DOI:10.3969/j.issn.1673-4785.201403019
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20150403.1732.001.html

社交网站中用户评论行为预测

孔庆超,毛文吉,张育浩
(中国科学院自动化研究所 复杂系统管理与控制国家重点实验室,北京 100190)

摘 要:社交网站为用户相互交流、发表意见和观点提供了非常便利的平台。对社交网站的用户行为进行建模和预测对于安全、商业等多个领域具有十分重要的社会意义和应用价值,近年来逐渐得到研究者的重视。面向社交网站中用户评论行为,预测用户是否会参与讨论。采用基于特征的机器学习方法,其中特征包括讨论帖子及其内容、用户行为特征和社交关系,并引入参数控制数据集的不平衡性。实验采用来自豆瓣小组的真实数据。实验结果表明,新提出的用户行为和社交关系特征以及对不平衡数据集的处理方法能够有效提高用户评论行为的预测效果,进一步说明用户的历史行为和所在的社交关系网络对当前的评论行为有较大影响。

关键词:社交网络;用户评论;机器学习;行为建模;行为预测;不平衡性数据集
中图分类号:TP391 **文献标志码**:A **文章编号**:1673-4785(2015)03-0349-05

中文引用格式:孔庆超,毛文吉,张育浩. 社交网站中用户评论行为预测[J]. 智能系统学报, 2015, 10(3): 349-353.
英文引用格式:KONG Qingchao, MAO Wenji, ZHANG Yuhao. User comment behavior prediction in social networking sites[J]. CAAI Transactions on Intelligent Systems, 2015, 10(3): 349-353.

User comment behavior prediction in social networking sites

KONG Qingchao, MAO Wenji, ZHANG Yuhao
(State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Science, Beijing 100190, China)

Abstract: Social networking sites provide a convenient way for users to communicate with others and to present opinions. Related researches on modeling and predicting user behaviors in social networking sites are of vital importance for many applications in the domains of security and business. The aim of this paper is to predict user comment behavior based on postings in social networking sites. A feature-based machine learning approach is employed, which includes features from the postings, content, user behaviors and social relations, and introduces a parameter to control the imbalanceness of the dataset. Real-world datasets from Douban Group were used in the experiments. The experimental results showed that the user behavior and social relation features and the imbalance processing technique effectively improved the prediction performance of user comment behaviors. This further demonstrates that the user comment behavior is largely affected by their behavior history and social network.

Keywords: social network; user comments; machine learning; behavior modeling; behavior prediction; imbalance dataset

社交网站如 Facebook、Twitter 等对人们的生活产生了巨大影响。人们在社交网站中更新状态或发送广播,以此来展现自己的生活状态、发表感想或与朋友们分享信息。社交网站已成为人们获取信息、参与讨论和表达观点的重要平台。另一方面,用户在社交网站中的行为则体现了用户的行为模式和兴趣。由于社交网站中用户众多,并且具有实时性的

收稿日期:2014-03-05. **网络出版日期**:2015-04-03.
基金项目:国家自然科学基金资助项目(61175040, U1435221).
通信作者:毛文吉. E-mail: wenji.mao@ia.ac.cn.

特点,人们讨论的话题能够迅速在网络上传播和演化,因此理解他们的行为,并对其进行建模和分析显得十分重要。例如,在信息检索领域,预测参与哪些话题的讨论可以帮助服务提供者了解更多有关用户兴趣和需求的信息;在情报安全领域,追踪和预测用户参与的话题讨论可以帮助决策者更好地了解和掌握用户的行为特点。

面向社交网站的讨论组用户,本文提出一种预测用户是否会参与讨论的计算方法。具体而言,采用基于特征的方法,考虑了影响用户评论行为的多种主要因素,并使用机器学习算法结合所有的特征,最终得到用户对某个帖子进行回复的概率。此外,由于一个帖子中的评论用户数量相对于讨论组中的用户总数而言总是很少,往往造成了数据集中的类不平衡问题,本文还针对这一问题做了相应处理。本文主要贡献在于:结合多类不同类型的特征,通过与评论行为相关联的社交关系特征,如用户活跃度、用户间的关注关系等预测评论行为,并通过对类不平衡问题的处理提升预测的效果。

