ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2018/12(05)-0730-11
doi: 10.3778/j.issn.1673-9418.1709035

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

在线招聘场景下的简历活跃度预测*

史舒扬[1],张智鹏[1],郭 龙[1],邵鉴侠[1],崔 斌[2+]
1. 北京大学 信息科学技术学院 高可信软件技术教育部重点实验室,北京 100871
2. 北京大学 深圳研究生院,广东 深圳 518055

Resume Activeness Prediction in Online Recruitment Scenarios*

SHI Shuyang[1], ZHANG Zhipeng[1], GUO Long[1], SHAO Yingxia[1], CUI Bin[2+]
1. Key Lab of High Confidence Software Technologies, Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
2. Peking University Shenzhen Graduate School, Shenzhen, Guangdong 518055, China
+ Corresponding author: E-mail: bin.cui@pku.edu.cn

SHI Shuyang, ZHANG Zhipeng, GUO Long, et al. Resume activeness prediction in online recruitment scenarios. Journal of Frontiers of Computer Science and Technology, 2018, 12(5): 730-740.

Abstract: In the era of Internet, a lot of recruitments happen on online recruitment platforms. These platforms recommend jobs for applicants and meanwhile recommend resumes to corporations. However, it is almost impossible for the platforms to know whether the applicant has found a job. As a result, resumes are still being recommended to corporations even if the applicant has found a job, which leads to a waste of the platform resource as well as unsatisfactory user experience. This paper formalizes the resume activeness prediction problem in online recruitment scenarios, which aims to find highly active applicants so that the platform recommends the active ones and discards the inactive ones and therefore escalates user experience. Current solutions for user activity level predication are often restricted to certain scenarios like social networks, and they utilize scenario-aware features. Unfortunately, these features are not applicable in online recruitment scenarios. With a careful study of real-world recruitment data, this paper summarizes four characteristics of online recruitment scenarios, which are hyper-dynamism, low user viscosity, bidi-

* The National Natural Science Foundation of China under Grant Nos. 61702016, 61702015, 61572039 (国家自然科学基金); the Postdoctoral Science Foundation of China under Grant Nos. 2017M610019, 2017M610020 (中国博士后科学基金); the Science Research Project of Shenzhen under Grant No. CYJ20151014093505032 (深圳市科技计划项目).
Received 2017-09, Accepted 2017-11.
CNKI网络出版: 2017-11-30, http://kns.cnki.net/kcms/detail/11.5602.TP.20171130.1535.002.html

rectional matching and the priority of recall over precision in the predication result. Based on these characteristics, this paper proposes a model named RAP (resume activeness prediction), which carefully handles these characteristics and also provides a parameter γ to deal with the priority of recall. The extensive experiments on real-world data from 58 Recruitment Website demonstrate that the AUC of RAP can achieve 0.817.

Key words: user activeness prediction; online recruitment; classification; data analysis

摘　要: 在信息时代,在线招聘平台承担了大量的招聘任务,平台向求职者推荐合适的职位,并向招聘者推荐合适的简历。但是在推荐简历的时候,平台难以获知用户是否已找到工作,常会在求职成功以后继续推送,导致平台资源的浪费和用户体验的损失。基于这一情况,提出了在线招聘场景下的简历活跃度预测问题,旨在通过预测未来活跃度高的求职者,对其重点推送,从而应对这一问题。现有的活跃度预测方案,大都在社交场景下,结合社交网络的特点设计适应性的模型,但特点不同导致这些方案在招聘场景下并不适用。结合真实数据分析了在线招聘场景的数据特征,提出4个场景特点——高度动态性、用户黏度低、双向匹配、召回优先等。据此,有针对性地提出了招聘平台下的简历活跃度预测方法(resume activeness prediction, RAP)。RAP能适应上述前3项特点,并通过调节筛选参数γ满足召回优先。在58招聘真实数据的实验中,RAP模型的AUC达到了0.817。

关键词: 用户活跃度预测;在线招聘;分类;数据分析

文献标志码: A　　**中图分类号:** TP391

1　引言

在信息时代的背景下,大量招聘与求职的工作被放在互联网上进行,各种各样的在线招聘平台相继出现,如58招聘、智联招聘、拉勾网等。

在以58招聘为例的求职招聘网站上,每天有大量的简历被求职者更新后展示在个人资料里或者投递到对应职位上,也有大量的职位被各招聘单位放出。通过数据统计发现,在2016年9月10日开始的一个月间,就一共添加了超过165万份新简历,放出了超过280万个职位,这样的数据量暗示着平台的重要性与潜力。在线招聘平台中招聘者和求职者有两种互动模式:其一是求职者浏览公开招聘的职位,选择合适的职位投递简历,然后等待招聘者联系并安排面试;其二是招聘平台向招聘者推送可能适合于该单位的简历,然后招聘者从中下载相应求职者的具体联系信息。从平台的角度出发,这里主要关注后者,即求职者的简历被推送给招聘者这一方式。

在这个方式中,尽管招聘平台提供了将求职者推送给招聘者的途径,但是也存在着一定的问题。因为求职者找到工作之后通常不会告知平台,可能只是不再登录平台进行浏览、点击、投递等,所以平台并不知道他已经不需要继续求职,会继续推送该求职者的简历。但是由于该求职者其实已经有了满意的工作,此时推送的简历就成为了"无效"简历,即使被下载,也不会再达成新的劳动协议。这样的情况,一方面降低了求职者和招聘者双方的用户体验,另一方面降低了推送简历的价值,也造成了平台的资源浪费。如图1所示,在2016年的一个月中,除去仅活跃一天的用户以后,有点击、投递行为不超过9天的用户占了97%,并且活跃只有两三天的用户超过70%,这意味着大部分求职者的活跃时间都很短暂,他们在求职成功以后不再活跃,继续推送他们的简历将造成平台资源的浪费和用户体验的损失。因此对简历活跃与否的甄别是有意义的,针对活跃简历重点推送能够很大程度上提升推送效果。

