基于用户行为的高校 BBS 热帖预测模型

于兴隆 李丽萍 吴 斌
(北京邮电大学计算机学院 北京 100876)

摘 要 校园 BBS 是高校网络舆论的主要载体,反应了大学生的舆论倾向以及生活的各个方面,高校 BBS 的实证研究具有重要的意义。如何高效地对帖子的热度进行预测是发现突发网络舆情的基础,对网络舆情的研究具有重要的意义。以一高校 BBS 实际的数据为研究对象,对帖子和用户进行深入分析,提出一种基于用户行为的高校 BBS 热帖预测模型,通过实验分析,该方法可以对论坛中的热帖进行较好的预测。

关键词 热帖预测 用户聚类 人类行为动力学 高校 BBS

中图分类号 TP391 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2013.01.011

PREDICTION MODEL FOR HOT COLLEGE BBS POSTS BASED ON USER BEHAVIOUR

Yu Xinglong Li Liping Wu Bin
(*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*)

Abstract Campus BBS is the main carrier of college networks public opinion, it reflects the tendency of college students' public opinions as well as various aspects of their life, the empirical study on college BBS has important significance. How to efficiently predict the hot degree of the posts is the base of the discovery of unexpected networks public opinion, and is of significant to the study of networks public opinions. In this paper, we take the actual data of a university BBS as the research object, and make in-depth analysis on the posts and the users. We then put forward a prediction model for hot college BBS posts which is based on user behaviour. Through the experimental analysis, this method is proved to be able to well predict the hot BBS posts.

Keywords Hot posts prediction User clustering Human behaviour dynamics College BBS

0 引 言

随着互联网技术的发展,高校网络舆论的载体变得越来越广泛,各类网站的公共论坛、校园 BBS、社交网站、学生个人网站、博客、微博等为个人观点的表达和传播提供了一个自由的公共平台,"权威"、"中心化"等得以消解,每个大学生都可以成为信息的发布者。但从目前的调查情况看,高校网络舆论的主要载体还是校园 BBS[1]。基于 BBS 的舆情研究已经成为当前一个研究热点和难点。

目前,BBS 舆情研究大部分都是围绕社会学、传播学以及心理学、文本挖掘等角度进行展开,分析手段的限制使得大部分研究仍停留在探索阶段。如文献[2,3]建立情感激励模型来模拟回复量的变化情况,对于预测起到一定的借鉴作用;文献[4]从网络拓扑结构进行分析,但没有涉及现实论坛情况;文献[5]提出了一种基于隐马尔科夫的网络舆情预测模型,但是不能对网络舆情突发事件进行预测。对于网络舆情的发现目前主要以文本分类和聚类为主要研究手段,文献[6]从帖子回复结构出发,利用词语的潜在影响力进行聚类,发现当日的"十大"热贴,但是准确率和查全率都偏低,而且这种方法有滞后性。文献[7]利用粗糙集结合集成学习的方法建立网络舆情分类模型,对网络舆情的预测进行了初步的探索。

本文针对 BBS 论坛帖子热度预测问题,提出了基于用户行为的热帖预测模型,其流程如图 1 所示,论坛数据采集模块用来收集高校 BBS 数据,作为研究的真实数据源;用户和帖子特征分析主要包括了基于用户行为的聚类分析、典型的长贴时间间隔序列模式分析,另外,本文对帖子回复时间间隔分布进行了实证研究;在大量数据分析的基础上,本文提出了 7 种帖子热度的预测属性,并结合集成分类器建立了预测模型,最终达到 BBS 热帖预测的目的。

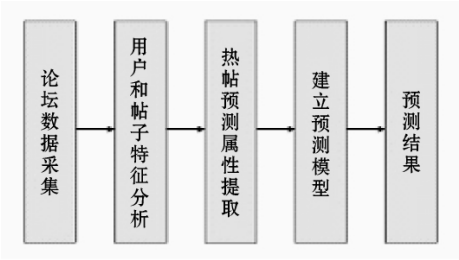


图 1 BBS 热帖预测模型流程图

收稿日期:2012-08-07。2012 中国计算机大会论文。国家自然科学基金项目(61074128)。于兴隆,硕士,主研领域:数据挖掘,社会网络。李丽萍,硕士。吴斌,教授。

1　相关概念及数据描述

1.1　论坛网络模型

在互联网论坛中,用户之间通过帖子的回复关系可建立明显的连接关系。随着时间的推移、帖子的发布和用户的增加,用户之间的关系也逐渐复杂,将会形成庞大的网络。我们可以发现用户之间的关系与言论和话题之间存在映射关系,如图 2 所示。其中 $Set(U)$ 表示用户的集合,$Set(O)$ 表示言论的集合,$Set(T)$ 表示话题的集合,言论是来自用户的集合,一定的言论又是针对特定的话题,可见它们之间存在一种映射关系。因而通过构建用户之间的关系网络,就可以反应用户的影响力以及间接地反应舆情的变化。

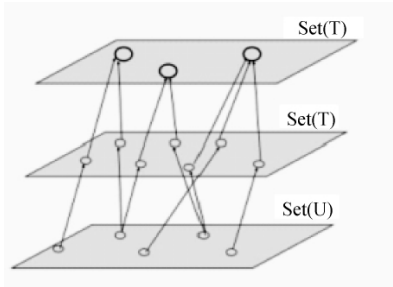


图 2　用户与话题关系

1.2　数据处理及描述

本文采用爬虫技术对截止 2011 年 11 月 24 日的北邮人论坛(http://bbs.byr.cn/#!Default)数据进行了的全站爬取。为了方便后面的研究,本研究中把抓取到的 html 格式的网页文件通过解析转化为设定的 XML 文件格式和 SQL 文件格式。

