文章编号:1007-5321(2018)02-0103-06 **DOI**:10.13190/j.jbupt.2017-187

基于网络结构与用户内容的动态兴趣识别方法

黄丹阳[1], 王菲菲[1], 杨 扬[2], 许 进[2]
(1. 中国人民大学 统计学院, 北京 100872; 2. 北京大学 高可信软件教育部重点实验室, 北京 100871)

摘要: 提出了将社交类服务中的两类极为重要的数据——社交网络结构数据和用户所发布的文本内容数据相结合的动态兴趣识别方法. 首先通过定义时间窗口,对社交网络用户的实时文本信息进行主题建模,识别用户实时兴趣概率特征;然后将微观网络结构信息与用户好友的兴趣信息相结合,构建预测特征;最后,建立逻辑回归、支持向量机等分类器,采用所构建的预测特征对用户兴趣进行动态预测. 在新浪微博中的应用表明,该方法具备一定的有效性.

关 键 词: 网络结构; 主题模型; 用户兴趣; 动态识别
中图分类号: TP391.1 **文献标志码**: A

Dynamic Interest Identification Based on Social Network Structure and User Generated Contents

HUANG Dan-yang[1], WANG Fei-fei[1], YANG Yang[2], XU Jin[2]
(1. School of Statistics, Renmin University of China, Beijing 100872, China;
2. Key Laboratory of High Confidence Software Technologies, Peking University, Beijing 100871, China)

Abstract: Two important data sources in social networks, i. e. the network structure and the user generated contents, were combined to dynamically identify user interest. When building topic models, the topic distributions of contents for each user at each time are obtained. And features used for prediction are extracted by summarizing the topical information based on the social network structure. Finally, these prediction features are exploited to dynamically predict user interest via several classification methods, such as logistic regression and support vector machine. The effectiveness of the proposed method is illustrated based on the Sina Weibo dataset.

Key words: network structure; topic model; user interest; dynamic identification

社交类服务中的个性化推荐一直是当前的研究热点. 挖掘用户的兴趣对于个性化推荐至关重要. 已有研究通常从两类极为重要的数据出发,研究用户同质性,挖掘用户的潜在兴趣. 第 1 类数据是社交数据,研究认为,2 个节点的相似性越强,他们之间产生链路的可能性就越高[1]. 所以, 网络结构信息经常被用于进行相似性评价. 另一类数据是社交网络中用户发布的文本内容数据. 其中用到的一类重要模型是潜在狄利克雷分配 (LDA, latent Dirichlet allocation) 模型[2],其出发点在于认为文本中隐含着丰富的主题. 学者通过研究用户之间主题分布的相似性对用户之间的相互关注关系进行预测

收稿日期: 2017-09-14
基金项目: 国家自然科学基金项目(11701560); 北京市社会科学基金项目(17GLC051); 中央高校建设世界一流大学(学科)和特色发展引导专项资金项目; 国家统计局一般项目(2017LY83); 中国博士后科学基金项目(2017M620985)
作者简介: 黄丹阳(1989—), 女, 讲师; 王菲菲(1988—), 女, 讲师, E-mail:feifei.wang@ruc.edu.cn.

和解释[3-4],或者直接利用微博文本及转发特性对于用户的兴趣进行挖掘[5]. 但这2种数据并未有效结合. 鉴于此,定义用户动态兴趣,结合文本与网络结构信息,同时考虑用户的历史兴趣信息,对于用户的下一期兴趣进行预测. 利用新浪微博数据验证了模型效果. 不同分类模型的实验效果表明,将用户网络结构信息与用户兴趣信息相结合,可以对用户的兴趣进行有效地动态识别,进而为用户进行个性化推荐.

1 动态兴趣识别模型

1.1 基本定义

假定能够收集到用户的网络结构信息和文本微博数据. 为了动态研究用户关注的话题,笔者将收集到的网络结构数据集定义为邻接矩阵,并且对于微博内容进行动态时间窗口的划分.