在现实中,求职者是否找到工作的真实情况是难以推测的,因为用户找到工作之后通常不会告知平台。尽管如此,显而易见的是,持续活跃的用户仍然有在平台求职的意愿。基于这一观点,本文重点关注那些近期内活跃的求职者,用求职者在最近一

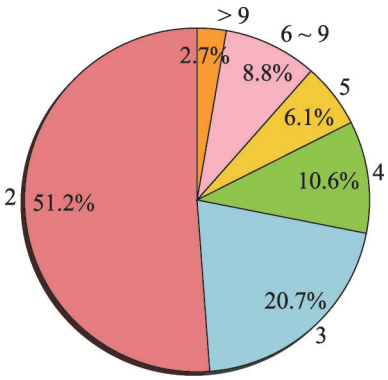


Fig.1 Ratio of users that click or deliver on different numbers of days

图1 不同天数发生过点击或者投递的用户所占比例

段时间内点击、浏览等行为定义简历活跃度的概念，旨在通过预测简历的活跃度，将未来近期内可能会活跃的简历筛选出来，在推送的时候赋予更高的优先级和更好的展示机会，提高平台简历推荐的效果，提升用户体验，减少平台的资源浪费。

现有的活跃度预测方案，大多是基于在线社交网络的方案，文献[1]采用生存时间预测场景下的Cox比例风险统计模型，在招聘场景下由于“生存时间”相对较短而难以保证其适用性，同时也会浪费场景信息资源。文献[2]从社交网络内外两方面进行考虑，也有工作从社交网络本身的特性出发，根据用户行为的多样性、动态性、社交影响等特点，基于现有模型提出了更具有适应性的预测模型[3-4]。这些模型适应了社交网络的特点，但是由于招聘场景与其的诸多不同，并不能很好地移植。例如在招聘场景下，用户的个体行为风格难以刻画，社交关系没有定义，多样性、社交影响就无从谈起。

为了做好简历的活跃度预测，本文对在线招聘的场景特点进行了探索，总结了以下4个特性：首先，数据具有高度动态性，每天有大量的求职者在平台上传并投递简历，也有大量的招聘职位被招聘者放出。其次，用户黏度低，求职者往往在短至几天的活跃之后就离开平台，在很长时间内不再有活跃行为。对用户连续行为的分析表明，在线招聘场景下用户往往不会表现出像社交网络那样长久的粘性，而是在一段时间内频繁产生登录、点击等行为，之后就会很长时间内不再活跃。第三，由于招聘平台的特殊性，其中的用户分为求职者（简历）一方和招聘者（职位）一方，前者可以点击、浏览、投递，而后者可以接受求职者投递的简历，也可以从招聘平台购买所需要的简历，这样的行为模式是达成雇佣关系的重要过程，也是数据中应该被考虑的内容。除此以外，招聘场景下活跃度预测的目标也有一定的特殊性，即召回优先（参见2.4节），因为与推送一份“无效”简历（求职结束的求职者）带来的资源浪费相比，损失一份活跃简历（仍在求职的求职者）会直接影响劳动协议的达成，有更严重的后果，所以在预测过程中需要保留足够多的有效简历。

在适应在线招聘场景数据特点的情况下，根据高度动态性、用户黏度低、双向匹配以及召回优先特点，本文结合随机森林模型（random forest，RF）和逻辑回归模型（logistic regression，LR），提出了简历活跃度预测模型（resume activeness prediction，RAP），并提供了可供平台选择的筛选参数 γ，用来调节召回率和准确率的相对重要程度。

本文的主要贡献如下：

（1）提出在线招聘平台中的简历活跃度预测问题，并基于真实数据分析，总结了在线招聘场景下数据具有高度动态性、用户黏度低、双向匹配等特点，明确了场景要求下预测问题的召回优先性。

（2）结合招聘场景下高度动态性、用户黏度低、双向匹配的特点，采用树模型和线性模型的混合模型，提出了简历活跃度预测方法RAP，并通过筛选参数 γ 来调节两类预测错误的损失函数比，从而适应在线招聘场景下召回优先的需求。

（3）在58招聘的真实数据上的实验表明，RAP可以获得0.817的AUC（area under curve）值，并且通过筛选参数 γ 能够有效地实现召回优先这个需求。

本文组织结构如下：第2章给出在线招聘场景下简历活跃度预测的问题定义；第3章总结在线招聘场景的特点；第4章根据总结的特点提出预测方法RAP；第5章在58招聘的真实数据上进行实验，证明RAP方法的有效性；第6章介绍目前已有的相关工作和研究；最后对本文结果进行总结。

2 问题定义

本章首先定义简历活跃度的概念，在此基础上，给出简历活跃度预测问题的定义，作为进一步研究简历活跃度预测问题的基础。

定义1（简历活跃度）若简历 r 在第 t 天被投递了 D_r^t 次，其用户在该天一共点击了 C_r^t 次招聘职位，则 r 在这一天的活跃情况为 (C_r^t, D_r^t)。为了方便进行比较等运算，简历 r 在第 t 天的活跃度 A_r^t 定义为：

$$A_r^t = \begin{cases} 1, & C_r^t \geqslant \theta_c \text{ or } D_r^t \geqslant \theta_d \\ 0, & \text{otherwise} \end{cases}$$

其中，θ_c、θ_d 是两个阈值参数。进一步地，简历 r 在第 t 天时，将未来一段时间（δ 天）内的活跃度定义为：

$$y_r^t = \begin{cases} 0, & A_r^{t+1} = A_r^{t+2} = \cdots = A_r^{t+\delta} = 0 \\ 1, & \text{otherwise} \end{cases}$$

