响力，其有可能越活跃。

用户 mu_i 的 SR_{mu_i} 计算方法如式(7)所示。

$$\text{SR}_{mu_i} = \text{num}(\text{follower} + \text{followee})/\text{maxnum} \tag{7}$$

其中，num(follower+followee)表示用户 mu_i 的关注和粉丝之和，maxnum 表示在收集的所有用户中，粉丝数和关注数之和最大的用户的粉丝数和关注数之和。

活跃用户集 AU 的 SR_{AU} 计算方法如式(8)所示。

$$SR_{AU} = \sum_{j=1}^{|AU|} SR_{u_j} / |AU| \tag{8}$$

SR_{mu_i} 与 SR_{AU} 的差异度计算方法如式(9)所示。

$$diversitySR(SR_{mu_i}, SR_{AU}) = |SR_{mu_i} - SR_{AU}| \tag{9}$$

2.3.3 用户发表信息质量及其差异度计算

用户在微博网络上发表信息时,有原创、评论两种方式。用户发表的一条信息的质量计算方法如式(10)所示。

$$IQ_{infor_1} = \begin{cases} 1 + \log_e^{len(infor_1)} \\ \quad + extra(infor_1) & infor_1 \text{ 是原创内容} \\ 1 + \log_e^{len(infor_1)} & infor_1 \text{ 是评论内容} \end{cases} \tag{10}$$

式(10)中,将用户发表信息区分为两种情况,对于原创内容,考虑信息的长度及信息的附加值,信息长度越长,包含的信息量越大,用户花费在微博平台的时间也越长。附加值包括发表内容引起的转发数、评论数及点赞数。

一条信息 $infor_1$ 的附加值的计算方法如式(11)所示。

$$extra(infor_1) = 1 + \log_e^{num(comment+forward+like)} \tag{11}$$

其中,comment、forward 及 like 分别表示信息 $infor_1$ 的评论数、转发数及点赞数。

用户 mu_i 发表信息的质量计算方法如式(12)所示。

$$IQ_{u_i} = \sum_{i=1}^{s} IQ_{infor_i} / maxIQ \tag{12}$$

其中,$\sum_{i=1}^{s} IQ_{infor_i}$ 表示用户 mu_i 发表的多条信息的质量之和,maxIQ 表示发表信息质量最大的用户的信息质量。