1) 邻接矩阵. 对于一个包含 n 个用户的网络,如果一个用户 $i(1\leqslant i\leqslant n)$ 关注了用户 $j(1\leqslant i\leqslant n)$,则定义 $a_{ij}=1$;否则 $a_{ij}=0$. 定义 $a_{ii}=0(1\leqslant i\leqslant n)$. 笔者为整个社交网络定义其邻接矩阵 $\boldsymbol{A}=(a_{ij})\in\mathbb{R}^{n\times n}$. 由于在数据收集期间,网络变动的边较少,假设收集到的社交网络结构不随时间变化.

2) 动态时间窗口. 微博收集到的总体时间区间为$[T_s,T_e]$,其中,T_s 为起始时间,T_e 为终止时间. 将时间区间等分为 T 份,其中第 t 个时间分割点为 $T_s+t(T_e-T_s)/T$,其中 $1\leqslant t\leqslant T$. 定义在 t 时间内,即动态时间窗口$[T_s+(t-1)(T_e-T_s)/T, T_s+t(T_e-T_s)/T]$之内,所发布的全部微博用来分析用户在 t 时间段内的兴趣. 注意,任何一个用户对某个主题的兴趣不是一成不变的,将随着时间 t 的变化发生变化.

3) 动态兴趣主题. 假定用户感兴趣的主题共有 K 个. 定义用户 i $(1\leqslant i\leqslant n)$ 在 $t(1\leqslant t\leqslant T)$ 时间段内对于第 $k(1\leqslant k\leqslant K)$ 个主题的兴趣为 $Y_{k,i,t}$. $Y_{k,i,t}=1$ 表示该用户 i 在 t 时间内具有第 k 个主题相关的兴趣;反之 $Y_{k,i,t}=0$ 表示用户 i 在 t 时间内不具有第 k 个主题相关的兴趣.

1.2 基于实时数据的 LDA 模型

使用基于实时数据的 LDA 模型. 定义 $\boldsymbol{\theta}_{i,t}=(\theta_{1,i,t},\cdots,\theta_{K,i,t})^{\mathrm{T}}$ 为第 i 个人在时刻 t 所发文章在 K 个主题上的概率分布,$\boldsymbol{\phi}_k=(\phi_{1,k},\cdots,\phi_{V,k})^{\mathrm{T}}$ 为主题 k 在 V 个词所构成的词典空间上的概率分布,α 和 β 为超参数.

对于第 i 个人在时刻 t 所发文章,其生成过程如下.

1) 从狄利克雷分布中产生该文档在 K 主题上的概率分布 $\boldsymbol{\theta}_{i,t}\sim\mathrm{Dir}(\alpha)$;

2) 对该文档中的第 c 个词:

① 从多项分布中产生它表达的具体的主题:$z_{i,t,c}\sim\mathrm{multi}(\boldsymbol{\theta}_{i,t})$;

② 从多项分布中产生表达该主题的具体的词 $w_{i,t,c}\sim\mathrm{multi}(\boldsymbol{\phi}_{z_{i,t,c}})$;

3) 从狄利克雷分布中产生主题 k 在所有词上的概率分布 $\boldsymbol{\phi}_k\sim\mathrm{Dir}(\beta)$.

基于上述文档生成过程,可以得到全部文档的联合似然函数,然后使用吉布斯抽样(Gibbs Sampling)的方法进行模型求解,得到 $\boldsymbol{\theta}_{it}(1\leqslant i\leqslant n,\ 1\leqslant t\leqslant T)$ 和 $\boldsymbol{\phi}_k(1\leqslant k\leqslant K)$ 的估计值.

1.3 网络结构特征

1) 动量特征

以往的研究表明,人们倾向于延续以往的行为模式[6]. 由此可知,如果用户 i 在 $t-1$ 时间内在某一个兴趣主题的概率取值较高,则在 t 时间内也有可能继续保持对于该主题的兴趣. 为此,定义动量特征为该用户上一期在该兴趣主题上表现出的概率值. 通过符号表达为

$$X_{1k,i,t}=\theta_{k,i,t-1} \tag{1}$$

其中 $\theta_{k,i,t-1}$ 为用户 i 在 $t-1$ 时间内在主题 k 上表现出的兴趣特征概率.

2) 动态网络结构相关变量

在微博中单向关注关系和双向关注关系表示不同的亲疏程度,而存在双向关注关系的用户往往具备更强烈的互相影响能力. 故可以从不同的关注关系入手定义动态网络结构相关变量.