即，如果简历 r 在未来 δ 天内至少有一天活跃，那么它在未来 δ 天内就被认为是活跃的。

在此基础上，将简历活跃度预测问题定义为一个传统的分类问题。

定义2（简历活跃度预测）给定简历集合 R 和每个简历的特征信息，给定未来区间 δ，对其中的每一份简历 $r \in R$，简历活跃度预测问题就是利用特征信息预测简历 r 在第 t 天时未来 δ 天的活跃度 $y_r^t \in \{0,1\}$，记预测值为 $\hat{y}_r^t \in \{0,1\}$，其中1代表活跃，0代表不活跃。

3 在线招聘场景特点

为了更好地了解在线招聘场景，更有针对性地设计模型来解决简历活跃度预测的问题，本章分析在线招聘场景的主要特点。通过对58招聘平台的数据分析与探究，对比社交网络等其他在线网络平台，本文总结了数据与用户行为的特性，包括高度动态性、用户黏度低、双向匹配等特点。同时，从场景动机出发，在线招聘对简历活跃度的预测也有召回优先的要求。

3.1 高度动态性

在招聘平台上，每天有大量的求职者在平台上传并投递简历，也有大量的招聘职位被招聘者放出。如图2所示，对58招聘平台的数据统计表明，每天大约产生数万的新简历和超过十万的新职位。这样的数据量展示了招聘场景数据的高度动态性，新用户、新简历、新职位的产生速度非常快，信息快速更迭。在活跃度预测问题中，高度动态性表示人们不仅需要处理大量的数据，还需要应对大量新简历的情况。

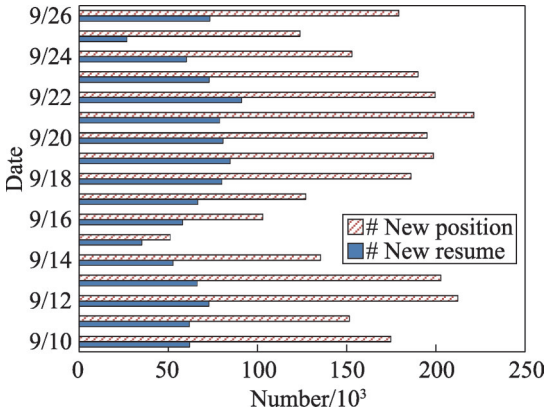


Fig.2 Number of new resumes and positions published every day on 58 Recruitment Website in Sep, 2016

图2 58招聘每天添加的新简历、新职位数目（2016年9月）

3.2 用户黏度低

求职者在求职期间会经常登录求职网站，浏览各种职位，并投递一些职位，而一旦找到了满意的工作，就不会继续登录网站产生上述行为。从而招聘平台往往有较低的用户黏度，尤其是与社交网络用户的强黏度相比。在这样的情况下，数据具有强时效性，判断用户的行为可能需要从近期的活跃度和行为情况入手，久远的信息可能只涉及完全不相关的工作状态。

本文从58招聘平台的真实数据中找到了相应的依据。如图1所示，根据用户在一个月内产生点击或者投递行为的天数分类统计，产生过点击或者投递行为的用户中，3天以下的占71.2%，5天以下的占87.9%，大于9天的只占2.7%，因此单个用户的行为数据量是非常有限的，经常活跃几天以后就不再产生用户行为。除了宏观统计之外，将这段时间内有活跃行为的用户取样，对他们的活跃行为次数按照日期的分布绘制了统计图。在图3的个案分析中，可以看到，比较典型的用户，例如用户B，在一段比较短的时间内（8天）活跃，在接下来的一段时间内不再登录平台产生行为；用户A在两个活跃峰值之间有约一

周的间隔,但是之后也不再继续活跃,行为集中;用户C活跃的日期范围相对分散,横跨18天,但只活跃了6天。在对数据的观察中发现这3类用户都存在,其中与用户A、B类似的用户尤其广泛,大部分用户都在一个比较短暂的区间内产生活跃行为。因此,用户黏度低,数据的时效性非常强,采用近段时间的数据进行预测是非常合理的选择。

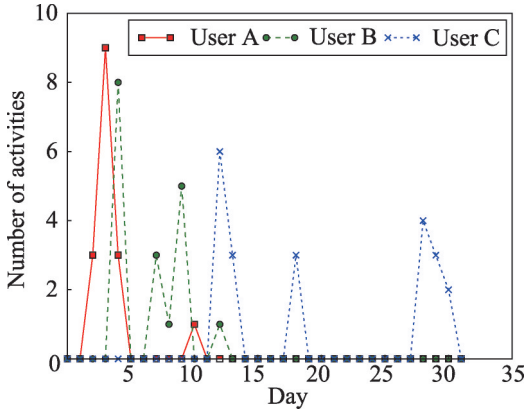


Fig.3 Statistics of user activities of several sample users in a month on 58 Recruitment website

图3 一个月内58招聘用户活跃次数按天统计示例

3.3 双向匹配

在现实职业场景的招聘会中,有招聘者和求职者双方的参与,线上招聘平台也是如此,双方都有对应的活动,分别发布简历和职位,然后求职者可以浏览职位,投递简历,而招聘者可以购买或下载简历。由于这是招聘场景的本质特征,这样的双向互动在所有线上招聘平台中都是存在且类似的。

在简历活跃度预测的问题中,很重要的内容是求职者能否通过平台找到工作,而能达到这一点的最主要的就是简历和职位在浏览、投递、购买等场景的交互中能否匹配得当。简单来说,求职者认为匹配得当就会投递简历,而招聘者觉得匹配得当就会购买简历或者约求职者面试。因此,尽管没有显式的公式进行衡量,这种双向匹配的活动是人们在判断的时候需要关注的重要内容。

