软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2767−2776 [doi: 10.13328/j.cnki.jos.004730]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于动态演化的讨论帖流行度预测*

孔庆超, 毛文吉

(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)

通讯作者: 毛文吉, E-mail: wenji.mao@ia.ac.cn

摘 要: 互联网用户间的交互行为,使得某些用户生成的内容(如讨论帖、微博话题)变得流行.对所关注内容的流行度进行建模和预测,在多个领域中具有十分重要的研究和应用价值.针对论坛讨论帖的流行度预测问题,基于早期的发展演化信息,探讨了影响讨论帖流行度的相关动态因素,并提出一种结合局部特性、融合多个动态因素的讨论帖流行度预测算法.以豆瓣小组的数据为例,对所提出的算法进行实验.实验结果表明,所提出的融合多种动态因素的方法与基准方法相比,能够较好地预测讨论帖的流行度.

关键词: 用户生成的内容;内容流行度;流行度预测;社会媒体分析;动态演化建模与预测

中图法分类号: TP181

中文引用格式: 孔庆超,毛文吉.基于动态演化的讨论帖流行度预测.软件学报,2014,25(12):2767−2776. http://www.jos.org.cn/1000-9825/4730.htm

英文引用格式: Kong QC, Mao WJ. Predicting popularity of forum threads based on dynamic evolution. Ruan Jian Xue Bao/ Journal of Software, 2014,25(12):2767−2776 (in Chinese). http://www.jos.org.cn/1000-9825/4730.htm

Predicting Popularity of Forum Threads Based on Dynamic Evolution

KONG Qing-Chao, MAO Wen-Ji

(State Key Laboratory of Management and Control for Complex Systems (Institute of Automation, The Chinese Academy of Sciences), Beijing 100190, China)

Corresponding author: MAO Wen-Ji, E-mail: wenji.mao@ia.ac.cn

Abstract: Web user's online interacting behavior with others usually makes some user generated content (e.g. forum threads and Weibo topics) popular. The modeling and prediction of the popularity of online content are of great research importance and practical value in many different domains. To predict the popularity of forum threads, this paper discusses several dynamic factors that might affect the popularity of online content based on the information of dynamic evolution at the early stage, and proposes a popularity prediction algorithm which makes use of the locality property and combines multiple dynamic factors. The proposed algorithm is further evaluated with the Douban group dataset. The experimental results show that, compared with the baseline methods, our method achieves relatively better performance in predicting the popularity of forum threads.

Key words: user generated content; popularity of online content; popularity prediction; social media analytics; modeling and prediction of dynamic evolution

近年来,互联网应用更强调用户与用户之间的互动以构成虚拟的在线社交网络.其中,典型的媒体形式包括关于各种主题的网络论坛、微博和社交网站等.用户之间的交互行为使得某些用户生成的内容成为受到关注的热门内容,拥有了较高的流行度(popularity).例如,论坛用户通过发帖表达对某个主题的关注,其他用户可以参与评论,进而形成互动的讨论组.当某个帖子被浏览或评论足够多次数时,则成为流行度较高的热帖.

由于互联网具有实时性和交互性的特点,所关注内容的流行度通常随时间变化明显,呈现出较强的动态演

* 基金项目: 国家自然科学基金(61175040, 71025001)
收稿时间: 2014-05-06; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

化特征[1,2].对所关注内容的流行度进行动态建模、分析和预测具有十分重要的研究和应用价值.在安全相关领域,对内容流行度的动态建模、分析和预测可以及时了解互联网舆情信息,有效把握网络化社会态势并有力支持安全预警和辅助决策.在经济和商业领域,准确预估网站内容的流行度可以及时了解用户需求和喜好,帮助商家更合理地进行商品推荐和广告投放[3].此外,针对重点关注的网络媒体内容进行流行度分析预测,还可以为政府部门决策和社会与公共管理提供重要依据.

由于网络论坛具有讨论主题丰富、内容覆盖广泛、用户数量庞大等特点,一直是最为活跃的反映社会热点话题的网络媒体形式之一.近年来蓬勃发展的社交网站,如人人网和豆瓣网,也多将论坛作为其重要组成部分.本文基于对论坛讨论帖早期的动态演化过程分析,提出一种预测讨论帖在未来某一时段后流行度的方法.我们根据讨论帖早期的流行度及其发展变化,定义了多个影响内容流行度的动态因素,并尝试融合多个动态因素预测讨论帖的流行度.为了验证所提出方法的有效性,我们以豆瓣小组的讨论帖数据为例进行了实验验证.与基准方法相比,我们提出的方法取得了比较好的预测效果.

本文的贡献包括以下方面:

(1) 首次针对讨论帖的流行度预测问题展开研究,并给出了该问题的明确定义;
(2) 提出两种新的动态特征,即评论结构特性和用户回复关系;
(3) 提出融合多个动态因素的流行度预测算法,利用局部特性计算各个动态因素的先验并改进 *kNN* 算法,有效提升了预测效果.