活跃用户集 AU 的 IQ_{AU} 计算方法如式(13)所示。

$$IQ_{AU} = \sum_{j=1}^{|AU|} IQ_{u_j} / |AU| \tag{13}$$

IQ_{u_i} 与 IQ_{AU} 的差异度计算方法如式(14)所示。

$$diversityIQ(IQ_{u_i}, IQ_{AU}) = |IQ_{u_i} - IQ_{AU}| \tag{14}$$

2.3.4 用户社交行为及其差异度计算

用户在微博网络上,可以对其他用户发表的微博进行转发、点赞、收藏等社交行为,这种行为虽然没有发表微博信息包含的信息量大,但也能反映用户在微博网络上的活跃程度。

目前,微博网络上任意一个用户的点赞数量及点赞内容对其他用户是开放的,但无法获取用户收藏的微博数量及转发的数量。所以,本文只使用了点赞这类社交行为。

用户 mu_i 的 SB_{mu_i} 计算方法如式(15)所示。

$$SB_{mu_i} = num(like)/maxnum \tag{15}$$

其中,num(like)表示用户 mu_i 的点赞次数,maxnum 表示所有收集到的用户中,点赞次数最多的用户的点赞次数。

用户集 AU 的 SB_{AU} 计算方法如式(16)所示。

$$SB_{AU} = \sum_{j=1}^{|AU|} SBu_j / |AU| \tag{16}$$

SB_{u_i} 与 SB_{AU} 的差异度计算如式(17)所示。

$$diversitySB(SB_{u_i}, SB_{AU}) = |SB_{u_i} - SB_{AU}| \tag{17}$$

2.3.5 微博用户活跃性的差异度计算

在 2.3.1 至 2.3.4 小节分别得到边缘用户 mu_i 与活跃用户集 AU 的背景差异度 diversityProfile($Profile_{mu_i}$, $Profile_{AU}$)、社交关系差异度 diversitySR(SR_{mu_i}, SR_{AU})、发表信息质量差异度 diversityIQ(IQ_{u_i}, IQ_{AU})和社交行为差异度 diversitySB(SB_{u_i}, SB_{AU})的基础上,可以得到用户 mu_i 与活跃用户集 AU 的总体差异度,如式(18)所示。

$$\begin{aligned} diversity(bui, AU) = & \\ & \lambda_1 \times diversityProfile(Profile_{mu_i}, Profile_{AU}) \\ & + \lambda_2 \times diversitySR(SR_{mu_i}, SR_{AU}) \\ & + \lambda_3 \times diversityIQ(IQ_{mu_i}, IQ_{AU}) \\ & + \lambda_4 \times diversitySB(SB_{mu_i}, SB_{AU}) \end{aligned} \tag{18}$$

式(18)中,λ_1、λ_2、λ_3 和 λ_4 用于调节用户背景、社交关系、发表信息质量及社交行为在计算差异度时的权重,$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$。如果 $\lambda_1 = 0$,$\lambda_2 = 0$,$\lambda_4 = 0$ 就转化为完全依靠用户发表的信息质量计算用户差异度。

在这四类指标里,发表信息质量及社交行为是用户直接在社交媒体上从事的活动,体现活跃性的作用更大,而用户背景、社交关系说明用户有可能在社交媒体上从事更多的活动,体现一定的活跃性。因此,四类指标的权重在进行实验时,进行了有指导性的设置。在 3.5 节对 λ_1、λ_2、λ_3 和 λ_4 的取值进行了实验比较。

3 实验及分析

3.1 实验数据

目前,还没有用于微博用户活跃度判定的公开

语料。本文选取了新浪微博进行实验的统计与分析。截止 2015 年 9 月，新浪微博用户已经达到 2.12 亿，在微博网络上发表信息已经成为用户日常的网络生活中的重要组成部分。

本文选取了学术研究、企业管理、教育、文化、军事五个领域进行实验数据的采集与分析。在新浪微博搜索框中输入领域关键字进行检索，然后单击“找人”按钮，选取了“个人认证”及“普通用户”两类用户，使用 HtmlUnit 进行采集。五个领域获取的认证及普通用户情况见表 1 所示，共计 8 188 个用户。

表 1　实验选用的五个领域①

序号	领域	关键字	认证及普通用户数
1	学术研究	信息检索	490
2	企业管理	互联网高管	45
3	教育	幼儿教育	6 049
4	文化	谍战	876
5	军事	歼 20	728

在这 8 188 个用户中，少数的用户是微博“大 V”，一个用户就有大量的粉丝，这类用户不具有普遍的代表性，所以删除掉粉丝数大于 1 000 的微博“大 V”用户，最终剩下 8 063 个用户。

对五个领域 8 063 个用户，进一步获取他们的背景、关注数、粉丝数、原创微博、评论内容、社交行为等信息。对原创微博、评论内容的采集时间限定在 2015 年 1 月 1 日至 2015 年 5 月 28 日，共计 5 个月。

在 8 063 个用户中，发表信息量小于等于 5 个的选取了 300 个，发表信息量大于 30 个的选取了 300 个，发表信息量大于 5 个且小于 30 个的选取了 300 个，共计选用 900 个用户参与实验分析。

3.2　四种实验方法

目前有关用户活跃性判定的研究文献较少，我们选用了四种方法进行实验的对比分析。

方法一　对于发表信息量小于 5 的用户直接判定为冷启动用户，其他的用户都判定为活跃用户，类似于文献[10-11,17,19]阐述的内容，简记为 InforNum5；