1 相关研究工作

在博客和论坛中,网站大多都会提供用户评论功能。用户评论能够促进用户之间的交流,发帖人也能够通过用户评论得到反馈。目前已经有一些预测博客评论数量的相关工作。M. Tsagkias 等[1]基于文本、语义和现实世界特征预测一篇博文是否会有评论以及评论数量的多少。T. Yano 和 N. A. Smith[2]采用主题模型预测博客的评论数量。预测微博的转发量[3-6]和博客中的评论数量这 2 项研究具有一定的相似性,如 L. Hong 等[4]通过抽取 Twtiter 中评论内容、时序信息、评论及用户的元数据以及用户社交网络的结构关系等,基于特征的模型预测微博的转发数量。其他的相关研究还包括视频[7-8]、图片[9-10]浏览数的预测等。

相对于预测用户的评论数量,预测用户评论行为(即用户是否会对某个帖子进行回复)是一项更具挑战性的工作。T. Yano 等[11]构建了一个能够同时生成评论和博文内容的概率模型预测用户会评论哪一篇博文。Tang 等[12]构建了一个用户兴趣和话题检测模型(UTD)。在给定已有部分用户对某个帖子进行回复的条件下,UTD 模型通过获取话题内容和发展趋势预测哪些用户会对新的帖子产生兴趣。

上述工作中提出了一系列影响用户在博客和讨论组评论行为的因素,如帖子和内容信息。然而,除了帖子本身和话题相关的内容信息以外,社交网络中还存在着多个与用户评论行为密切相关的社会关系、用户行为特征等因素,充分利用这些信息可以更好地预测用户行为。例如,用户之间的“粉丝”关系可能会影响用户对发帖人所发帖子的回复。

本文分析了影响用户评论行为的主要因素,并构建了一个结合多个关键影响因素、基于特征的评论行为预测模型。这些因素不仅包括话题内容特征,也包括社会关系和行为特征。采用逻辑回归模型进行分类,并通过采样方法解决类不平衡问题,最终预测用户评论某个帖子的可能性。

2 问题定义

用 U 表示讨论组中的用户集合,D 表示帖子的集合,每个帖子 $d(d \in D)$ 包括标题、内容和发帖者 $u_d(u_d \in U)$ 的信息,其中标题和内容都采用词袋(bag of words, BOW)模型表示。除了帖子本身的内容外,已知的信息还包括用户间的关注关系和用户的历史行为信息。将问题定义为:给定帖子 d 的信息,预测目标用户 $u(u \in U)$ 对帖子 d 进行评论的概率。

3 用户评论行为建模与预测

针对用户评论行为,详细介绍特征选取、如何计算内容相似度以及构建逻辑回归模型进行预测,并讨论对类不平衡问题的处理。

3.1 特征选取和内容相似度计算

考虑采用 4 类可能影响用户评论行为的因素作为特征,包括帖子本身、内容相似度、用户行为和社交关系等特征。

1)帖子特征:包括标题长度、正文长度、图片和外链的数量,共 4 个特征。

2)内容相似度特征:包括帖子内容和目标用户兴趣之间的相似度以及发帖者兴趣和目标用户兴趣之间的相似度,共 2 个特征。

3)用户行为特征:包括发帖者和目标用户各自发过的帖子数量以及评论回复数量,共 4 个特征。

4)社交关系特征:包括发帖者和目标用户各自的关注和粉丝的数量,以及目标用户是否关注了发帖者,共 5 个特征。

在计算内容相似度特征时,应用 LDA 模型[13]得到用户关注内容在不同主题(topic)的分布情况来刻画用户兴趣。LDA 模型是近年来非常受关注的主题模型,其数学表达简洁,而且文本建模效果很好。具体来说,首先,在整个数据集上训练得到一个

LDA 模型;然后,将每个用户曾经发布的所有帖子和评论组成一个文档;最后,将该文档输入 LDA 模型中得到用户感兴趣的主题分布,即一定长度的向量(长度为训练 LDA 模型时指定的主题个数)。通过应用 LDA 模型,所有的文档和用户兴趣可以表示为主题分布向量,而文本内容之间的相似度则定义为 2 个主题分布之间的欧氏距离。

将以上 4 类特征(共 15 个)组成 15 维向量作为特征向量。需要注意的是,特征向量中的特征具有不同的数值类型,如"标题长度"为离散型特征,而内容相似度为连续型特征,取值范围也不尽相同,所以需要在模型训练前对特征进行归一化处理。本文分别对于每个特征进行归一化:

$$f_{\text{normalized}} = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{1}$$