为了研究论坛群体用户发帖时间间隔的分布规律和长帖的典型模式,抽取了每个话题所有帖子的发帖时间,并计算了相邻发帖时间的间隔,选取了几个典型的版面作为研究的对象,数据整理结果如表 1 所示。

表 1　回帖时间间隔整理结果

版面	总话题数	总间隔数	最长贴回复数
燕赵情怀	410	6714	71
情感天空	1226	35582	524
足球吧	2045	32687	257
缘来如此	256	7900	856
笑口常开	3501	72389	784
谈天说地	1770	24523	529

根据定义的用户行为数据格式,对 57584 注册用户的行为数据进行了统计。用户行为数据的格式为“用户名,总发帖量,发主贴数,发回帖数,总共使用过版块数量,2011 年发贴量,其他年份发帖量,一天中每六个小时的发帖数汇总”。其中,发主贴数和发回帖数反应了用户的使用倾向,总共使用过的板块数量反应了用户兴趣的集中程度。2011 年发贴量反应了用户近期活跃度,一天中每六个小时的发帖数反应了用户的发帖的时间规律。以上定义的数据格式从用户兴趣和使用时间分布等方面反应了用户行为特点。

为了对基于用户行为的角色识别的有效性进行验证,本文整理了论坛注册用户的等级数据库,全论坛共 14 名站务用户账号,225 个管理员用户账号,其他为普通用户账号。为了对热帖进行预测,本文整理了一周(2011 年 10 月 20 日 – 2011 年的 10 月 26 日)的论坛几个热门版面的热帖以及普通帖子数据,帖子的特征将在 3.4.1 节进行详细说明。

2　算法和模型

2.1　BBS 动力学实证研究

人类的动力学行为具有非线性,并且高度复杂,一直是社会心理学研究的中心问题之一。由于缺少人类行为的定量数据,绝大多数以前的研究常常把人类的行为简化为可以使用泊松过程描述的稳态随机过程[8],即人类行为发生的速率是近似均匀的,两个相继行为之间存在极大的时间间隔的概率很小。2005 年 Barabasi 通过挖掘分析人类活动的历史数据,发现人类从事特定活动的行为具有阵发和胖尾的特性,也就是说,这些行为的发生过程是不能用泊松过程描述的[9]。之后,引发了学者们对这一问题极为广泛的研究。已有学者的研究涉及网页浏览[10]、短消息发送[11]、微博[12]、博客评论[13]、在线电影点播[14]等。包含了商业行为、娱乐行为、日常使用习惯等众多的人类行为。在这些行为中,普遍发现有偏离泊松过程的特性。这些现象显示,除了受到生理周期强烈影响的部分行为外,时间间隔统计所显示的非泊松特性可能是在人类行为中普遍存在的。在人类行为动力学研究的初始阶段,实证研究具有特别重要的意义。本文针对论坛的发帖时间间隔进行了如下研究。

2.1.1　概率分布模型

设 X 为独立同分布的随机变量,如果其概率密度函数满足:

$$p(x) \propto x^{-\alpha} \tag{1}$$

其中,α 为标度指数,称 X 服从幂律分布。对于连续幂律分布:

$$p(x)\,\mathrm{d}x = P(x \leqslant X \leqslant x + \mathrm{d}x) = Cx^{-\alpha}\mathrm{d}x \tag{2}$$

其中 C 为标准化常量,当 $x \to 0$ 时发散,因此存在 x 的下限 $x_{\min}$ 使得式(2)成立。根据标准化条件 $\int_{x_{\min}}^{\infty} Cx^{-\alpha}\mathrm{d}x = 1$,解得 $C = (\alpha - 1)x_{\min}^{\alpha-1}$,且 $\alpha > 1$。概率密度函数可以写为:

$$p(x) = \frac{\alpha - 1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha} \tag{3}$$

其累计概率分布函数为:

$$P(x) = \int_{x_{\min}}^{\infty} p(x)\,\mathrm{d}x = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1} \tag{4}$$

对于离散幂律分布,$p(x) = P(X = x) = Cx^{-\alpha}$,同理,取 $x_{\min}$ 并标准化后,可以表示为:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{5}$$

其中,$\zeta(\alpha, x_{\min}) = \sum_{x=0}^{\infty}(x + x_{\min})^{-\alpha}$。回帖时间间隔是离散型的随机变量,本文采用后者建模。

2.1.2　参数估计

本文采用极大似然法对计算幂指数,根据文献[15]提出的方法,先求出其似然函数:

$$p(x \mid \alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}}\left(\frac{x_i}{x_{\min}}\right)^{-\alpha} \tag{6}$$

对其求对数后对参数求导,可以得到参数方程:

$$\hat{\alpha} = 1 + n\left(\sum_{i=1}^{n} \ln\frac{x_i}{x_{\min}}\right)^{-1} \tag{7}$$

这个方程给出了估计的幂指数的值。然而,在现实网络应用中,更多的是离散的度序列值,简单地把离散的度看成是连续的度来分析计算,得出的结果就会存在偏差。对于离散的度来说:

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \tag{8}$$

其中,$\zeta(\alpha, x_{\min}) = \sum_{x=0}^{\infty}(x + x_{\min})^{-\alpha}$ 为 zeta 函数,累计分布为:

$$P(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{\min})} \tag{9}$$

同理可以推导出:

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{10}$$

理论上可以从方程解出幂指数。事实上精确的解很难获得,文献[15]通过一个模糊的整数函数 $F(x)$ 获得近似解法($F(x)' = f(x)$),可得到参数的近似估计值:

$$\hat{\alpha} \approx 1 + n\left(\sum_{i=1}^{n}\ln\frac{x_i}{x_{\min} - 0.5}\right)^{-1} \tag{11}$$

用连续的度的逼近来代替离散度值的偏差很小,经验证明 $x_{\min} \geqslant 6$ 时偏差极小。

2.1.3 假设检验

依对于 $x_{\min}$ 的求解过程中,或者假定其为已知的值,比如假设其为 X 序列的最小值,或者从图中直观地观测得到,这样处理往往会给计算带来误差。文献[15]引进了 KS 统计量,较好地解决了这个问题,KS 统计量的基本算法在于寻找 2 个分布的最大距离,它既可以用来检验样本是否来源于给定分布,也可以检验 2 个样本分布是否相似。

$$K = \max_{x \geqslant x_{\min}} |F^*(x) - S(x)| \tag{12}$$

式中,$F^*(x)$ 是拟合的累计分布,$S(x)$ 是基于样本的经验累积分布。基于最小的 K 值的 $x_{\min}$ 就是所求的统计量。这种方法不仅使用于检验幂律分布,对于检验其他的分布也是适用的。

对于检验幂律分布的拟合优度而言,这个时候的 $F^*(x)$ 和 $S(x)$ 的代表的分布变为:$F^*(x)$ 为拟合最好的累积分布,$S(x)$ 为满足给定幂律条件下随机产生的分布(包括 $x_{\min}$ 在内)。本文采用 $p \leqslant 0.05$ 作为对于 KS 统计量拒绝原假设检验标准。

2.2 基于 FCM 算法用户行为聚类

2.2.1 模糊 C 均值聚类

本文采用了 FCM 算法,即模糊 C 均值算法[16]。它是一种基于划分的聚类算法,基本思想为通过聚类过程,使得被划分到同一簇的对象之间相似度最大,而不同簇之间的相似度最小。它是对普通 C 均值聚类算法(K-MEANS 算法)的改进,后者对于数据的划分是硬性、非此即彼的。而 FCM 则通过加入模糊集合的概念,实现了一种柔性的模糊划分。FCM 把 n 个向量 $x_i(i = 1,2,\cdots,n)$ 分为 c 个模糊组,并求每组的聚类中心,使得非相似性指标的价值函数达到最小。FCM 与 HCM 的主要区别在于 FCM 用模糊划分,使得每个给定数据点用值在 0,1 间的隶属度来确定其属于各个组的程度。与引入模糊划分相适应,隶属矩阵 U 允许有取值在 0,1 间的元素。不过,加上归一化规定,一个数据集的隶属度的和总等于 1:

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j = 1,\cdots,n \tag{13}$$

那么,FCM 的价值函数(或目标函数)为:

$$J(U, c_1, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{14}$$

这里 u_{ij} 介于 0,1 间;c_i 为模糊组 i 的聚类中心,$d_{ij} = \|c_i - x_j\|$ 为第 i 个聚类中心与第 j 个数据点间的欧几里德距离;且 $m \in [1,\infty)$ 是一个加权指数。构造如下新的目标函数,可求得使式(13)达到最小值的必要条件:

$$\bar{J}(U, c_1, \cdots, c_c, \lambda_1, \cdots, \lambda_n) = J(U, c_1, \cdots, c_c) + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c} u_{ij} - 1\right)$$

$$= \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c} u_{ij} - 1\right) \tag{15}$$

这里 $\lambda_j(j = 1,2,\cdots,n)$ 是式(13)的 n 个约束式的拉格朗日乘子。对所有输入参量求导,使式(14)达到最小的必要条件为:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{16}$$

和

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \tag{17}$$

由上述两个必要条件,模糊 C 均值聚类算法是一个简单的迭代过程。在批处理方式运行时,FCM 用下列步骤确定聚类中心 c_i 和隶属矩阵 U。

算法 1 FCM 算法

输入: 聚类样本数据和要求的聚类数目。

输出: 隶属矩阵和聚类中心。

步骤 1 用值在 0,1 间的随机数初始化隶属矩阵 U,使其满足式(13)中的约束条件。

步骤 2 用式(16)计算 c 个聚类中心 $c_i, i = 1,\cdots,c$。

步骤 3 根据式(14)计算价值函数。如果它小于某个确定的阈值,或它相对上次价值函数值的改变量小于某个阈值,则算法停止。

步骤 4 用式(17)计算新的 U 矩阵。返回步骤 2。

2.2.2 聚类评价指标

为了对聚类的结果进行正确评价,确定最佳的聚类数,本文通过分析已提出的 5 种模糊聚类有效性指标,即:划分系数[17],改进的划分系数[18],XB 指标[19],PBMF 指标[20] 和 ZWJ 指标[21],对比发现近年提出的 PBMF 指标和 ZWJ 指标可以较好的反应聚类的质量,本文以两者的乘积作为聚类的综合评价指标。其中,划分系数定义为:

$$V_{pc}(U; c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{18}$$

改进的划分系数:

$$V'_{pc}(U; c) = \frac{1}{n}\sum_{j=1}^{n}\left(\max_{i=1}^{c} u_{ij}\right) \tag{19}$$