1 相关工作

1.1 流行度定义

目前,关于流行度的研究对象主要集中于在线视频[3−12]、微博[13−17]、话题标签(hashtag)[18,19]、Digg 链接分享[20,21]、图像[22,23]等.内容流行度的定义通常与具体应用相关.以往的研究工作通常将内容流行度定义为某种数量,如视频浏览数、微博评论或转发数、话题标签的出现次数等.其中,Ma 等人[14]研究并比较了微博的转发数和浏览数两种流行度定义,发现尽管两者之间存在正相关关系,但是这种相关关系并不强,因此建议将浏览数和转发数当作两种流行度的度量来研究.在描述 Digg 分享的流行度时,Yin 等人[9]并没有直接使用总得票数,而是综合考虑了正面和负面得票数.类似地,Khosla 等人[22]为了更好地刻画图像的流行度,考虑了其浏览数随时间逐渐增长的特点,将流行度定义为当前浏览数与其上传时间的比值.

本研究工作以论坛讨论帖为研究对象,由于相较于浏览数,讨论帖的评论数更能体现用户关注情况,因而,本文将讨论帖的评论数作为流行度的度量.

1.2 流行度预测

Szabo 和 Huberman[20]对大量 Digg 分享和 Youtube 视频流行度的研究发现,早期的流行度和一段时间后的流行度之间呈现出简单的线性关系(比例系数记为α),并由此提出 S-H 模型.作为较早的流行度预测模型, S-H 模型非常简洁,并且在大规模真实数据集中保持了不错的有效性.但是,S-H 模型仍有明显的缺陷:首先,S-H 模型只考察所关注内容早期和一段时间后的流行度的相关关系,没有考虑更丰富的特征;其次,一些所关注内容在早期具有相近的浏览数,而一段时间后它们的流行度可能相差很大[12],S-H 模型无法处理这种情况,因为模型假设所有的样本共享相同的比例系数α;最后,尽管 S-H 模型在大规模真实数据集中表现良好,但其主要原因是数据集中不同对象的流行度的分布极不均衡[24],而 S-H 模型只是对数据集中流行度较低的内容预测效果较好,反之亦然.

为了克服以上 S-H 模型的缺陷,近期的研究工作充分利用了具体问题情景下更丰富的特征.如果按照特征的使用方法分类,近期的工作主要可以分成两类:基于静态特征的方法[3-5,7,9,13−17,19,21]和基于动态特征的方法[10−12,18].

基于静态特征的方法通常通过寻找可能跟流行相关的因素,然后训练回归模型,最后用得到的模型来预测

所关注内容未来的流行度.例如,为了预测在线视频的流行度,Figueiredo[3]考察了内容特征,包括视频的分类、上传时间、视频已存在的时间等,以及链接分享信息,包括首次被分享的时间和分享的观看次数等;为了预测Twitter上微博的流行度,Hong等人[17]使用的特征包括内容特征、转发图的拓扑特征、时序特征以及其他元信息特征;此外,Yin 等人[9]考察用户在投票时与大众保持统一或者相反两种心理,对所关注内容的流行度建立生成式模型;其他采用生成式模型的文献还包括文献[5,13,14,21].

由于所关注内容的流行度通常随时间变化明显,呈现出较强的时序动态特征,因而相较于静态方法,以刻画随时间变化的演化过程为特点的动态方法近几年成为流行度预测的重要方法.例如,对于某个关注的内容,Ahmed 等人[10]分别使用相对于其他关注内容流行度的比例的变化和自身流行度变化两个特征来描述各个所关注内容间的相似程度,并在不同的时间段进行聚类,然后分析对象在不同时间段所在的聚类簇之间的转换关系;考虑到在社交网站上用户的分享行为能够极大地影响视频的流行度,Li 等人[11]主要考虑浏览数和视频分享率两个特征,然后基于这两者随时间演化的情况预测未来一段时间后的视频流行度;Pinto等人[12]则直接使用视频在不同时间段的浏览数相较于前一个时间段的增量作为特征,训练回归模型预测未来的浏览数.

基于动态特征的方法能够较好地反映流行度演化过程中随时间变化的特性,但目前,相关工作中所基于的动态因素主要是与数量相关的因素,如评论数、用户数、分享率等,缺乏与所关注对象和用户相关的结构特征和关联因素.此外,目前绝大多数相关工作仅预测所关注内容的浏览数量,其所采用的方法在数据分布上往往存在不均衡性,这使得现有工作存在与 S-H 模型类似的缺陷.缺乏描述流行度分布的客观全面的问题定义,也带来了测试标准上的种种偏差.

本文针对讨论帖的流行度预测问题,首先给出了关于该问题的明确定义.在动态特征的构建上,不仅采用了评论数和参与用户数,还利用了评论内在的结构特性和参与用户间的关系描述等因素.在流行度定义的基础上,我们将流行度预测问题转化为时间序列分类问题.由于对时间序列分类问题而言,*kNN* 是相对简单而且有效的方法[25],本文在 *kNN* 算法的基础上,利用局部特性计算各个动态因素对样本的影响大小,进而提出一种融合多个动态因素的讨论帖流行度预测算法.