方法二　对于发表信息量小于 20 的用户直接判定为冷启动用户，其他的用户都判定为活跃用户，类似于文献[12]的论述，简记为 InforNum20；

方法三　本文提出的方法，对于发表信息质量小于阈值 $\alpha(\alpha = 5)$ 的用户直接判定为冷启动用户，其他的用户都判定为活跃用户，简记为 InforQuality5；

方法四　本文提出的方法，对于发表信息质量小于阈值 $\alpha(\alpha = 5)$ 的用户直接判定为冷启动用户，大于阈值 $\beta(\beta = 30)$ 的用户直接判定为活跃用户，其他的都作为边缘用户，进而使用边缘用户与冷启动用户集、活跃用户集的差异度判定方法，简记为 UA4Index，参照 3.5 节的实验结果，四类指标的权重设置为 $\lambda_1 = 0.1$、$\lambda_2 = 0.2$、$\lambda_3 = 0.4$、$\lambda_4 = 0.3$。

3.3　评价指标

使用准确率 P、召回率 R 及 F 值对实验结果进行评价，三类评价指标分别介绍如下。

(1) 准确率 $p = C/N$，N 是使用某种方法判定出的活跃或者是冷启动用户的个数，C 是 N 个用户中判定结果正确的个数(需参考标准答案)。比如有 600 个用户，某种方法判定出 120 个活跃用户，480 个冷启动用户，其中，120 个活跃用户中判定正确的是 60 个，480 个冷启动用户中判定正确的是 360 个，则 $p = \frac{60+360}{120+480} = \frac{420}{600} = 0.7$，准确率为 70%。

(2) 召回率 $R = C/A$，A 是标准答案中活跃或者冷启动用户的个数，C 是判定结果正确的个数。比如，600 个用户中，标准答案有 120 个是活跃用户，其他的 480 个是冷启动用户，某种方法识别出的真正是活跃用户的是 80 个，冷启动用户是 400 个，则 $R = \frac{80+400}{120+480} = \frac{480}{600} = 0.8$，召回率为 80%。

(3) $F = 2 \times P \times R/(P + R)$，则 $F = \frac{2\times 0.7\times 0.8}{0.7+0.8} = \frac{1.12}{1.5} = 0.75$，F 值为 75%。

本文对实验的标准答案的确定方法如下：

(1) 对于发表信息小于等于 2 的 263 个用户，直接作为冷启动用户；

(2) 对于发表信息大于等于 30 的 300 个用户，作为活跃用户；

(3) 对于发表信息在 3～29 之间的 337 个边缘用户，使用 Pooling 技术确定标准答案，具体步骤如下：

① 选取三名与本研究工作相关的研究生，使其理解微博网络上活跃用户与冷启动用户的基本概念；

① 2015 年 5 月 28 日执行完采集。

② 为三名研究生提供337个用户的背景(微博等级、标签)、社交关系(粉丝数、关注数)、发表的信息(原创微博、评论)、社交行为(点赞),每名研究生根据自己的理解,将337个用户划分到两个集合,一个是活跃用户集AU,另一个是冷启动用户集IAU;

③ 获得三名研究生关于活跃用户集及冷启动用户集的并集,得到两个Pool;

④ 再由本文的两位作者,对同一用户划分到不同Pool的情况进行人工区分,即每个用户最终只能划分到一个类别中。

经过Pooling过程后,337个边缘用户中有126个是活跃用户,211个是冷启动用户。最终900个用户中活跃用户集AU=474,冷启动用户集IAU=426。

3.4 四种方法判定用户活跃度的效果比较

使用3.2节介绍的四种实验方法,3.3节介绍的评价指标,对900个用户,得到的用户活跃性判定结果如表2所示。

表2 四种方法得到的用户活跃性判定结果

方法	P	R	F
InforNum5	0.71	0.65	0.68
InforNum20	0.56	0.60	0.58
InforQuality5	0.77	0.74	0.75
UA4Index	0.92	0.78	0.84