XB 指标是目前广泛使用的指标,其定义为:

$$V_{XB}(U, V, c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \|v_i - x_j\|^2}{n \times \min_{i \neq j}\|v_i - v_j\|^2} \tag{20}$$

PBMF 指标是最近提出来的较好的聚类有效性指标,其定义形式如下:

$$V_{PBMF}(U, V, c) = \frac{1}{c} \times \frac{E_1 \times \max_{i \neq j}\|v_i - v_j\|}{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \|x_j - v_i\|} \tag{21}$$

ZWJ 指标:

$$V_{zwj}=\frac{\sum_{i=1}^{c}\left[\frac{1}{n_i}\sum_{x_j\in v_i}^{n}(u_{ij})^m\|x_j-v_i\|^2\right]}{\frac{1}{c}(\sum_{i=1}^{c}\|v_i-v_0\|^2)}+\left[1-\frac{1}{n}\sum_{j=1}^{n}(\max_{i=1}^{c}u_{ij})\right] \quad (22)$$

其中 $v_0=\frac{1}{c}\sum_{i=1}^{c}v_i$, $\sum_{i=1}^{c}\|v_i-v_0\|^2$ 反映类间总变差, $\sum_{i=1}^{c}\left[\frac{1}{n_i}\sum_{x_j\in v_i}^{n}(u_{ij})^m\|x_j-v_i\|^2\right]$反映类内总差变,$1-\frac{1}{n}\sum_{j=1}^{n}(\max_{i=1}^{c}u_{ij})$ 包含改进的划分系数,反映划分结果是否分明,其值愈小,则划分愈分明,同时该值也可做为校正值使用。

通过对比分析,本文选定表现较优的后两个指标的乘积作为本文的综合指标。即 $V_{ot}=V_{PBMF}\times V_{zwj}$,该指标越小,聚类效果越好。

2.3　基于关键点的回贴时间间隔序列聚类

本文首先采用了文献 [22] 中的关键点选择算法 KPSegmentation(key points segmentation) 对时间间隔序列进行了关键点的抽取。该方法将时间序列中重要的信息点全部提取出来,有效地去除了时间序列中的噪声与重要性小的数据点。设 $X_1=(x_{11},x_{12},\cdots,x_{1n})$ 与 $X_2=(x_{21},x_{22},\cdots,x_{2n})$ 为两条原始时间序列,则通过 KPSegmentation 算法处理后,将得到两条时间关键点序列 $X'_1=(x_{m1},x_{m2},\cdots,x_{ms})$ 与 $X'_2=(x_{n1},x_{n2},\cdots,x_{np})$。

在找出时间序列的关键点之后,需对时间序列间求相似度。研究中本文采用欧氏距离这一最常用的度量方法计算时间序列之间的相似度。该方法具有运算速度快,复杂度低的优点。对于 n 维空间的两个点,它们之间的欧氏距离为:

$$d=\sqrt{(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+\cdots+(x_{1n}-x_{2n})^2} \quad (23)$$

关键值时间序列 X'_1 与 X'_2 在大多数情况下是不等长的,也就无法直接求出两者间的欧氏距离。本文采用了对两者的时间点取并集,只留下在关键点序列 X'_1 与 X'_2 上都存在的时间点上的数据,这样再次简化后的时间序列就变为等维的了。之后便可使用上式即可求得二者之间的欧氏距离,作为两条时间序列相似程度高低的评判依据。最后,在获取任意两个时间序列的相似度后,采用基于 FCM 聚类算法进行聚类分析。

2.4　论坛热帖预测

2.4.1　热帖预测属性

本文利用数据挖掘的方法来对高校 BBS 中的热帖进行预测。鉴于网络中帖子的变化是一非线性运动,成为热帖需要一个过程,因此,在研究中本文并没有关注帖子的内容,而是在分析了大量数据的基础上,从帖子参与者和回复帖子的属性上提取了 7 个特征作为热帖预测属性。

1) 发帖人影响力　研究发现在论坛中影响力较大的用户的帖子往往容易引起大家的关注,因此,发表的帖子是否会成为热帖是和发帖人有一定的联系。本文从两个方面来定义用户的影响力,根据用户聚类的等级 $UserClass(i)$ 和回复网络拓扑结构中用户影响力的评定系数 $PageRank(i)$ 的乘积作为综合评定指标; 则用户 i 的综合影响力:

$$Influence(i)=UserClass(i)\times PageRank(i) \quad (24)$$

2) 回复者平均影响力　帖子参与者的影响力也会影响到帖子的发展,活跃的回复者往往会推动帖子的快速发展。和发帖人影响力的影响力计算相同,本文将参与帖子 j 的回复者的平均影响力作为一个特征;

$$Avg\mathrm{Re}Influence(j)=\frac{1}{|U_i|}\sum_{i\in U_j}UserClass(i)\times PageRank(i) \quad (25)$$