3.4 活跃度预测的召回优先

在线招聘场景下简历活跃度预测问题,是为了挑选出活跃简历进行重点推送,因此准确率和召回率[5]的重要性有明显的差异。当一份不会再活跃的简历被当作活跃简历重点推送时,最坏的结果是浪费少量的平台资源,降低一些求职者和招聘者的用户体验;但是另一方面,如果一份活跃简历被当作不会再活跃的简历,基本不被推送,则可能会导致一个正在求职的用户无法获得对应的招聘,妨碍劳动协议的达成,直接影响平台核心功能,进而影响招聘平台对就业市场的贡献。两相比较,召回率比准确率在此场景下更为重要一些,即活跃度预测具有召回优先的特性,预测方法也需要与这样的指标相适应。同时,召回率相对准确率优先的程度需要根据具体平台的实际情况和需求来调节,理想的预测方法应该能满足这一要求。

3.5 小结

根据以上分析,在线招聘场景具有高度动态性、用户黏度低、双向匹配、召回优先等特点,根据这些特点,在预测简历活跃度中,本文选取两方面的特征:近期历史活跃信息(对应高度动态性、低用户黏度)和双向匹配信息(对应双向匹配特点),而召回优先的需求将从模型的角度控制。具体的特征参见5.2节。

4 活跃度预测方法RAP

根据第2章总结的在线招聘的场景特点,即高度动态性、用户黏度低、双向匹配和召回优先,本章提出对应的简历活跃度预测方法RAP。

如图4所示,RAP模型的执行流程分为特征转换和模型训练两部分。第一部分是特征转换部分,RAP通过树模型将数据特征 x_r^t 中的近期历史活跃信息部分转化为新的向量特征 z_r^t;第二部分用转换后的特征和双向匹配特征拼接,然后训练模型RAP。

接下来具体介绍模型的两部分,即如何对随机森林特征进行转化,以及如何定义RAP模型。

4.1 特征转换

通过树模型对特征进行一定的选择和过滤,能够选取一些比较重要的特征,使得模型具有一定程度的可解释性。在线招聘场景中,同样要求用于活跃度预测模型的特征具有一定的可解释性,因为这样有利于用户和平台结合自己的策略进行特征的调

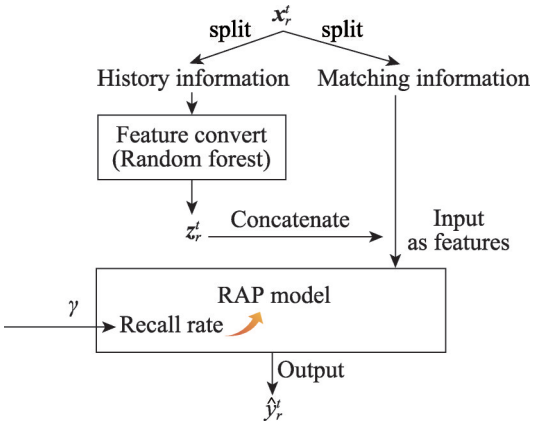


Fig.4 A schematic diagram of RAP model

图4 RAP模型的结构示意图

控[6]。因此，本文采用能适应较大数据量的树模型随机森林对近期历史活跃信息进行特征的转换。

利用随机森林进行特征转换[6]指的是：对输入的特征（近期历史活跃信息），通过随机森林模型进行分类训练，然后将现有特征 $\boldsymbol{x}_r^t$ 转化为只包含0、1的新特征 $\boldsymbol{z}(\boldsymbol{x}_r^t)=\boldsymbol{z}_r^t$，具体的流程如算法1所示。在用随机森林训练得到树模型之后，新特征 $\boldsymbol{z}_r^t$ 的维数就是所有树的叶子结点个数（第2~3行），而对于特征 $\boldsymbol{x}_r^t$，其在 $\boldsymbol{z}(\boldsymbol{x}_r^t)$ 中第 k 维的值由 $\boldsymbol{x}_r^t$ 在树模型上分类的结果决定，第 k 维为1当且仅当分类后落在第 k 个叶子结点上。计算的时候，遍历森林中的所有树（第4~15行），找到特征对应的叶子结点（第5~12行），并标记新特征的对应值（第13~14行）。例如，图5中的 $\boldsymbol{x}_r^t$ 经过树模型之后，假设分类经过的结点为图中的画“X”结点，并且一共有两棵树，5个叶子结点，那么得到的新特征为 $\boldsymbol{z}_r^t=(0,0,1,1,0)^{\mathrm{T}}$。

算法1 特征转换

输入：训练得到的树模型 *trees*，输入的特征向量 $\boldsymbol{x}$。

输出：转换后的特征 $\boldsymbol{z}$。

1. *leaves* ← 包含 *trees* 中所有叶结点的数组；
2. $\boldsymbol{z}$ ← 长度与 *leaves* 相同且全为0的数组；
3. for $T\in trees$ do
4. $node \leftarrow T.root$；
5. while not *node.isLeaf* do
6. if x 属于 *node* 的左子树 then
7. $node \leftarrow node.leftChild$
8. else
9. $node \leftarrow node.rightChild$
10. end
11. end
12. 找到 k 使得 $leaves[k]=node$
13. $z[k] \leftarrow 1$
14. end
15. return $\boldsymbol{z}$

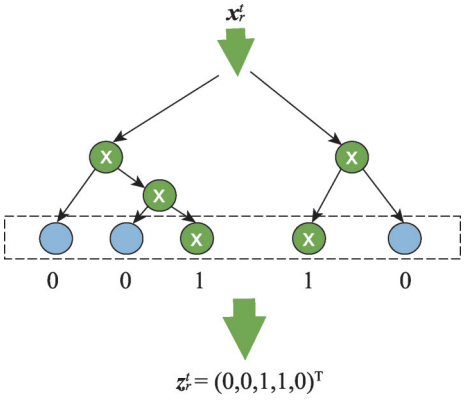


Fig.5 An example of feature conversion

图5 特征转换样例

4.2 模型定义

如第1章所述，将活跃度预测问题形式化为一个二分类问题。这里基于逻辑回归模型提出RAP模型。RAP定义了全新的损失函数，能够很好地适应在线招聘场景中召回优先的需求。本文以逻辑回归（LR）为基本模型的原因是：LR的输出结果表征求职者是否活跃的概率，而不是简单的0、1类别。在在线招聘场景中，以概率作为输出值是很重要的，因为平台可以根据各个简历的活跃情况进行排序，进而获得相对活跃的简历。接下来，介绍活跃度预测模型RAP。