① 同质性特征. 传统社交行为中的同质性理论表明,具备双向链接的人们更倾向于具备相似的兴趣取向[7]. 上一期与该用户存在双向链接的用户在某一主题上具备的概率特征可能对于当期该用户在该主题上的概率特征有一定的预测能力. 故可定义同质性特征 $X_{2k,i,t}$ 如下:

$$X_{2k,i,t}=\frac{\sum_{j=1}^{n}a_{ij}a_{ji}\theta_{k,j,t-1}}{\sum_{j=1}^{n}a_{ij}a_{ji}} \tag{2}$$

② 同兴趣特征. 人们的行为能够受到某些共同外生因素,即共同兴趣的影响[8]. 在上一期该用

户关注的所有人(除双向关注关系链接)在一定程度上反应了该用户的阅读兴趣,而该用户关注的所有人在某主题上的平均概率特征一定程度上影响了该用户在当期于该主题上的概率特征. 故定义同兴趣特征 $X_{3k,i,t}$ 为

$$X_{3k,i,t}=\frac{\sum_{j=1}^{n}a_{ij}(1-a_{ji})\theta_{k,j,t-1}}{\sum_{j=1}^{n}a_{ij}(1-a_{ji})} \tag{3}$$

③ 同因子特征. 用户通过微博文字表达出的兴趣特征往往能够通过关注他们的用户阅读兴趣得到体现,即关注该用户的人在某主题上的平均概率特征一定程度上能够反映该用户对于该主题的概率特征. 由于模型用于预测,当期关注该用户人的平均主题概率特征不能用于预测当期概率,故采用前一期指标作为预测特征. 定义同因子特征 $X_{4k,i,t}$ 为

$$X_{4k,i,t}=\frac{\sum_{j=1}^{n}a_{ji}(1-a_{ij})\theta_{k,j,t-1}}{\sum_{j=1}^{n}a_{ji}(1-a_{ij})} \tag{4}$$

值得注意的是,随着每一期用户特征的体现,兴趣特征概率 $\theta_{k,i,t}$ 不断变化,故对于不同特征的计算是不断更新的动态过程.

1.4　用户内容特征

1) 兴趣广度

用户在某一时间段内兴趣的广度也会影响用户对某一具体兴趣的关注程度. 用户的兴趣越广泛,越有可能在下一期关注新的主题或者不再关注原有感兴趣的主题;相反,用户的兴趣越集中,越有可能继续关注原有感兴趣的内容. 因此,使用熵来刻画用户在主题分布上的离散程度,从而衡量用户兴趣的广度. 对于用户 i 在 $t-1$ 时间内所发的全部内容在 K 个主题上的分布,其主题熵的计算公式为

$$X_{5,i,t-1}=-\sum_{k=1}^{K}\theta_{k,i,t-1}\log\theta_{k,i,t-1} \tag{5}$$

2) 用户相似性

相邻用户往往具有更高的内容相似性. 首先基于 LDA 模型得到主题分布,计算用户的主题相关性系数. 已知用户 i 和用户 j 的文本内容在 k 个主题上的分布 $\theta_{k,i,t-1}$ 和 $\theta_{k,j,t-1}$,相关性系数的计算方法为

$$C_{i,j,t-1}=\frac{\sum_{k=1}^{K}(\theta_{k,i,t-1}-\bar{\theta}_{i,t-1})(\theta_{k,j,t-1}-\bar{\theta}_{j,t-1})}{\sqrt{\sum_{k=1}^{K}(\theta_{k,i,t-1}-\bar{\theta}_{i,t-1})^2\sum_{k=1}^{K}\theta_{k,j,t-1}-\bar{\theta}_{j,t-1})^2}} \tag{6}$$

其中:$\theta_{k,i,t-1}$ 表示用户 i 在 $t-1$ 时间内的内容在第 k 个主题上的概率取值,$\bar{\theta}_{i,t-1}$ 表示 $\theta_{k,i,t-1}$ 在全部 K 个主题熵的均值;$\theta_{k,j,t-1}$ 和 $\bar{\theta}_{j,t-1}$ 表示的含义类似.