2 基于动态演化的流行度预测

2.1 问题定义

基于以上考虑,我们给出本研究工作中对流行度预测问题的定义:给定一个讨论帖,以讨论帖的发表时间为0时刻,记在 t 时刻的评论数为 $C(t)$,即讨论帖在 t 时刻的流行度.在 t_r 时刻时,模型将讨论帖当前的演化情况作为输入,预测 $t_t(t_t>t_r)$时刻讨论帖是否流行,如图 1 所示.具体来说,令:

$$r=\frac{C(t_r)}{C(t_t)} \tag{1}$$

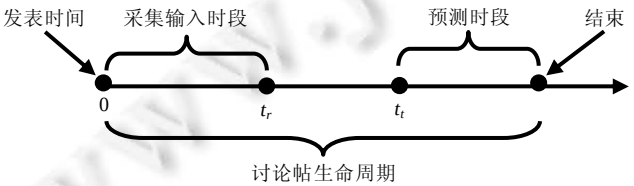

Fig.1 Thread life cycle and description of t_r and t_t

图 1 讨论帖生命周期及 t_r 和 t_t 变量含义

如果 $r<p$(p 为预先设定的阈值,满足 $0<p<1$),说明此讨论帖在 t_r 时刻后仍会有相当数量的评论,可认为其在未来一段时间后会流行(记为 L_1 类);否则,认为其在未来一段时间后不会流行(记为 L_2 类).在实际应用中,我们可以限制上述问题定义中讨论帖分别在 t_r 和 t_t 时刻的最少评论数量 $\min_C(t_r)$和 $\min\_C(t_t)$,以保证有足够的评论信

息用于该预测问题.

2.2 动态因素的构建

讨论帖在一段时间后能否流行,主要依赖于静态和动态两方面因素,其中,静态因素包括讨论帖内容、发表时间、作者等相对不变的信息,而动态信息是在流行度的动态演化过程中起决定作用的因素.考虑到流行度预测问题的研究对象(即评论)和生成评论内容的主体(即用户)的重要性,我们在构建动态因素时主要采用了评论及其参与用户的数量和结构方面的信息,后者包括由评论形成的评论树(comment tree)结构和由参与评论的用户形成的回复图(reply graph)结构.具体来说,动态因素包括如下几个方面:

- 评论数和参与评论的用户数.

需要说明的是,我们此时记录的是每个采样时间内新增的评论数和参与评论的用户数,而不是累计评论数和参与评论的用户数.另外,我们参照文献[18]中的方法对两个动态因素对应的时间序列进行变换.

- 评论树的结构特性.

首先构造评论树:将原讨论帖看作根节点 Root:

1) 如果有评论 *A* 直接回复原帖,则新增节点 *A* 并创建链接使得 *A* 指向 Root;
2) 如果有评论 *A* 回复评论 *B*(评论 *B* 节点已经存在且非 Root 节点),那么新增节点 *A* 并创建链接使得 *A* 指向 *B*,如图 2 所示.

我们将评论树的结构特性作为动态因素,如评论树深度(即根节点到叶节点的最长路径长度)和平均节点间路径长度.上述两个结构特性分别描述评论树的两个方面:评论树的深度越大,说明用户间的讨论越深入,表明讨论帖能够吸引用户参与讨论,进而提高自身的流行度.平均节点间路径长度在相关文献中也定义为 Wiener 系数[23],在我们的方法中用于刻画评论树的相对平衡程度.

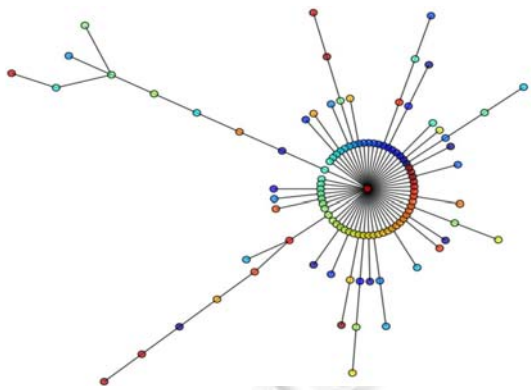

Fig.2 Comment tree

图 2 评论树

- 回复图的结构特性.

与评论树不同,回复图的节点是用户,其构造方法如下:讨论帖的作者是回复图的第 1 个节点,记作 *FN*:

1) 如果用户 *A* 发表评论直接回复原帖,则新增节点 *A* 并创建链接使得 *A* 指向 *FN*;
2) 如果用户 *A* 发表评论回复的是用户 *B* 的某个评论,那么新增节点 *A* 并创建链接使得 *A* 指向 *B*.

对于回复图,我们考虑的结构特性有链接密度(link density)和平均度(mean degree)等.回复图的链接密度和平均度越高,表明用户节点间交互越频繁,也就意味着该讨论帖的流行度会更高.