定义3（RAP预测模型）假设需要在第 t 天预测简历 r 在未来 δ 天内是否活跃，用 $\boldsymbol{z}_r^t$ 表示数据特征，$\hat{y}_r^t$ 表示预测值，则预测值为：

$$\hat{y}_r^t=\sigma(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t}}$$

其中，σ 为sigmoid函数；$\boldsymbol{\omega}$ 为需要学习得到的回归系数。

对于每一天、每一份简历，最小化它们预测值和真实值之间的误差，因此优化目标 J 定义为：

$$\min_{\omega} J=\sum_{r\in R}\sum_{t\leqslant T} l(y_r^t,\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t)+\gamma_0\left\|\boldsymbol{\omega}\right\|_2^2$$

其中，γ_0 是正则化项 $\left\|\boldsymbol{\omega}\right\|_2^2$ 的参数；T 是训练数据的天数 $l(y_r^t,\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t)$ 是损失函数。损失函数定义为：

$$l(y_r^t,\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t)=\ln(1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t})+(1-y_r^t)(1-\gamma)\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{z}_r^t$$

其中，$\gamma\in[0,1]$ 为筛选参数。这个预测模型记为RAP模型。通过简单的求导可知，该模型的损失函数是凸函数，因此能够使用梯度下降函数求解参数并获得全局最优解。

下面证明RAP中召回率和准确率的相对重要程度由筛选参数 γ 决定。考虑分类过程中出现的两类错误，FP（false positive，预测值为1而实际值为0）与FN（false negative，预测值为0而实际值为1），要判断召回率和准确率的相对重要程度，只需判断两者对损失函数的贡献的比值即可。

命题1 如果分别记 FP、FN 两类错误对目标函数 J 的贡献（即损失）为 $L(FP)$ 和 $L(FN)$，则两者之比为 $1-\gamma$。

证明 直接计算：

$$\frac{L(FP)}{L(FN)}=\frac{\lim\limits_{y=0,\boldsymbol{\omega}^{\mathrm{T}}z\to+\infty} l(\gamma,\boldsymbol{\omega}^{\mathrm{T}}z)}{\lim\limits_{y=1,\boldsymbol{\omega}^{\mathrm{T}}z\to-\infty} l(\gamma,\boldsymbol{\omega}^{\mathrm{T}}z)}=$$

$$\frac{\lim\limits_{\boldsymbol{\omega}^{\mathrm{T}}z\to+\infty}\ln(1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}z})+(1-\gamma)\boldsymbol{\omega}^{\mathrm{T}}z}{\lim\limits_{\boldsymbol{\omega}^{\mathrm{T}}z\to-\infty}\ln(1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}z})}=$$

$$\lim_{\boldsymbol{\omega}^{\mathrm{T}}z\to+\infty}\frac{\ln(1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}z})+(1-\gamma)\boldsymbol{\omega}^{\mathrm{T}}z}{\ln(1+\mathrm{e}^{\boldsymbol{\omega}^{\mathrm{T}}z})}=$$

$$\lim_{\boldsymbol{\omega}^{\mathrm{T}}z\to+\infty}\frac{(\ln(1+\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}z})+(1-\gamma)\boldsymbol{\omega}^{\mathrm{T}}z)'}{(\ln(1+\mathrm{e}^{\boldsymbol{\omega}^{\mathrm{T}}z}))'}=$$

$$\lim_{\boldsymbol{\omega}^{\mathrm{T}}z\to+\infty}1-\gamma-\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}z}=1-\gamma\in[0,1]\qquad\square$$

可以看到，筛选参数 γ 的实际意义就是两类错误对损失函数的贡献相差的百分比。因此，RAP可以方便简单地通过控制参数 γ 达到召回优先的需求。

5 实验

本章通过58招聘平台的真实数据，分别探究、展示近期历史活跃信息、双向匹配信息在简历活跃度预测问题中的利用情况，最后结合这两类特征，证明RAP模型在在线招聘场景中预测简历活跃度的有效性。

5.1 实验环境

本文实验运行在一个包含7台机器的Spark 1.6.1集群上，每个机器配有两个AMD Opteron 4180处理器，40 GB内存。实验代码均使用Scala语言编写。本文采用的数据来自58招聘平台（2016年9月10日至2016年10月10日），包括用户、简历、职位、企业等项目的基本信息，以及他们的行为数据，数据的特点参见表1。在实验中，不失一般性，采用前26天的数据作为训练集，后5天的数据作为测试集（其中简历用户都是近期 L 天内有过点击或者投递行为的）。

Table 1 Basic characteristics of 58 Recruitment data

表1 58招聘数据的基本特点

Property	Value
Number of resumes	1 513 066
Average number of resumes that have activities in last 5 days	282 643
Number of positions	2 804 917
Daily positive sample quantity	17 241
Daily negative sample quantity	161 108
Length of time interval/day	31

相关社交网络的活跃度模型[1-4]大多从社交网络本身的特性出发，根据用户行为的多样性、动态性、社交影响等特点预测，但是由于招聘场景与其的诸多不同，例如在招聘场景下，用户的个体行为风格难以刻画，社交关系没有定义等，已有的活跃度预测方法难以移植。因此，本文只展示RAP模型在实验数据中的有效性。