从表2可见,四种方法中,本文提出的方法UA4Index得到的效果最为理想,F值达到0.84,说明在微博类社交媒体上综合地考虑各类指标比单纯的使用用户发表信息的数量判定用户的活跃性更为有效。方法InforNum20得到的F值为0.58,效果最差,主要原因是把冷启动判定的阈值设置过大,导致本来属于边缘的用户武断地判定为了冷启动用户。方法InforNum5得到的结果为0.68,方法InforQuality5得到的结果为0.75,说明了直接使用发表信息的质量比使用用户发表信息的数量效果来得更好,提高了7%。

3.5 四类指标权重取值对判定结果的影响

在这四类度量指标里,发表信息质量及社交行为是用户直接在社交媒体上从事的活动,体现活跃性的作用更大,而用户背景、社交关系说明用户有可能在社交媒体上从事更多的活动,体现更大的活跃性。因此,四类度量指标的权重系数在进行实验时,进行了有指导性的设置。如果没有任何指导,四类指标,每类指标的变化范围即使取{0,0.1,0.2,…,1.0}(步长为0.1)共11种情况,四类指标共需进行$11^4=14\ 641$次实验。

我们设置的四类指标取值范围分别为$\lambda_1=[0,0.3]$、$\lambda_2=[0,0.3]$、$\lambda_3=[0.3,0.6]$、$\lambda_4=[0.3,0.6]$(步长分别为0.1),共进行了$4^4=256$次实验。

表3列出了评价指标F值得分比较高的12条数据对应的四类指标的权重取值情况。

表3 12条数据对应的四类指标的权重取值

λ_1的取值(背景)	λ_2的取值(社交关系)	λ_3的取值(信息质量)	λ_4的取值(社交行为)	F值
0	0.2	0.4	0.4	0.81
0	0.2	0.5	0.3	0.81
0	0.1	0.5	0.4	0.81
0.1	0	0.5	0.4	0.82
0.1	0	0.4	0.5	0.82
0.1	**0.1**	**0.4**	**0.4**	**0.83**
0.1	0.1	0.3	0.5	0.82
0.1	**0.1**	**0.5**	**0.3**	**0.83**
0.1	0.1	0.6	0.2	0.82
0.1	**0.2**	**0.4**	**0.3**	**0.84**
0.1	0.2	0.3	0.4	0.82
0.1	0.2	0.5	0.2	0.81

对四类指标不同的变化组合的256次实验中,F值最高为0.84,可以认为F值偏差在0.1范围的都是非常合理的。因此,关于四类指标的权重给出三组参数的取值建议,第一组:$\lambda_1=0.1$、$\lambda_2=0.1$、$\lambda_3=0.4$、$\lambda_4=0.4$;第二组:$\lambda_1=0.1$、$\lambda_2=0.1$、$\lambda_3=0.5$、$\lambda_4=0.3$;第三组:$\lambda_1=0.1$、$\lambda_2=0.2$、$\lambda_3=0.4$、$\lambda_4=0.3$。当然,拓宽F值的偏差范围,四类指标的权重组合形式更多一些。

3.6 用户活跃性判定对用户推荐的影响

为了进一步检验本文所提用户活跃性判定方法的有效性,以微博网络用户推荐为出发点进行实验分析。

微博网络上用户之间的推荐受到多个因素的影响,我们前期的研究成果综合地考虑了用户发表内容相似性、交互相关性、社交关系(粉丝和关注)相关

性等指标，提出了新颖的相似用户计算方法，具体方法详见文献[3]，该方法记为 URNoActive。在此基础上，我们将用户活跃性作为其中的指标引入，即只有当用户是活跃的，系统才会进一步计算用户的相似度并进行推荐，该方法记为 URByActive。因为，用户在微博网络上构建关注关系，本意是建立社交圈子、分享生活体验和增进社交能力。对于冷启动用户而言，难以达到在微博网络上社交目的，在推荐时价值不大。