随着评论数和参与评论用户的增加,评论树和回复图的结构都在不断发生变化:节点数和链接数逐渐增加,两者的结构特性也在变化,即动态因素不断变化.此外,在记录动态因素的信息时,我们采用的策略是:从讨论帖发表开始,每隔一定的时间(称为采样时间)便记录各个动态因素的值.这样,最终我们得到的是一系列间隔一定

时间的动态因素的值,即一个时间序列.如果考虑多个动态因素,则每个样本对应一个多维时间序列.

2.3 IPW算法

如果只考虑单个动态因素,那么对应每个讨论帖的则是一个由该动态因素组成的时间序列.此时,流行度预测问题便转化为经典的时间序列分类问题.根据本文第 1 节分析,我们采用 *kNN* 算法进行分类,其中,时间序列间的相似度采用欧几里德距离来定义.

显然,单个动态因素只能描述讨论帖动态演化的某一个方面,我们的目标是融合多个动态因素以提高预测准确率.对于单个动态因素预测结果的分析表明,不同动态因素对于不同类型样本的分类结果影响不同.即:对于不同类型的样本,不同的动态因素的分类效果不同.基于以上考虑,我们改进 *kNN* 算法,提出 IPW(instance prior weighting)算法.

IPW 算法包括训练和测试两个过程:

- 首先,通过训练过程得到不同动态因素对不同训练样本分类结果的影响大小,我们用 IP 矩阵(*IP_matrix*)来描述这种影响(具体构建过程在下文中说明),其中,IP 矩阵中的数值表示影响作用的大小,如图 3 所示.

	样本 1	样本 2	…	样本 N
动态因素 1	P_{11}	P_{12}	…	P_{1N}
动态因素 2	P_{21}	P_{22}	…	P_{2N}
…	…	…	…	…
动态因素 M	P_{M1}	P_{M2}	…	P_{MN}

Fig.3 IP matrix

图 3 IP 矩阵

- 在测试过程中,对于一个测试样本,IPW 算法最终得到对于每一类的分类置信值,并据此给出分类结果.为了得到此分类置信值,首先分别考察在单个动态因素下使用 k 近邻方法进行分类,并根据距离的大小计算权重;然后,对于每个近邻,根据其所属的类别、权重以及对应 IP 矩阵中的项,计算此近邻对于测试样本的分类置信值的贡献;最后,对于每个类别,累加所有动态因素下的分类置信值.

为构建 *IP_matrix*,我们提出 IPCreate 算法,见表 1.IPCreate 算法具体过程是:分别对应于训练集合 T 中的每个样本 i 和动态因素集合 D 中的每个动态因素 m(第 1 步和第 1.1 步),在动态因素 m 下,使用加权 *kNN* 算法对样本 i 的 k 个近邻,其中,属于 L_1,L_2 类的近邻个数分别记为 n_1 和 n_2(第 1.1.1 步),并设 *diff* 为 n_1 和 n_2 之差的绝对值(第 1.1.2 步).根据 n_1 和 n_2 的大小对样本 i 进行分类(第 1.1.3 步),分别按照分类正确与否两种情况对 *IP_matrix* 中对应动态因素 m 和样本 i 的对应项进行设置(第 1.1.5 步和第 1.1.6 步):如果分类正确,表示此动态因素对于训练样本的正确分类影响较大,那么其在 *IP_matrix* 中对应的元素值将大于 1;否则小于 1.注意到,此值的大小跟 *diff* 相关:*diff* 值越大,说明该动态因素 m 对于此分类结果越自信,那么如果分类正确,显然动态因素 m 对于训练样本 i 的影响应该越大;反之越小.

IPW 算法过程详见表 2.首先调用 IPCreate 算法构建 *IP_matrix*(第 1 步).对于目标测试样本 o,我们用 *score* 来存储各个动态因素在类别 L_1 和 L_2 上的分类置信值之和,并将 *score* 初始化为[0,0](第 2 步).分别对于动态因素集中 D 中的动态因素 m(第 3 步),计算测试样本 o 与训练样本集 T 中每个样本的距离,并按照距离从小到大的顺序取前 k 个近邻(第 3.1 和 3.2 步).分别对于 k 个近邻中的每个样本 i(第 3.3 步):利用样本 i 与测试样本 o 在动态因素 m 下的距离计算样本 i 的权重 w_i(第 3.3.2 步);然后,从 *IP_matrix* 中取出对应动态因素 m 和样本 i 的先验信息 p_{mi}(第 3.3.3 步);最后,根据样本 i 的所属类别,累加动态因素 m 在该类上的分类置信值(第 3.3.4 步).IPW 算法最终将 *score* 向量中对应置信值较大的类别作为测试样本 o 的流行度预测结果(第 4 步).

需要说明的是,在 IPW 算法的第 3.3.5 步,如果 p_{mi} 的值为 1,即,不考虑动态因素对样本的影响,那么 IPW 算

法退化为带有权重的 *kNN* 算法;如果 p_{mi} 和 w_i 的值都为 1,意味着既不考虑动态因素对样本的影响也不考虑样本的权重,那么 IPW 算法则退化为无权重的 *kNN* 算法.