式中:f 为样本的特征取值,$f_{\min}$ 为该特征在所有样本(包括训练集和测试集)中的最小取值,$f_{\max}$ 为该特征在所有样本中的最大取值。

3.2　模型建立及类不平衡的处理

逻辑回归(LR)模型是一种线性分类模型,可以得到样本属于每个类别的概率。对于每个目标用户和帖子,抽取出上面列举的所有特征,组成一个特征向量。令 $\boldsymbol{x}$ 表示特征向量,$\boldsymbol{w}$ 表示特征的权重向量。Y 表示预测结果,为二值随机变量,当目标用户评论时 $Y=1$,目标用户不评论时 $Y=0$。

$$P_w(Y=1 \mid \boldsymbol{x}) = \frac{1}{1+\mathrm{e}^{-g(\boldsymbol{x})}} \tag{2}$$

式中:$g(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$,$P(Y=1 \mid \boldsymbol{x})$ 表示目标用户评论帖子的概率。根据式(1),$g(\boldsymbol{x})>0$ 时模型预测 $y(\boldsymbol{x})=1$,$g(\boldsymbol{x})>0$ 时模型预测 $y(\boldsymbol{x})=0$。

在构建分类器的训练集和测试集时,对于一个帖子来说,可以认为所有真正参与评论的用户为正样本,而没有参与评论的用户为负样本。一般而言,由于负样本的数量远大于正样本数量,这就造成数据集中存在类不平衡问题。在给定特征集合的条件下,如果数据集中类不平衡性较大,那么采用分类学习算法最终将预测所有样本为负样例。这样的预测结果虽然准确率很高,但实质上完全忽略了正样本的重要性,即人们真正关心的是哪些用户会参与评论,而不是哪些用户不参与评论。

解决类不平衡问题的方法有多种,其中采样法是最常采用的一种方法。具体来说,本文采用的采样算法包括随机上采样和下采样、EasyEnsemble[14] 和 SMOTE 算法[15],其中 EasyEnsemble 的性能相对较好[14]。需要说明的是,在使用 EsayEnsemble 算法时,我们将原文献中 EasyEnsemble 的基本分类器 AdaBoost 替换为 LR 模型。

综上,构建分类预测框架的主要步骤是:首先构建训练集与测试集并抽取和计算特征,然后采用以上采样方法提高正样例在训练集中所占的比例,最后建立 LR 模型得到用户评论某个帖子的预测结果。

4　实验结果与分析

4.1　数据集和预处理

实验的数据集来自豆瓣小组。豆瓣是国内流行的社交网站,有超过七千万用户。作为豆瓣网站的一部分,豆瓣小组允许用户建立不同主题的小组。小组成员可以发布帖子,其他人可以对帖子进行评论,点击"喜欢"或者推荐给关注自己的用户。实验抓取的是豆瓣"美剧 fans"小组所有的帖子和评论。

训练集包括从 2012-8-1—2012-12-1 期间发布的所有帖子及评论,测试集包括从 2012-12-1—2013-1-1 发布的所有帖子及评论。在测试集中,实验移除了一些用户和帖子,以保证测试集中每个用户至少发表过 2 次评论,每个帖子至少有 5 个用户评论。

表 1 展示了预处理后训练集和测试集中帖子、评论和用户数量的统计数据。

表 1　训练集和测试集统计数据

Table 1　Statistics of training and test set

数据集	帖子数	评论数	用户数
训练集	535	20 473	3 467
测试集	112	4 226	708

4.2　目标用户集合

在进行模型训练和测试时,由于豆瓣小组中成员众多(截止到目前,"美剧 fans"小组共有 178 298 个成员),因此需要选择构建一个较小的且大小可控的目标用户集。

用 T_{S_d} 表示帖子 d 的目标用户集合。对于每个帖子 d,通过如下方式构建 T_{S_d}:首先将 S_d 中的所有用户加入 T_{S_d},于是 $S_d \in T_{S_d}$,然后从不在 S_d 中的其他小组成员中随机选择 $R \times |S_d|$ 个用户放入 T_{S_d} 中,其中 $|S_d|$ 代表集合 S_d 中的用户数目,R 为实验中设定的正整数。于是,对数据集中的每个帖子 d,都有 $|S_d|$ 个正样例(真正参与评论的用户)和 $|T_{S_d}| - |S_d|$ 个负样例(没有参与评论的用户)。显然,R 越大,数据集不平衡问题就越严重,换句话说,R 控制着数据集的不平衡性。