其中,U_i 代表了参与帖子 j 的前 n 个回复帖的用户集合,$|U_j|$ 代表了用户的数目。

3) 用户交互率　通过对帖子的观察分析,一些热门帖子往往会引起大家的讨论和意见的交流,从而用户之间的引用和交互的频率较大。话题 j 的参与者之间的交互率定义为:

$$InteractiveRate(j)_t=\frac{QuoteNum(j)_t}{ReplyNum(j)_t} \quad (26)$$

其中,$QuoteNum(j)_t$ 代表在 t 时刻帖子 j 参与者之间的引用次数,$ReplyNum(j)_t$ 表示在 t 时刻帖子 j 的总共回复数量。

4) 回复率　属性 $R(\Delta t)_i$ 反映了帖子在一段时间内的发展变化,热帖一般会在短时间内有较大的波动。

$$R(\Delta t)_i=\frac{R(t_1)_i-R(t_0)_i}{t_1-t_0}\quad t_1\geqslant t_0 \quad (27)$$

$R(t_1)_i$ 表示帖子 i 在 t_1 时刻的回复量,$R(t_0)_i$ 表示帖子 i 在 t_0 时刻的回复量。

5) 平均回帖时间间隔　热帖会在相对短的时间内吸引大家的关注,表现在前 n 帖的平均回复时间间隔与普通帖相比会相对较小。定义为:

$$ReplyInterval(n)_j=\frac{T(n)_j-T(0)_j}{n} \quad (28)$$

其中,$T(n)_j$ 表示帖子 j 的回复量达到 n 的时刻。本文选取帖子的前 20 回帖的平均回复时间间隔作为预测指标。

6) 系统回复密度　参照文献 [23] 定义系统的回复密度,此值越大说明当时 BBS 系统有大量的用户在关注最新发表的帖子。BBS 系统越有可能存在新的热帖。

$$RM(t)=[rb(\Delta t)+ob(\Delta t)]\times\frac{\lg 2[rb(\Delta t)+ob(\Delta t)+1]}{rb(\Delta t)+ob(\Delta t)+1} \quad (29)$$

其中,$ob(\Delta t)$ 表示 Δt 时间内到达 BBS 系统的主贴数量,$rb(\Delta t)$ 表示 Δt 时间内到达 BBS 系统的新帖的回帖数量。

7) 回帖占有率　参照文献 [23] 定义某话题的回复集中度,此值越大说明当时 BBS 系统有大量的用户在关注该帖子,该话题越有可能成为新的热帖。

$$RP(\Delta t)_j=\frac{rb_j(\Delta t)\times\lg 2[orb(\Delta t)+1]}{orb(\Delta t)+1} \quad (30)$$

其中,$orb(\Delta t)$ 表示 Δt 时间内被回复过的新帖的数量; $rb_j(\Delta t)$ 表示 Δt 时间内帖子 j 的回复量。

本文将第 3,第 4,第 6 和第 7 属性的观测时间设定为主帖发帖后的 20 分钟。

2.4.2　分类器

本文中选用 KNN、C4.5 和基于 Adaboost 的集成分类器作为实验分类器。

1) KNN(K 最近邻分类算法) : KNN 是一种基于样本密度评估的简单机器学习分类算法。算法采用向量空间模型来分类,概念为相同类别的样本,彼此的相似度高,而可以借由计算与已知类别样本的相似度,来评估未知类别样本可能的分类。

2) C4.5(决策树算法) : 决策树算法是一种逼近离散函数值的方法。它是一种典型的分类方法,首先对数据进行处理,利用归纳算法生成可读的规则和决策树,然后使用决策对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程。

3) Adaboost: Adaboost 是一种迭代算法,其核心思想是针对

同一个训练集训练不同的分类器(弱分类器),然后把这些弱分类器集合起来,构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的,它根据每次训练集之中每个样本的分类是否正确,以及上次的总体分类的准确率,来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练,最后将每次训练得到的分类器最后融合起来,作为最后的决策分类器。使用 Adaboost 分类器还可以排除一些不必要的训练数据特征,并将重点放在关键的训练数据上面。

本文选用 C4.5 作为基分类器,通过 AdaBoost 算法进行多轮迭代,每次迭代增加错分样本的权重,最终通过投票产生强分类器。

2.4.3 评价指标

本文选择的精度指标、召回率指标,以及 F1 值的得分作为系统评价指标。这些指标定义如下:

$$precision = \frac{|\{truehotPosts\} \cap \{detectedPosts\}|}{|\{detectedPosts\}|} \quad (31)$$

$$recall = \frac{|\{truehotPosts\} \cap \{detectedPosts\}|}{|\{truehotPosts\}|} \quad (32)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (33)$$

其中 $\{truehotPosts\}$ 是热帖的集合,$\{detectedPosts\}$ 是所有被检测帖子的集合。在本文中,帖子被分类成热帖和普通帖子。

3 实验与分析

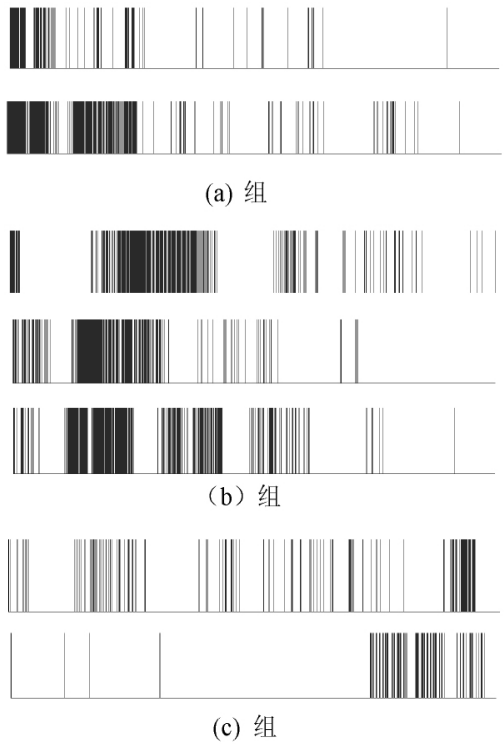
3.1 时间间隔幂律分布和聚类

表 2 时间间隔幂律分布

版面	该版最长贴幂律	全版块幂律
燕赵情怀	1.4783	1.6104
情感天空	2.1366	2.1471
足球吧	1.6406	2.3932
缘来如此	1.8284	1.4890
笑口常开	2.0455	2.2304
谈天说地	1.7849	2.5396