Table 1 IPCreate algorithm

表 1 IPCreate 算法

IPCreate 算法.

输入:
训练样本集合 *T*,
动态因素集合 *D*,
近邻个数 *k*;
输出:
IP 矩阵 *IP_matrix*.
1. FOR 样本 $i \in T$
  1.1. FOR 动态因素 $m \in D$
    1.1.1. 在动态因素 *m* 下,使用加权 *kNN* 算法找到样本 *i* 的 *k* 个近邻,其中,属于 L_1,L_2 类的近邻个数分别记为 n_1 和 n_2
    1.1.2. 令 *diff* 为 n_1 和 n_2 之差的绝对值
    1.1.3. 根据 n_1 和 n_2 的大小对样本 *i* 进行分类,记分类结果为 *pred*
    1.1.4. IF 样本 *i* 的所属类别和 *pred* 相一致
    1.1.5. THEN 令 *IP_matrix* 中动态因素 *m* 和样本 *i* 的对应项取值为 exp(+1***diff*)
    1.1.6. ELSE 令 *IP_matrix* 中动态因素 *m* 和样本 *i* 的对应项取值为 exp(−1***diff*)
  END-FOR
END-FOR
2. 返回 *IP_matrix*

Table 2 IPW algorithm

表 2 IPW 算法

IPW 算法.

输入:
目标测试样本 *o*,
训练样本集合 *T*,
动态因素集合 *D*,
近邻个数 *k*;
输出:
测试样本 *o* 的流行度预测结果.
1. 调用 IPCreate 算法(见表 1),得到 *IP_matrix*
2. 令 *score* 为二维向量,并初始化为[0,0].*score* 的作用为存储各个动态因素在类别 L_1,L_2 上的分类置信值之和
3. FOR 动态因素 $m \in D$
  3.1. 在动态因素 *m* 下,计算测试样本 *o* 与训练样本集 *T* 中每个样本的距离
  3.2. 按照距离从小到大的顺序取出前 *k* 近邻,记结果集合为 *K_list*
  3.3. FOR 样本 $i \in K\_list$
    3.3.1. 在动态因素 *m* 下,样本 *i* 与测试样本 *o* 的距离记为 d_i
    3.3.2. 利用 d_i 计算样本 *i* 的权重 $w_i=\exp(-1*d_i)$
    3.3.3. 从 *IP_matrix* 中取出动态因素 *m* 对样本 *i* 的先验影响值 p_{mi}
    3.3.4. 根据样本 *i* 所属类别 L_j,累加动态因素 *m* 在该类上的分类置信值,即 $score[j]=score[j]+w_i*p_{mi}$
  END-FOR
END-FOR
4. 将 *score* 向量中对应置信值较大的类别作为测试样本 *o* 的流行度预测结果

3 测试实验

3.1 数据集与实验设定

实验所用到的数据集来自两个豆瓣小组:buybook(“买书如山倒 读书如抽丝”)和 qiong(“穷游天下”).通常,每个豆瓣小组都会设定一个公共话题,例如,buybook 小组的讨论帖多是关于读书,而 qiong 小组的讨论帖多是关于旅游.小组的成员可以在小组内发布讨论帖,其他成员可以浏览、评论或者推荐给自己的“粉丝”等.

我们设计并实现了一个简易的爬虫程序,从以上两个小组的网页中抓取数据.为了在预测时有足够多的信

息,我们过滤掉在 t_r 时刻评论数少于 10 的讨论帖,并仅考虑在 t_t 时刻评论数不少于 50 的讨论帖,最终得到包括 2 300 个讨论帖样本的数据集.每个讨论帖样本数据包括讨论帖的发表者、发表时间、原帖内容、讨论帖中每个评论的发表者、发表时间、评论内容以及评论之间的回复关系.我们设定测试集和训练集大小的比例为 1:4,采样时间设为 5 小时;同时,为避免类不均衡问题,我们采用了下采样方法[26]使两类样本数大致相同.

图 4 是数据集中讨论帖评论数的分布情况,横坐标为评论数,纵坐标为讨论帖的数量.图 5 是数据集中讨论帖生命周期长度的分布情况,横坐标为讨论帖的生命周期长度(以天为单位),纵坐标为讨论帖的数量.