在数据集的使用上，3.1 节介绍的数据集是围绕学术研究、企业管理、教育、文化、军事五个领域采集的，采集的用户之间已经有明显的领域相关性。对参与实验的 900 个用户，我们进一步地采集了这些用户发表的信息、背景、粉丝、关注等信息。

评价指标同样使用准确率 P、召回率 R 及 F 值。假设用户 u_i 在微博网络上真实的关注集为 u_i_Follower，使用某种方法推荐给用户的关注集为 u_i_RFollower，要求 $|u_i\_\text{RFollower}| = |u_i\_\text{Follower}|$，则 $P=\frac{|u_i\_\text{RFollower}\cap u_i\_\text{Follower}|}{|u_i\_\text{Follower}|}$，$R$ 和 F 值指标计算方法和 3.3 节介绍的相同。

表 4 列出了使用两种方法（URNoActive 和 URByActive）得到的关注用户推荐结果。

表 4　两种方法得到的用户推荐结果

方法	P	R	F
URNoActive	0.64	0.55	0.59
URByActive	0.69	0.57	0.62

由表 4 可见，添加了用户活跃性判定再进行用户相似度计算和推荐的方法 URByActive，可以在一定程度上改善用户推荐的效果，P、R 和 F 值分别提高了 0.05、0.02 和 0.03。主要原因是在选取推荐用户时，考虑到用户在微博网络上发表信息、构建社交圈子的情况，这些行为暗示了被推荐用户可能是活跃的。但同时，我们也发现，两种方法在推荐效果上都不是非常好，主要原因是微博网络用户众多，即使在话题兴趣、社交圈子上有较高的相似性，但彼此之间并不知道，并没有构建关注关系，这也是微博网络上研究用户推荐的原因之一。

4　总结

本文针对已有方法简单地根据用户发表信息的数量判定用户是否活跃的粗糙方式，在综合的分析影响微博网络用户活跃性的背景、社交关系、发表信息质量及社交行为等各类指标的基础上，提出了系统的用户活跃性判定流程、全面的用户活跃性的差异度计算模型，并选取了主流的新浪微博进行了实验分析比较。

本文主要是面向微博网络进行了所提方法的理论分析与实验比较，在实际应用中，针对不同的社交媒体，需考虑不同媒体的差异，对四类指标可以进行压缩或者扩展，从而为不同的社交媒体给出合适的用户活跃性计算模型。

文献[3]在进行微博相似用户推荐时，没有考虑用户的活跃性，将本文提出的活跃用户判定方法应用于用户推荐的前期，即先判定用户的活跃性、选取活跃用户，然后再计算用户的相似度进行推荐，明显地改善了推荐的效果。文献[22]在挖掘用户兴趣、计算用户兴趣相似度时，根据社交媒体的特点，将用户兴趣区分为基于背景信息的静态兴趣和基于生成内容的动态兴趣，更合理地揭示了用户的兴趣特征。其提出的用户背景信息中的标签对于判定用户的活跃性有一定的参考价值，本文将用户标签的丰富程度作为判定用户活跃性的指标之一。

对于该问题的研究，我们认为如下内容还需进一步加深：(1)影响用户活跃性的指标，除了本文提出的四类指标外，是否还有其他指标能够体现用户的活跃性，从而更加全面地评价用户的活跃性；(2)将本文所提方法计算出的用户活跃性，与实际的应用场景结合，在实践中进一步检验所提方法的有效性，比如应用到多个社交媒体的推荐系统中。

致谢　感谢江苏金鸽网络科技有限公司为本研究提供的实验数据集。

参考文献

[1] 高明，金澈清，钱卫宁，等. 面向微博系统的实时个性化推荐[J]. 计算机学报，2014，37(4)：963-975.

[2] Chen K L, Chen T Q, Zheng G Q, et al. Collaborative personalized tweet recommendation[C]//Proceeding of the 35th International ACM SIGIR Conference on Research and Development Information Retrieval. Portland, OR, USA, 2012: 661-670.