从表 2 可以看出,该论坛各个版的回复时间间隔大体满足 1.5 至 2.5 的幂律分布。结合长贴时间间隔序列的聚类结果,长贴的整体演变过程,可以分为以下几类:帖子的关注度会随着时间逐渐降低,表现在回帖的密集程度逐渐降低,如图 3 中(a)组,基于兴趣随时间不断衰减的模型[13]可以较好地解释此类帖子现象;(c)组中贴子的回复开始非常少,后面却引起了大量的回复,这种情况的帖子数量较少;大部分帖子的回复频率会随着时间先逐渐增加,后逐渐减少,如图 3 中(b)组。

人们的回帖行为与人们的兴趣和对事物的关注度有着密切联系,通过分析大量的数据,发现帖子被回复趋势并不都是随时间严格递减的,而已有的人类动力学模型[9,13,24,25]基于的假设都不能很好的解释该论坛回帖行为。



(a) 组

(b) 组

(c) 组

图 3 典型的长贴时间间隔序列模式
(其中一个竖线表示一次帖子回复)

3.2 用户聚类结果描述

为了对聚类结果进行有效评估,本文用 PBMF 指标和 ZWJ 指标的乘积作为有效性的综合指标 V_{ot}。根据两个指标的性质,可知当 V_{ot} 值越小聚类的效果越佳。本文总共做了 11 组实验,即聚类数 K 的实验范围定在 2 – 12 之间。表 3 是实验的结果。

表 3 聚类指标记录

聚类数	$V_{PBMF}(*10^{-3})$	V_{ZWJ}	$V_{ot}(*10^{-3})$
2	0.92878	1.1056	1.0278
3	1.4	0.5489	0.76846
4	2.0	0.3822	0.7644
5	2.2	0.3408	0.74976
6	2.3	0.3717	0.85491
7	2.3	0.3650	0.8395
8	2.5	0.3658	0.9145
9	2.5	0.3782	0.9455
10	2.7	0.3775	1.01925
11	2.9	0.3816	1.10664
12	3.1	0.3829	1.18699

从图 4 可以看出,第 4 组,即聚类数为 5 时,综合有效性指标值最小,聚类的效果最佳。此时 5 个聚类中心属性值分布如图 5 所示。

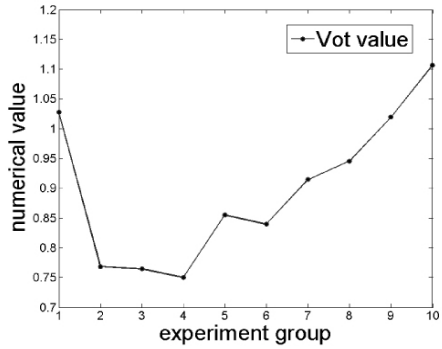


图 4　聚类指标变化曲线

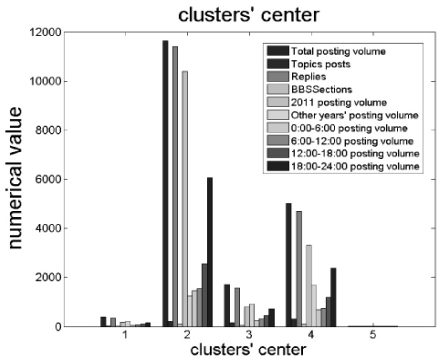


图 5　聚类中心

基于聚类结果,本文对每类用户进行了分析,并基于行为进行了用户论坛角色识别。结果如下:

1) 超级管理员类型　即站务人员,人数最少仅 11 人,他们的累计发帖量最多,回帖数量与主贴数量的比例约为 49。关注版面数的平均值接近 100 个,涉及了论坛 60% 以上的版块。在一天内不同时段发帖累计频率依次上升,且即使在 0 – 6 时这段时间,也保持了最大的累计发帖量。这一类用户属于论坛最活跃的用户。

2) 积极管理员类型　人数较少为 56 人,平均发帖数相较于第一类用户要减少一半,平均值接近 5000 贴,其中回帖数量与主贴数量的比例约为 15。关注的版面数为 85 个左右。一天之内发帖平稳,且在晚间较为积极。

3) 活跃用户与管理员混合类型　人数稍多达到 299 人。此类用户包含普通用户中最为活跃的一部分人,总发帖量在 1700 左右,回帖数量与主贴数量的比例约为 10。关注和参与的版面数在 54 个左右。

4) 稍活跃的普通用户　人数到达近 2000 人,占论坛用户人数 3.36%。这类用户较为活跃,参与的版面数达到 30 个左右,各个时间段发帖较为平稳。回帖数量与主贴数量的比例约为 9。