对于RAP模型效果的衡量指标，本文采用分类模型中通常使用的AUC（area under curve）、准确率（Precision）和召回率（Recall）。其中AUC是在ROC曲线下的面积[7]，用来衡量正样本预测结果比 负样本高的概率，而Precision和Recall的定义参见文献[5]。

本文实验中所提及的AUC、Precision、Recall是测试集中每天结果的平均值。

5.2 特征选择

根据第2章分析的在线招聘场景的特点，抽取近期历史活跃信息和双向匹配信息两方面的特征。

首先，根据在线招聘场景数据的高度动态性和

用户黏度低两个特点，抽取用户最近 L 天的近期历史活跃信息作为一部分特征。例如 $L=5$ 时，用户 A 在最近 L 天的活跃情况为（点击 5 次投递 3 次，点击 2 次投递 1 次，不点击也不投递，点击 6 次，不点击也不投递），那么对应的向量为 $(5,3,2,1,0,0,6,0,0,0)^{\mathrm{T}}$。

其次，根据双向匹配特性，分别抽取简历和职位两方面的信息以及二者的双向匹配信息，比如求职者用户点击/浏览职位，招聘者用户下载简历等行为。表 2 详细列出了这些信息的具体内容。

Table 2 Features extracted from 58 Recruitment data

表 2 58 招聘数据中提取的特征具体内容

Category	Feature
Daily activity	Click count
	Delivery count
	Download count
Recent activity	Recent clicks
	Recent deliveries
	Recent downloads
Resume-job match/ difference in delivery activities	Target salary-job salary
	Target salary-current salary
	Resume gender-job gender
	Resume complete level
	Resume position-job position
Resume-job match/ difference in download activities	Target salary-job salary
	Target salary-current salary
	Resume gender-job gender
	Resume complete level
	Resume position-job position

注：表中部分内容（recent activity，resume-job match/difference in delivery and download activities）采用的是平均值、标准差/方差、最大值等统计量。例如，recent click 包含求职者最近几天点击次数的平均值、标准差和最大值等。

5.3 基本参数设置

本节考虑实验中用到的基本参数。

在历史活跃信息中，本文取 $L=5$ 天的近期历史活跃信息来刻画高度动态性和用户粘度低等特点。因为根据对场景数据的统计分析，近 90%的用户活跃天数不超过 5 天，所以 5 天的历史活跃情况能对用户近期活跃做一个表征。考虑简历活跃的行为定义，由“活跃”的含义（参见第 1 章）可知，用户的投递行为可以明确表示其求职的意愿，而点击行为在达到一定数量之后也能体现出足够的求职意愿。不失一般性，实验中活跃度参数取 $\theta_c=3,\theta_d=1$。

接下来确定 RAP 模型中的基本参数，包括模型训练的迭代轮数和特征转换的随机森林中树的个数、深度等。通过对 RAP 的预实验来确定使用的迭代轮数。图 6 展示了模型训练中 SGD 的迭代轮数与 AUC 之间的关系曲线，可以看到在 8~10 轮后 AUC 基本不变，因此在后续实验中选取迭代轮数为 10 轮。

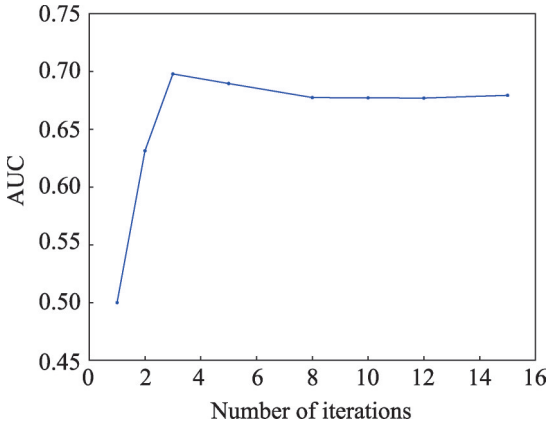


Fig.6 Relation between AUC and the number of SGD iterations when training RAP

图 6 RAP 模型训练中 AUC 值与 SGD 迭代轮数的关系

与之类似，用随机森林模型进行预实验，根据对 AUC 的影响，确定后续实验中使用的随机森林的参数，具体为 50 棵树，最大深度 10 层。

5.4 特征有效性评估

本节通过不同的基本模型考察近期历史活跃信息和双向匹配信息在简历活跃度预测问题中的不同作用。基本模型包括逻辑回归模型（LR）、随机森林模型（RF）和支持向量机（SVM）。

5.4.1 近期历史活跃信息特征

针对高度动态性、用户黏度低的场景特点，即在只有近期历史活跃信息的情况下，通过实验考察三种基本模型对信息的利用情况，探究数据和模型的特点，实验结果如表 3 所示。从表中可以观察到，三者的 AUC 值并不高，RF 相对略高一些，因为三者模型、特征都比较简单，无法很好地刻画活跃简历的特点，但是树模型 RF 在准确度上达到了 0.95 的水平，远远高于前两者的准确率和召回率，可能是通过对非

线性特征的过滤直接地筛选出了典型的活跃简历。因此,在RAP模型执行流程的设计中,也希望通过树模型对历史信息进行处理。

Table 3 Accuracy of models using history activeness information

表3 不同模型使用近期历史活跃信息的预测精度

Method	AUC	Precision	Recall
LR	0.565	0.296	**0.818**
SVM	0.555	0.294	0.771
RF	**0.571**	**0.951**	0.142

5.4.2 双向匹配信息特征

类似地,针对在线招聘市场的场景特点,在只有互动匹配信息的情况下,通过实验考查三种基本模型对信息的利用情况,探究数据和模型的特点,结果如表4所示。从表中可以看到,RF的表现尽管准确率相对较好,但是召回率很低,总体AUC效果也不如人意;与之相对,LR对互动匹配信息的利用能有相对较好的效果,尽管准确率相对较低,但是召回率相对较好,整体AUC也相对突出;SVM在召回率上表现较好,但是AUC表现一般,总体表现不如LR。因此,在RAP的设计中,希望用类似的线性模型处理互动匹配信息,达到较好的利用效果。