4.3 评价标准

由于数据集中存在类不平衡问题,传统的评价标准,如准确率和召回率并不适合于评估文中的实验结果。用户评论行为的预测可以看作一种信息检索问题(本文是“用户检索”),即检索哪些用户会以较高的概率对特定的帖子进行回复,所以选取信息检索中的常用评价指标 Precision@K。针对某一个帖子 d,预测框架会返回最有可能对帖子进行回复前 K 个用户,记为 K_d^{top}。对于帖子 d,其 Precision@K 计算方式如下:

$$\text{Precision@}K_d = \frac{|K_d^{\text{top}} \cap S_d|}{K} \tag{3}$$

4.4 实验结果

4.4.1 用户行为和社交关系特征的作用

首先考察用户行为和社交关系这 2 类特征对 Precision@K 的作用。图 1(a)和(b)分别展示了 $R=5$ 和 $R=10$ 时,测试集中每个帖子的 Precision@5 分布情况,其中黑色线表示加入了用户行为和社交关系特征的 Precision@5 分布,浅色线表示没有考虑这 2 类特征的 Precision@5 分布。图中所有帖子按 Precision@5 值降序排列。

从图 1 中可以看出,在不考虑用户行为和社交关系这 2 类特征时,虽然对于一小部分帖子,其 Precision@5 比考虑这 2 类特征时要高,但整体而言后者的效果更好。表 2 中的数据同样支持这个结论。表 2 展示了当 R 取不同值时,包含和不包含用户行为和社交关系 2 类特征下的平均 Precision@5。

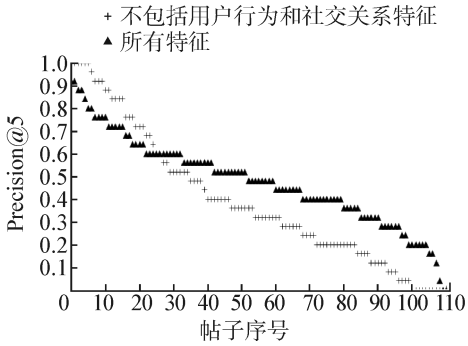
表 2 平均 Precision@5 对比

Table 2 Average Precision@5

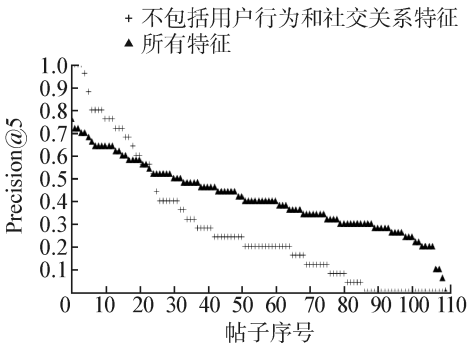
特 征	$R=5$	$R=10$	$R=15$	$R=20$
不含用户行为和社交关系特征	0.39	0.28	0.23	0.20
包含所有特征	0.47	0.36	0.31	0.27

4.4.2 类不平衡性的影响

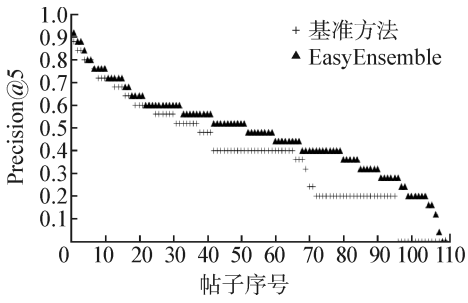
这里考察类不平衡问题对预测结果的影响。图 1(c)和(d)分别展示了 $R=5$ 和 $R=10$ 时,测试集中每个帖子的 Precision@5 分布情况,其中黑色线表示采用了 EasyEnsemble 采样方法后的 Precision@5 分布,浅色线表示没有对数据集进行采样处理的帖子的 Precision@5 分布。图中所有帖子按 Precision@5 值降序排列。从图中可以看出,在对数据集进行采样处理后,预测结果得到明显提升。从表 3 中同样可以看出,在对数据集进行采样处理之后,当 R 取不同值时,平均的 Precision@K 都有显著提高。



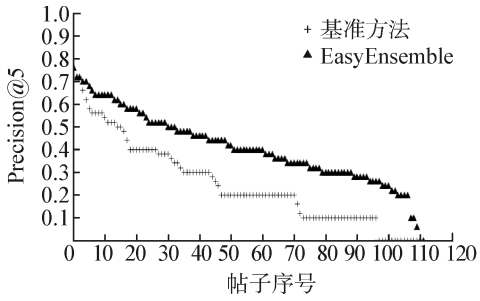
(a)不包括用户行为和社交关系特征 vs. 所有特征($R=5$)



(b)不包括用户行为和社交关系特征 vs. 所有特征($R=10$)



(c)基准方法 vs. EasyEnsemble($R=5$)



(d)基准方法 vs. EasyEnsemble($R=10$)

图 1 当 R 取 5 和 10 时不同方法的实验结果比较

Fig. 1 Experimental results comparison of different methods when $R=5$ and $R=10$