5) 普通用户　人数达到用户总数 96% 以上。他们的特点是参与关注的版数较少,仅为 5 个版块左右。发布和回复帖子的数量在 10 贴左右这个级别。这类用户的行为很不积极,极少参与话题讨论。回帖数量与主贴数量的比例约为 3.76。

从以上分析中,我们发现论坛的 96% 注册用户的活动并不积极,少数用户创造了论坛上的大部分信息。为了验证角色识别的有效性,参照建立的用户等级数据库,对推测的用户角色进行验证: 11 个超级管理员类型账户全部为站务账户,而其他 3 个站务账号被划分到了管理员级别。第二类和第三类总共包含了 213 名真实管理员和 142 名行为活跃的普通用户。第四类和第五类为普通用户,正确率达到 99% 以上。充分说明了基于用户行为聚类的角色识别的有效性。

3.3　论坛热帖预测结果

首先,本文将 9 组实验结果做了一个对比,结果如图 6 所示,包括 AdaBoost 集成分类器,C4.5,KNN 算法以及基于主成分分析后的实验结果。本文采用主成分分析试图发现向量空间降维后对分类预测的影响。例如,图 6 中 3 – C4.5 代表在数据降到 3 维后应用 C4.5 算法的分类预测效果。

从图 6 中可以看出,热帖的预测识别可以达到较高的准确率和召回率。其中,AdaBoost 和 C4.5 在所有的实验组中均保持了较高识别效率。AdaBoost 集成分类器在所有组实验中表现得最优,充分说明了采用基于 AdaBoost 的集成分类器的有效性。C4.5 算法对于本文的分类是一种具有健壮性的可选用的分类器,虽然在实验效果上比 AdaBoost 显的略差。在使用主成分分析之后,所有的分类器的效果都开始变差。

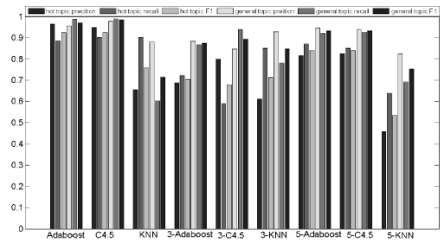


图 6　论坛热帖预测的 9 组实验结果

之后,本文分析了不同类别的属性对分类预测结果的影响,即帖子参与者的特征属性和帖子的特征属性,实验结果如表 4、表 5 所示。从表中我们可以看出帖子的特征属性是预测帖子热度的重要指标,而帖子参与者的特征在热帖演变过程中也起到了重要的作用。

表 4　帖子参与者的属性特征实验结果

	AdaBoost	C4.5	KNN
热帖准确率	0.727	0.684	0.443
热帖召回率	0.656	0.639	0.705
热帖 F1 值	0.69	0.661	0.544
普通帖准确率	0.865	0.856	0.841
普通帖召回率	0.899	0.879	0.638
普通帖 F1 值	0.882	0.868	0.725

表 5　帖子的特征属性实验结果

	AdaBoost	C4.5	KNN
热帖准确率	0.85	0.825	0.5
热帖召回率	0.836	0.852	0.738
热帖 F1 值	0.843	0.839	0.596
普通帖准确率	0.933	0.939	0.867
普通帖召回率	0.94	0.926	0.698
普通帖 F1 值	0.936	0.932	0.773

另外,同吴焕政等提出的方法[7]相比,尽管两者用到的数据集不一样,但都是针对 BBS 论坛热帖展开的研究,结论具有一定相似性,与文献[7]中对于天涯杂谈中的突发舆情帖子预测的准确率 0.467 和召回率 0.667 相比,本文对于高校 BBS 热帖的预测的准确率 0.933 和召回率 0.883 均取得了一定的进步。

4 结 语

校园 BBS 是高校网络舆论的主要载体，反应了大学生的舆论倾向以及生活的各个方面，高校 BBS 的舆情研究具有重要的意义。本文针对高校 BBS 热帖预测问题，提出的基于用户行为的热帖预测模型。该模型以一个高校 BBS 的实际数据为研究对象，从用户和贴子的特征分析入手，研究发现帖子回复时间间隔服从幂律分布，通过基于用户行为的聚类研究，对用户的论坛角色进行了有效的识别，并从用户和帖子两个角度，提出了 7 种帖子热度预测的特征属性，结合集成分类器建立了预测模型，通过实验证明了对高校 BBS 热帖预测的有效性。

本文下一步将以多个 BBS 为研究对象，如果把研究对象扩展为多个 BBS 将更能体现 BBS 热帖预测的普遍意义，并结合一定的文本分析技术，进一步提高热贴预测的效率。

参 考 文 献

[1] 蒋研川，肖铁岩，凌晓明，等. 新媒体环境下高校校园网络舆论的现状及引导策略研究[J]. 重庆大学学报：社会科学版，2012，18(1)：136-142.

[2] 曾祥平，方勇，等. 基于元胞自动机的网络舆论激励模[J]. 计算机应用，2007，27(11)：2686-2688，2714.

[3] Naruse K M. Lognormal distribution of BBS article and its social and generative mechanism[C]//Web intelligence: proceedings of the 2006 IEEE/WIC/ACM international conference on Web intelligence. Washington: IEEE Computer Society，2006：103-112.

[4] Green D G，Leishman T G，Sadedin S. The emergence of social consensus in boolean networks[C]//Proceedings of the 2007 IEEE symposium on Artificial life. Washington: IEEE Computer Society，2007：402-408.