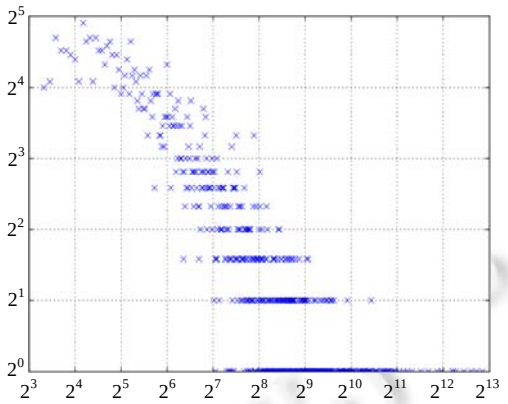

Fig.4 Distribution of number of comments

图 4 评论数的分布情况

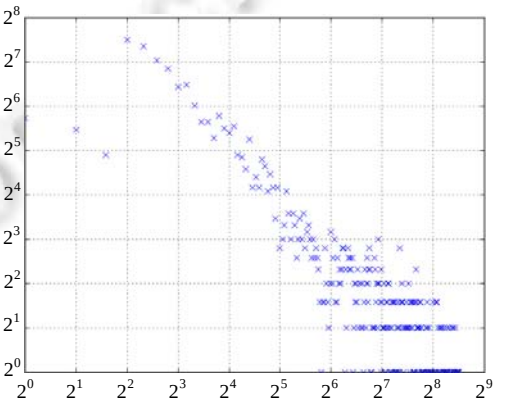

Fig.5 Distribution of life cycle of threads

图 5 讨论帖生命周期的分布情况

3.2 基准方法

在实验验证部分,首先考察不同的单个动态因素的分类效果,并使用 *kNN* 算法进行分类.另外,投票法[27]是一种简单且使用广泛的集成(ensemble)方法,用于综合各个不同动态因素的分类结果.作为早期的流行度预测模型,我们同样将 S-H 模型[20]作为基准方法.在使用动态特征的方法中,由于文献[10,11]都使用了具体问题相关的因素,所以本文的基准方法不考虑这两者,而只包含了 ML(multivariate linear)方法[12].

- *kNN* 算法:通过交叉验证的方法确定近邻数 k;
- 投票法[27]:先使用各个动态因素进行分类,然后根据分类结果进行投票,得票最多的那一类即为最终分类结果;
- S-H 模型[20]:先通过训练得到 t_t 时刻的评论数和 t_r 时刻的评论数之间的比值 α.在当前的问题设置下,根据 S-H 的模型假设,那么分类结果可以按照如下方法得到:如果 α 小于阈值 p,则所有的样本都会被归为不流行一类;否则,所有的样本都会被归为流行一类;
- ML 方法[12]:对于每个讨论帖,ML 方法首先记录在每个采样时间段内新增的评论数,记为 $N(i)$,其中,i 为采样时间段的顺序值,满足 $1\leqslant i\leqslant n$,n 为采样时间段的个数.ML 方法按照如下方式构造特征:

$$x(i):x(i)=N(i+1)-N(i).$$

详细的模型求解方法见文献[12].

3.3 实验结果及分析

本节比较使用单个动态因素的算法、IPW 算法以及基准算法的流行度预测性能,并考察参数的变化对基于 *kNN* 方法的流行度预测的影响.我们通过 3 折交叉验证将以下实验中基于 *kNN* 方法的 k 值选定为 5.本文提出的方法与基准方法相比较的实验结果见表 3(设定参数 p 为 0.6,t_r 为 40 小时,Δt 为 25 小时,其中 $\Delta t=t_t-t_r$).

Table 3 Experimental results

表 3 实验结果

方法名称		准确率 (%)
基于单个动态因素的 *kNN* 算法	评论数	56
	参与评论用户数	54
	评论树深度	53
	平均节点路径长度	53
	回复图连接密度	55
	平均度	56
投票法		56
S-H 方法		54
ML 方法		55
IPW 算法		**58**

从表 3 的实验结果中可以看出:单个动态因素的效果各不相同,且相对于基于单个动态因素的方法,投票法并未有效提升流行度预测的效果.对于 S-H 模型,由于所有的样本共享从数据集中计算得到的线性比例系数α,所以在当前的问题定义下,对于所有的测试样本只能得到相同的分类结果.ML 方法使用评论数的增量作为特征,能够较好地描述评论数的增长(或减少)趋势,所以其分类效果较好.与单个动态因素分类方法的缺点类似,ML 方法只是使用了单个动态因素——评论数,所以无法利用其他更为丰富的动态因素信息.在目前的参数设定下,相较于以上基准方法,本文提出的 IPW 算法获得了更好的流行度预测效果.

以下我们考察 IPW 算法及其他基于 *kNN* 的方法在流行度预测上受各参数变化的影响.

首先,变化阈值 p 的取值,同时保持其他参数不变(t_r 为 40 小时,Δt 为 25 小时).p 的变化影响讨论帖数据集的分类情况:当 p 变小时,L_1 类中的数据变少,L_2 类的数据变多;反之,L_1 类中的数据变多,L_2 类的数据变少.实验结果见表 4.

Table 4 Prediction results with varying threashold p (%)

表 4 p 值变化时的预测结果(%)

阈值 p		0.3	0.4	0.5	0.6	0.7	0.8
基于单个动态因素的 *kNN* 算法	评论数	55	52	50	56	53	**60**
	参与评论用户数	54	51	52	54	54	57
	评论树深度	56	**58**	54	53	52	54
	平均节点路径长度	59	54	53	53	54	55
	回复图连接密度	60	52	54	55	53	56
	平均度	59	53	53	56	53	55
IPW 算法		**63**	56	**56**	**58**	**54**	56

从表 4 实验结果中可以看出:对比使用单个动态因素的 *kNN* 算法,本文中提出的 IPW 算法的预测效果相对更稳定,并且在绝大多数 p 的取值下预测准确率更高;但当 $p \geqslant 0.8$ 时,即讨论帖的评论数目变化较小时,IPW 算法的预测效果减弱.一个可能的原因是:当 p 的取值接近边界值时,数据的不均衡性的影响较大.