Table 4 Accuracy of models using bidirectional matching information

表4 不同模型使用双向匹配信息的预测精度

Method	AUC	Precision	Recall
LR	**0.763**	0.436	0.615
SVM	0.556	0.294	**0.769**
RF	0.538	**0.628**	0.096

5.5 RAP模型评估

本节考察RAP模型预测简历活跃度的效果。通过分析筛选参数γ对准确率、召回率、AUC等指标的影响,考察它的作用与灵敏性,然后通过与基线模型的比较证明RAP模型的有效性。

5.5.1 筛选参数γ

筛选参数γ的选取,依赖平台对问题目标的要求,可能与平台的运营目标、用户规模、质量要求等相关,平台也可以通过选取不同的γ提供不同质量层次的简历给招聘者。根据3.2节的理论分析可知,参数γ可以控制准确率和召回率的相对重要程度。因此,这里用实验来衡量参数γ对预测结果指标的影响。

从图7中可以看到,当筛选参数γ从零开始逐渐增加时,AUC基本不变,模型保持了对正负样本的区分能力,有效性得到了保持。而准确率和召回率随γ的改变往不同方向改变,召回率逐渐上升,准确率逐渐下降,达到筛选参数的调节作用。从图7来看γ在0.1至0.2左右时有一个比较不错的结果。

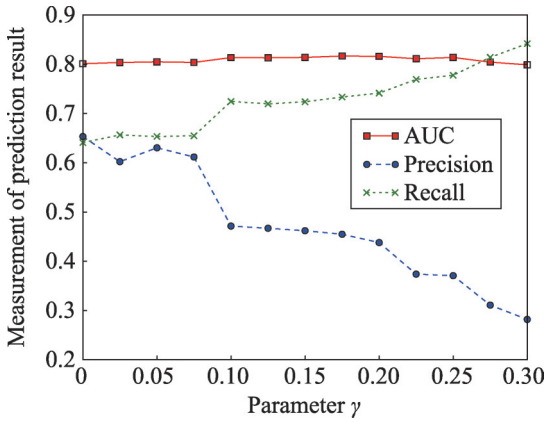


Fig.7 Influence of parameter γ on prediction accuracy

图7 参数γ对预测精度的影响

5.5.2 总体实验效果分析

下面通过实验证明RAP模型在综合利用近期历史活跃信息、互动匹配信息,解决预测问题方面的有效性。这里选取的基线模型是4.4节表现较好的两个模型,分别是使用互动匹配特征的逻辑回归(LR),以及使用近期历史活跃信息特征的随机森林(RF)。

如表5所示,RAP模型在AUC上达到了0.817,相比于两个基线模型有明显的优势,因为它综合了这两个模型对特征利用的优势,分别针对互动匹配

Table 5 Accuracy of models on activeness prediction

表5 不同模型在活跃度预测问题上的精度

Method	AUC	Precision	Recall
LR(Matching)	0.763	0.436	0.615
RF(History)	0.571	**0.951**	0.142
RAP($\gamma=0$)	0.801	0.653	0.641
RAP($\gamma=0.1$)	0.816	0.471	0.724
RAP($\gamma=0.2$)	**0.817**	0.438	**0.741**

信息和近期历史活跃信息，能更好地区分正负样本；RAP的准确率比LR高，但是比RF低，在一个可以接受的水平上，并随着筛选参数 γ 的改变而变化，召回率也能通过参数进行合理调节，满足此处对预测模型的要求，比如随着 γ 从0增长到0.2，RAP的Precision在降低，Recall在升高。

6 相关工作

本章回顾和分析两方面的工作：一般场景下的用户活跃度预测问题和在线招聘场景相关的研究。

（1）关于用户活跃度预测的研究。在用户活跃度预测方面，社交网络等平台上有较多的研究。文献[2]就用户可能退出社交平台的内、外部两方面因素进行考虑，提出了社交网络中用户活跃的总体性预测方法。文献[1]就免费网络服务的用户，采用Cox比例风险模型来拟合预测活跃情况。文献[4]就人人网实名社交平台对社交网络的用户数据特性进行了探索，根据多样性、动态性、社交影响这三方面的特点，提出了适应社交网络的活跃度预测方法。用户活跃度预测还有进一步的社会意义，文献[3]根据社交网络与肥胖症传递的相关特性[8]，结合文献[4]提出的社交网络三种主要特征，提出了社交限制性玻尔兹曼机（social restricted Boltzmann machine，SRBM）的深度学习方法，以预测健康网络中的用户行为。这些利用了社交网络特性的预测方法可以作为招聘场景的参考。另外，问答论坛[9]、网络游戏[10]、通讯行业[11-13]等场景中的用户活跃预测研究也都利用了各自内部用户之间的黏性，这和用户黏度低的招聘平台大有不同。

（2）关于在线招聘场景的研究。随着在线招聘的普及，关于在线招聘市场数据的研究也变得热门。文献[14]为了分析在线招聘市场中招聘的宏观趋势变化，提出了线性隐语义模型MTLVM（market trend latent variable model），该模型采用多级狄利克雷过程来捕获宏观过程中招聘主题的变化。文献[15]研究如何在大量的招聘数据中有效地找到具有特定技能的人才。该研究通过构建“职业变迁网络”，然后在该网络中进行稠密子图的发现，进而将网络划分成联通的“才能”网络。

7 总结

本文提出了在线招聘场景下的简历活跃度预测问题，总结了该场景下的数据具有的高度动态性、用户黏度低、双向匹配等特点，明确了预测目标的召回优先性，并在此基础上，提出了一种既能适应上述在线招聘场景特点，又能通过简单参数调节实现召回优先需求的简历活跃度预测方法RAP。通过真实数据集上的实验，RAP模型的有效性得到了证实。

References:

[1] Kapoor K, Sun Mingxuan, Srivastava J, et al. A hazard based approach to user return time prediction[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, Aug 24-27, 2014. New York: ACM, 2014: 1719-1728.