表 3 平均 Precision@5 对比

Table 3 Average Precision@5

数据集	$R=5$	$R=10$	$R=15$	$R=20$
不采样处理	0.38	0.25	0.18	0.14
EasyEnsemble	0.47	0.36	0.31	0.27

4.4.3　Precision@K 分布

图 2 展示了当 R 取不同值时,预测结果 Precision@5 的分布情况。从图中可以看出,本文的预测框架的性能尚不够稳定,对于一些帖子的 Precision@5 接近 1,而对于另一些帖子的 Precision@5 却较低。

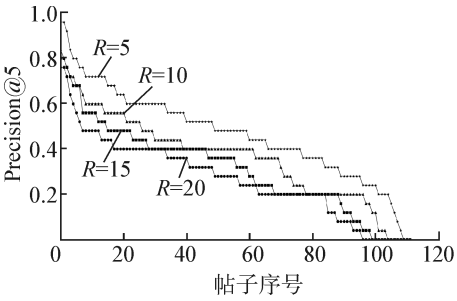


图 2　当 R 取不同值时的实验结果

Fig. 2　Experimental results with varying R

5　结束语

本文以用户评论行为为例,给出一种基于特征的行为预测方法。实验结果表明,新提出的 2 种新特征,即用户行为特征和社交关系特征,以及控制数据集的不平衡性的参数能够有效提升行为预测准确度。同时,进一步说明用户在社交网站中的评论行为受到其历史行为和社交关系的影响。未来的研究工作将尝试分析用户评论行为的生成过程,探讨其中起关键作用的因素并建立生成式模型,以提高预测结果的准确度和可解释性。

参考文献:

[1] TSAGKIAS M, WEERKAMP W, De RIJKE M. Predicting the volume of comments on online news stories[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, USA, 2009: 1765-1768.

[2] YANO T, SMITH N A. What's worthy of comment? Content and comment volume in political blogs[C]//Proceedings of the Fourth International Conference on Weblogs and Social Media. Washington, DC, USA, 2010: 359-362.

[3] ZAMAN T, FOX E B, BRADLOW E T. A bayesian approach for predicting the popularity of tweets[J]. The Annals of Applied Statistics, 2014, 8(3): 1583-1611.

[4] HONG L, DAN O, Davison B D. Predicting popular messages in Twitter[C]//Proceedings of the 20th International Conference Companion on World Wide Web. New York, USA, 2011: 57-58.

[5] MA Haixin, QIAN Weining, XIA Fan, et al. Towards modeling popularity of microblogs[J]. Frontiers of Computer Science, 2013, 7(2): 171-184.

[6] JENDERS M, KASNECI G, NAUMANN F. Analyzing and predicting viral tweets[C]//Proceedings of the 22nd International Conference on World Wide Web Companion. Geneva, Switzerland, 2013: 657-664.

[7] FIGUEIREDO F. On the prediction of popularity of trends and hits for user generated videos[C]//Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. New York, USA, 2013: 741-746.

[8] PINTO H, ALMEIDA J M, GONCALVES M A. Using early view patterns to predict the popularity of Youtube videos[C]//Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. New York, USA, 2013: 365-374.

[9] KHOSLA A, DAS SARMA A, HAMID R. What makes an image popular? [C]//Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea, 2014: 1-10.

[10] CHENG J, ADAMIC L, DOW P A, et al. Can cascades be predicted? [C]//Proceedings of the 23rd International Conference on World Wide Web. Seoul, Korea, 2014: 925-936.

[11] YANO T, COHEN W W, SMITH N A. Predicting response to political blog posts with topic models[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA, 2009: 477-485.

[12] TANG X N, YANG C C, ZHANG M. Who will be participating next? Predicting the participation of dark web community[C]//Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics. New York, USA, 2012: 1-7.

[13] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

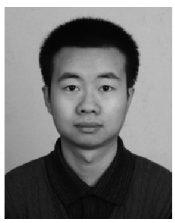
[14] LIU Xuying, WU Jianxin, ZHOU Zhihua. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 539-550.

[15] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

作者简介:

孔庆超,男,1987 年生,博士研究生,主要研究方向为社会媒体信息分析与处理、数据挖掘。

毛文吉,女,1968 年生,研究员,博士生导师,主要研究方向为智能信息处理、人工智能、社会计算。曾获国家科技进步二等奖,"吴文俊人工智能科学技术奖"创新二等奖,"中国自动化学会科学技术进步奖"一等奖,发表学术论文 40 余篇。



张育浩,男,1989 年生,博士研究生,主要研究方向为社会建模与计算。