其次,我们令 t_r 变化,同时固定其他参数(p 为 0.6,Δt 为 25 小时),观察各个动态因素以及 IPW 算法对在不同时间段的讨论帖流行度预测效果的影响.t_r 的变化影响算法的信息输入量:当 t_r 变小时,算法的信息输入变少;反之,信息输入变多.实验结果见表 5.

从表 5 的实验结果可以看出:相对于使用单个动态因素的 *kNN* 算法,IPW 算法的预测结果较为稳定,预测正确率也保持在前两位;但当 t_r 的取值较小时,IPW 算法的预测效果弱于基于单个动态因素的方法.一个可能的原因是信息输入较少带来预测结果波动.

最后,我们考察Δt 的取值对预测效果的影响.令Δt 变化,同时固定其他参数(p 为 0.6,t_r 为 40 小时).实验结果见表 6 ,可以看出:随着Δt 取值的变大,IPW 算法的效果相对更稳定.

以上实验结果表明:在变化各个参数的取值时,相较于使用单个动态因素的方法,总的来说,IPW 算法的性

能更稳定,预测效果更好.

Table 5 Prediction results with varying t_r (%)

表 5 t_r 值变化时的预测结果(%)

t_r		25	30	35	40	45	50
基于单个动态因素的 *kNN* 算法	评论数	58	52	**56**	56	54	53
	参与评论用户数	60	**54**	54	54	55	54
	评论树深度	54	47	53	53	51	52
	平均节点路径长度	57	49	54	53	54	54
	回复图连接密度	**61**	53	54	55	55	53
	平均度	**61**	53	55	56	55	53
IPW 算法		56	**54**	55	**58**	**57**	**55**

Table 6 Prediction results with varying Δt (%)

表 6 Δt 值变化时的预测结果(%)

Δt		15	20	25	30	35	40
基于单个动态因素的 *kNN* 算法	评论数	**57**	55	56	54	56	55
	参与评论用户数	54	51	54	52	59	55
	评论树深度	50	51	53	52	51	53
	平均节点路径长度	54	53	53	49	55	55
	回复图连接密度	56	55	55	53	54	51
	平均度	55	**56**	56	53	54	53
IPW 算法		56	53	**58**	**57**	**60**	**57**

4 结论及进一步的工作

所关注内容的流行度建模、分析和预测的相关研究在商业、安全等相关领域中有重要研究和应用价值.本文基于论坛讨论帖早期的发展演化过程,首次针对其流行度建模和预测问题进行了初步研究并给出明确的问题定义.针对该预测问题,我们总结了影响流行度的相关动态因素,并提出一种融合多个动态因素的讨论帖流行度预测算法.实验结果表明:本文中所提出的算法相对于基准方法具有良好的性能,能够较好地预测讨论帖在一段时间后的流行度.

由于目前所解决的应用问题的特点,本工作采用了论坛数据和二分类的问题定义,但本文所采用的方法可以直接推广到微博等其他类似媒体形式以及预测多类的情形.进一步的研究工作包括:深入分析各个动态因素如何相互作用和影响讨论帖的流行度,并综合考虑静态因素和动态因素,以进一步改进算法和提升预测效果.

References:

[1] Ugander J, Backstrom L, Marlow C, Kleinberg J. Structural diversity in social contagion. Proc. of the National Academy of Sciences, 2012,109(16):5962−5966. [doi: 10.1073/pnas.1116502109]

[2] Yang J, Leskovec J. Patterns of temporal variation in online media. In: Proc. of the fourth ACM Int'l Conf. on Web Search and Data Mining. Hong Kong: ACM Press, 2011. 177−186. [doi: 10.1145/1935826.1935863]

[3] Figueiredo F. On the prediction of popularity of trends and hits for user generated videos. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. Rome: ACM Press, 2013. 741−746. [doi: 10.1145/2433396.2433489]

[4] Chatzopoulou G, Sheng C, Faloutsos M. A first step towards understanding popularity in YouTube. In: Proc. of the INFOCOM IEEE Conf. on Computer Communications Workshops. San Diego: IEEE, 2010. 1−6. [doi: 10.1109/INFCOMW.2010.5466701]

[5] Borghol Y, Mitra S, Ardon S, Carlsson N, Eager D, Mahanti A. Characterizing and modelling popularity of user-generated videos. Performance Evaluation, 2011,68(11):1037−1055. [doi: 10.1016/j.peva.2011.07.008]