[2] Zhu Yin, Zhong Erheng, Pan S J, et al. Predicting user activity level in social networks[C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, Oct 27-Nov 1, 2013. New York: ACM, 2013: 159-168.

[3] Oentaryo R J, Lim E P , Lo D, et al. Collective churn prediction in social network[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Aug 26-29, 2012. Washington: IEEE Computer Society, 2012: 210-214.

[4] Phan N H, Dou Dejing, Piniewski B, et al. Social restricted Boltzmann machine: human behavior prediction in health social networks[C]//Proceedings of the 2015 International Conference on Advances in Social Networks Analysis and Mining, Paris, Aug 25-28, 2015. New York: ACM, 2015: 424-431.

[5] Baeza-Yates R A, Ribeiro-Neto B. Modern information retrieval [M]. Boston: Addison-Wesley Longman Publishing Co, Inc, 1999.

[6] He Xinran, Pan Junfeng, Jin Ou, et al. Practical lessons from predicting clicks on ads at Facebook[C]//Proceedings of the 8th International Workshop on Data Mining for Online Advertising, New York, Aug 24, 2014. New York: ACM, 2014: 1-9.

[7] Friedman J H, Hastie T, Tibshirani R. The elements of statistical learning (Vol.1)[M]. Berlin, Heidelberg: Springer, 2001.

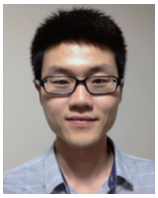
[8] Fichman R G, Kemerer C F. The illusory diffusion of innovation: an examination of assimilation gaps[J]. Information Systems Research, 1999, 10(3): 255-275.

[9] Yang Jiang, Wei Xiao, Ackerman M S, et al. Activity lifespan: an analysis of user survival patterns in online knowledge

sharing communities[C]//Proceedings of the 4th International Conference on Weblogs and Social Media, Washington, May 23-26, 2010. Menlo Park: AAAI, 2010: 186-193.
[10] Kawale J, Pal A, Srivastava J. Churn prediction in MMORPGs: a social influence based approach[C]//Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, Vancouver, Aug 29-31, 2009. Washington: IEEE Computer Society, 2009: 423-428.
[11] Dasgupta K, Singh R, Viswanathan B, et al. Social ties and their relevance to churn in mobile telecom networks[C]//Proceedings of the 11th International Conference on Extending Database Technology, Nantes, Mar 25-29, 2008. New York: ACM, 2008: 668-677.
[12] Guyon I, Lemaire V, Boullé M, et al. Analysis of the KDD Cup 2009: fast scoring on a large orange customer database [C]//Proceedings of KDD- Cup 2009 Competition, Paris, Jun 28, 2009: 1-22.
[13] Richter Y, Yom-Tov E, Slonim N. Predicting customer churn in mobile networks through analysis of social groups[C]//Proceedings of the 2010 SIAM International Conference on Data Mining, Columbus, Apr 29- May 1, 2010. Philadelphia: SIAM, 2010: 732-741.
[14] Zhu Chen, Zhu Hengshu, Xiong Hui, et al. Recruitment market trend analysis with sequential latent variable models [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, Aug 13-17, 2016. New York: ACM, 2016: 383-392.
[15] Xu Huang, Yu Zhiwen, Yang Jingyuan, et al. Talent circle detection in job transition networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, Aug 13-17, 2016. New York: ACM, 2016: 655-664.

SHI Shuyang was born in 1994. He received the B.S. degree in computer science and technology from Peking University in 2017. Currently he is pursuing the M.S. degree at Stanford University.
史舒扬(1994—),男,浙江杭州人,2017年于北京大学计算机科学与技术专业获得学士学位,目前在斯坦福大学就读硕士学位。

ZHANG Zhipeng was born in 1993. He is a Ph.D. candidate at Peking University. His research interest includes graph computing, machine learning and general distributed data processing frameworks, etc.
张智鹏(1993—),男,江苏淮安人,北京大学博士研究生,主要研究领域为图计算,机器学习,通用分布式计算框架等。



GUO Long was born in 1988. He received the Ph.D. degree from National University of Singapore in 2015. Now he is a postdoctoral researcher at School of Electronics Engineering and Computer Science, Peking University. His research interests include spatial and temporal data mining, spatial databases, machine learning and natural language processing, etc.
郭龙(1988—),男,山东寿光人,2015年于新加坡国立大学获得博士学位,现为北京大学信息科学技术学院博士后,主要研究领域为时空数据挖掘,空间数据库,机器学习,自然语言处理等。

SHAO Yingxia was born in 1988. He received the Ph.D. degree from Peking University in 2016. Now he is a postdoctoral researcher at School of Electronics Engineering and Computer Science, Peking University. His research interests include large-scale graph analysis, knowledge graph and natural language processing, etc.
邵蓥侠(1988—),男,浙江宁波人,2016年于北京大学获得博士学位,现为北京大学信息科学技术学院博士后,主要研究领域为大规模图数据分析,知识图谱,自然语言处理等。

CUI Bin was born in 1975. He received the Ph.D. degree from National University of Singapore in 2004. Now he is a professor and Ph.D. supervisor at Peking University. His research interests include database system architectures and performance optimization, data mining and big data management, etc.
崔斌(1975—),男,浙江宁波人,2004年于新加坡国立大学获得博士学位,现为北京大学教授、博士生导师,主要研究领域为数据库系统设计和性能优化,数据挖掘,大数据管理等。