[3] 仲兆满，胡云，李存华，等. 微博中特定用户的相似用户发现方法[J]. 计算机学报，2016，39(4)：765-779.

[4] 徐志明，李栋，刘挺，等. 微博用户的相似性度量及其应用[J]. 计算机学报，2014，37(1)：207-218.

[5] 彭泽环,孙乐,韩先培,等. 基于排序学习的微博用户推荐[J]. 中文信息学报, 2013, 27(4): 96-102.
[6] Zeng W, Zeng A, Liu H, et al. Uncovering the information core in recommender systems[J]. Scientific Reports, 2014(4): 6140.
[7] 汪祥,贾焰,周斌,陈儒华,韩毅. 基于交互关系的微博用户标签预测[J]. 计算机工程与科学, 2013, 35(10): 44-50.
[8] Akcora C G, Carminati B, Ferrari E. User similarities on social networks[J]. Social Network Analysis and Mining, 2013(3): 475-495.
[9] Lin J, Sugiyaman K, Kan M Y, et al. Addressing cold-start in app recommendation: Latent user models constructed from Twitter followers[C]//Proceedings of the SIGIR'13, 2013: 283-293.
[10] Massa P, Avesani P. Trust-aware recommender systems[C]//Proceedings of the 2007 ACM Conference on Recommender Systems, 2007: 17-24.
[11] Guo G B, Zhang J, Thalmann D. Merging trust in collaborative filtering to alleviate data sparsity and cold start[J]. Knowledge-Based Systems, 2014(57): 57-68.
[12] Pereira A L V, Hruschka E R. Simultaneous co-clustering and learning to address the cold start problem in recommender systems[J]. Knowledge-Based Systems, 2015(82): 11-19.
[13] 古万荣,董守斌,曾之肇,等. 基于微博用户模型的个性化新闻推荐. 中文信息学报, 2016, 30(1): 93-100.
[14] 王占,林岩. 基于信任与用户兴趣变化的协同过滤方法研究[J]. 情报学报, 2017, 36(2): 197-205.
[15] Meyffret S, Medini L, Laforest F. Trust-based local and social recommendation[C]//Proceedings of the 2012 ACM Conference on Recommender Systems. Dublin, Ireland, ACM, 2012: 53-60.
[16] Yuan W W, Yang X W, Steck H, et al. Circle-based recommendation in online social networks[C]//Proceedings of the 2012 ACM Conference on Recommender Systems. Beijing, China, ACM, 2012: 1267-1275.
[17] Ocepek U, Rugelj J, Bosnic Z. Improving matrix factorization recommendations for examples in cold start [J]. Expert Systems with Applications, 2015(42): 6784-6794.
[18] 于洪,李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报,2015,26(6):1395-1408.
[19] 杨圩生,罗爱民,张萌萌. 基于信任环的用户冷启动推荐[J]. 计算机科学, 2013, 40(11a): 363-366.
[20] Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases[J]. Proceedings of the National Academy of Sciences, 2009, 106(51): 21484-21489.
[21] Liang C, Liu Z Y, Sun M S. Expert finding for Microblog misinformation identification[C]//Proceedings of the 24th ACL International Conference on Computational Linguistics. Mumbai, 2012: 703-712.
[22] 仲兆满,管燕,胡云,等. 基于背景和内容的微博用户兴趣挖掘[J]. 软件学报, 2017, 28(2): 278-291.

仲兆满(1977—),博士研究生,副教授,主要研究领域为信息检索、人工智能和社交网络分析。
E-mail: zhongzhaoman@163.com

戴红伟(1975—),博士研究生,副教授,主要研究领域为人工智能。
E-mail: 3974523025@qq.com

管燕(1976—),硕士研究生,讲师,主要研究领域为模式识别与人工智能。
E-mail: gy764@sohu.com