[6] Figueiredo F, Benevenuto F, Almeida JM. The Tube over time: Characterizing popularity growth of YouTube videos. In: Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining. Hong Kong: ACM Press, 2011. 745−754. [doi: 10.1145/1935826.1935925]

[7] Borghol Y, Ardon S, Carlsson N, Eager D, Mahanti A. The untold story of the clones: Content-Agnostic factors that impact YouTube video popularity. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Beijing: ACM Press, 2012. 1186−1194. [doi: 10.1145/2339530.2339717]

[8] Brodersen A, Scellato S, Wattenhofer M. YouTube around the world: Geographic popularity of videos. In: Proc. of the 21st Int'l Conf. on World Wide Web. Lyon: ACM Press, 2012. 241−250. [doi: 10.1145/2187836.2187870]

[9] Yin P, Luo P, Wang M, Lee WC. A straw shows which way the wind blows: Ranking potentially popular items from early votes. In: Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. Seattle: ACM Press, 2012. 623−632. [doi: 10.1145/2124295.2124370]

[10] Ahmed M, Spagna S, Huici F, Niccolini S. A peek into the future: Predicting the evolution of popularity in user generated content. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. Rome: ACM Press, 2013. 607−616. [doi: 10.1145/2433396.2433473]

[11] Li HT, Ma XQ, Wang F, Liu JC, Xu K. On popularity prediction of videos shared in online social networks. In: Proc. of the 22nd ACM Int'l Conf. on Information & Knowledge Management. San Francisco: ACM Press, 2013. 169−178. [doi: 10.1145/2505515.2505523]

[12] Pinto H, Almeida JM, Gonçalves MA. Using early view patterns to predict the popularity of YouTube videos. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. Rome: ACM Press, 2013. 365−374. [doi: 10.1145/2433396.2433443]

[13] Zaman T, Fox EB, Bradlow ET. A Bayesian approach for predicting the popularity of Tweets. arXiv e-print 1304.6777. 2013.

[14] Ma HX, Qian WN, Xia F, He XF, Xu J, Zhou AY. Towards modeling popularity of microblogs. Frontiers of Computer Science in China, 2013,7(2):171−184. [doi: 10.1007/s11704-013-3901-9]

[15] Kupavskii A, Umnov A, Gusev G, Serdyukov P. Predicting the audience size of a Tweet. In: Proc. of the 7th Int'l Conf. on Weblogs and Social Media. Cambridge: The AAAI Press, 2013.

[16] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral Tweets. In: Proc. of the 22nd Int'l Conf. on World Wide Web Companion. Republic and Canton of Geneva: ACM Press, 2013. 657−664.

[17] Hong LJ, Dan O, Davison BD. Predicting popular messages in Twitter. In: Proc. of the 20th Int'l Conf. on Companion on World Wide Web. Hyderabad: ACM Press, 2011. 57−58. [doi: 10.1145/1963192.1963222]

[18] Chen GH, Nikolov S, Shah D. A latent source model for nonparametric time series classification. In: Proc. of the Advances in Neural Information Processing Systems. 2013. 1088−1096.

[19] Ma ZY, Sun AX, Cong G. Will this #hashtag be popular tomorrow? In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Portland: ACM Press, 2012. 1173−1174. [doi: 10.1145/2348283.2348525]

[20] Szabo G, Huberman BA. Predicting the popularity of online content. Communications of the ACM, 2010,53(8):80−88. [doi: 10.1145/1787234.1787254]

[21] Lerman K, Hogg T. Using a model of social dynamics to predict popularity of news. In: Proc. of the 19th Int'l Conf. on World Wide Web. Raleigh: ACM Press, 2010. 621−630. [doi: 10.1145/1772690.1772754]

[22] Khosla A, Sarma A, Hamid R. What makes an image popular? In: Proc. of the 23rd Int'l Conf. on World Wide Web Companion. Seoul: ACM Press, 2014. 867−876. [doi: 10.1145/2566486.2567996]

[23] Cheng J, Adamic LA, Dow PA. Can cascades be predicted? In: Proc. of the 23rd Int'l Conf. on World Wide Web Companion. Seoul: ACM Press, 2014. 925−936. [doi: 10.1145/2566486.2567997]

[24] Wu F, Huberman BA. Novelty and collective attention. Proc. of The National Academy of Sciences, 2007,104(45):17599−17601. [doi: 10.1073/pnas.0704916104]

[25] Ye L, Keogh E. Time series shapelets: A new primitive for data mining. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Paris: ACM Press, 2009. 947−956. [doi: 10.1145/1557019.1557122]

[26] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent Data Analysis, 2002,6(5):429−449.

[27] Zhou ZH. Ensemble Methods: Foundations and Algorithms. Chapman & Hall/CRC, 2012. 72−73.

孔庆超(1987－),男,河北邢台人,博士生,主要研究领域为社会媒体分析,数据挖掘.
E-mail: qingchao.kong@ia.ac.cn

毛文吉(1968－),女,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为智能信息处理,人工智能,多 Agent 技术,社会计算.
E-mail: wenji.mao@ia.ac.cn