[5] Zeng X P，Zhang S Y，Wu C Y. Predictive model for internet public opinion[C]//Fourth international conference on fuzzy systems and knowledge discovery(FSKD 2007). Washington: IEEE Computer Society，2007：7-11.

[6] 高俊波，安博文，王晓峰，等. 在线论坛中潜在影响力主题的发现研究[J]. 计算机应用，2008，28(1)：140-142.

[7] 吴焕政，吴渝，肖开州. 基于粗糙集和集成学习的 BBS 网络舆情分类[J]. 广西大学学报：自然科学版，2009，34(5)：696-699.

[8] Haight F A. Handbook of the Poisson distribution[M]. New York，1967.

[9] Barabasi，Albert-Laszlo. The origin of bursts and heavy tails in human dynamics[J]. Nature，2005，435(7039)：207-211.

[10] Dezso Z，Almaas E，Lukacs A，et al. Dynamics of information access on the web[J]. Phys. Rev. E，2006(73)：066132.

[11] Han X P，Zhou T，et al. Heavy-tailed statistics in short-message communication[J]. Chinese Physics Letters，2009，26(2)：028902.

[12] Yan Q，Yi L，Wu L. Human dynamic model co-driven by interest and social identity in the microblog community[J]. Physica A: Statistical Mechanics and its Applications，2012，391：1540-1545.

[13] 郭进利. 博客评论的人类行为动力学实证研究和建模[J]. 计算机应用研究，2011，28(4).

[14] Zhou T，Kiet H，Kim B，et al. Role of activity in human dynamics[J]. Europhys Lett，2008，82：28002-28006.

[15] Clauset A，Shalizi C R，Newman M E J. Power-Law Distributions in Empirical Data[J]. SIAM Review，2009，51(4)：661-703.

[16] 齐淼，张化祥. 改进的模糊 C 均值聚类算法研究[J]. 计算机工程与应用，2009，45(20)：133-135.

[17] Bezdek J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics，1974，3(3)：58-72.

[18] 范九伦，吴成茂，丁夷. 基于样本最大分类信息的聚类有效性函数[J]. 模糊系统与数学，2001，15(3)：69-74.

[19] Xie X L，Beni G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)，1991，13(8)：841-847.

[20] Pakira M K，Bandyopadhyay S，Maulik U. Validity index for crisp and fuzzy clusters[J]. Pattern Recognition，2004，37：487-501.

[21] 朱文婕，吴楠，胡学钢. 一个改进的模糊聚类有效性指标[J]. 计算机工程与应用，2011，47(5)：206-206.

[22] 杜奕，卢德唐，李道伦，等. 一种快速的时间序列线性拟合算法[J]. 中国科学技术大学学报，2007，37(3)：310-314.

[23] 杨梅. 网络舆情热点发现的研究[D]. 北京：北京交通大学，2008.

[24] 邓竹君，张宁，李季明. 截止时间对人类动力学模型的影响[M]//郭进利，周涛，张宁，等. 人类行为动力学模型. 香港：上海系统科学出版社，2008：29-34.

[25] 戴双星，陈冠雄，周涛，等. 兴趣驱动的人类动力学模型研究[M]//郭进利，周涛，张宁，等. 人类行为动力学模型. 香港：上海系统科学出版社，2008：4-58.

(上接第 47 页)

种不同并行模式实现的算法进行了相应的比较，通过分析结果可以发现在数据量较大时基于 MapReduce 模式的并行算法处理效果更为理想。

尽管通过本文中的一系列对比试验中可以发现，本文提出的算法在分类准确性上相比于其它算法有所提高，但仍有进一步提高的空间，另外在对比实验中发现并行化的方法分类效果有所降低，并且对于不同数据量规模的数据，种并行模式的完成时间略有差异，如何充分发挥两种并行框架的优势，提高并行算法的性能是以后工作研究中的重中之重。

参 考 文 献

[1] Sebastiani F. Text Categorization[M]. Encyclopedia of Database Technologies and Applications，2005：683-678.

[2] Su Jinshu，Zhang Bofeng，Xu Xin. Advances in Machine Learning Based Text Categorization[J]. Journal of Software，2006，17(9)：1848-1859.

[3] Yang Y. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval[C]//SIGIR-94，1994.

[4] Cortes C，Vapnik V. Support vector networks[J]. Machine Learning，1995，20：273-297.

[5] Han E H，Karypis G. Centroid-based document classification algorithms: Analysis & experimentalresults[R]. Technical Report TR-00-017，Department of Computer Science，University of Minnesota，Minneapolis，2000.

[6] Tan S. An improved centroid classifier for text categorization[J]. Expert Systems with Applications，2008，35(1-2)：279-285.

[7] Tan Songbo，Cheng Xueqi. An Effective Approach to Enhance Centroid Classifier for Text Categorization[C]//11th European Conference on Principles and Practice of Knowledge Discovery in Databases，Proceedings：581-588.

[8] Shankar S，Karypis G. Weight Adjustment Schemes for a Centroid Based Classifier[R]. Army High Performance Computing Research Center，2000.

[9] Tom White. Hadoop: The Definitive Guide[M]. O'Reilly Media，2009.

[10] Dean J，Ghemawat S. MapReduce: simplified data processing on large clusters[C]//OSDI'04: Sixth Symposium on Operating System Design and Implementation，San Francisco，CA，December，2004：107-113.

[11] Edward J Yoon. Apache Hama (v0.2): User Guide—a BSP-based distributed computing framework[EB]. Apache.