基于相关性系数,根据网络结构进一步定义了 3 个指标,分别用于度量用户 i 在 $t-1$ 时间内,与该用户具有双向链接所有用户的平均相似性系数、该用户关注的所有人的平均相似性系数以及与所有关注该用户全部用户的平均相似性系数. 这 3 个预测特征的计算公式如下:

$$X_{6,i,t}=\frac{\sum_{j=1}^{n}a_{ij}a_{ji}C_{i,j,t-1}}{\sum_{j=1}^{n}a_{ij}a_{ji}} \tag{7}$$

$$X_{7,i,t}=\frac{\sum_{j=1}^{n}a_{ij}(1-a_{ji})C_{i,j,t-1}}{\sum_{j=1}^{n}a_{ij}(1-a_{ji})} \tag{8}$$

$$X_{8,i,t}=\frac{\sum_{j=1}^{n}a_{ji}(1-a_{ij})C_{i,j,t-1}}{\sum_{j=1}^{n}a_{ji}(1-a_{ij})} \tag{9}$$

在后续分析中,会将上述定义的 8 个预测特征直接作为分类器的输入,对不同时间内用户的动态兴趣进行实时预测.

2　实验设置

2.1　实验数据集

收集的原始数据采集自新浪微博,采集时间段为 2014-01-01—2014-12-31,其中包含 34 086 个用户,有向连接 594 365 条以及每个用户在数据收集时间段内原创和转发的微博共 649 428 条. 为了动态刻画用户关注的兴趣主题情况,并克服文本内容较短的问题,将同一个用户 i 在时间段 t 内的所有微博内容进行拼接,并对文本进行了预处理.

使用第 1 节中介绍的数据来验证不同预测特征对于用户兴趣的预测效果. 将 2014 年 12 个月按季度划分为 4 个时间段($T=4$):1 ~ 3 月为第 1 时间段,即 $t=1$;4 ~ 6 月为第 2 时间段,即 $t=2$;7 ~ 9 月为第 3 时间段,即 $t=3$;10 ~ 12 月为第 4 时间段,即 $t=4$. 实际应用中,可以进行更为细致的时间段划

分,此处只按照此种划分方式进行结果展示. 将每个用户在每个时间段所发的全部微博内容进行拼接,得到最终用于建模的文本内容.

2.2 兴趣主题设置

对全部文本内容进行预处理,然后统一建立主题模型. 选用的模型估计方法为吉布斯抽样,并使用开源软件包 GibbsLDA + + 进行模型实现. 为了验证实验的有效性和稳定性,在基于实时数据的 LDA 模型中选用了 2 个不同的兴趣主题数,$K=50$ 或 $K=100$,由于结果相似,只展示 $K=50$ 的模型结果. 模型中控制超参数 $\alpha=50/K$,$\beta=0.1$. 确定主题数后,可以得到每个用户 i 在 t 时间内所发文本内容在每个兴趣主题 k 上的概率分布 $\theta_{k,i,t}$ 以及每个兴趣主题 k 在整个词典空间上的概率分布.

在不同主题数的 LDA 模型下,计算了整个文档集合中每个主题 k 出现的平均概率值,即

$$\bar{\theta}_k=\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\theta_{k,i,t} \tag{10}$$

然后选择 $\bar{\theta}_k$ 取值最高的前 10 个主题作为“热门兴趣主题”.

2.3 分类器设置

在后续的实验中,笔者将对选出的每个热门兴趣主题进行用户兴趣的动态预测,即考察用户 i 在 t 时间内是否会对某个热门主题感兴趣,并分析其影响因素. 对于任意主题 k,选取所有用户在该主题上概率取值的 90% 分位数作为阈值 ε_k. 如果用户 i 在 t 时间内在主题 k 上的概率取值 $\theta_{k,i,t}>\varepsilon_k$,则定义 $Y_{k,i,t}=1$;反之,定义 $Y_{k,i,t}=0$. 为了探索各个预测特征对于用户兴趣的影响情况,建立了逻辑回归模型说明预测特征的有效性. 此外,还分别建立了逻辑回归(LR, logistic regression)、支持向量机(SVM, support vector machine)、随机森林(RF, random forest)和朴素贝叶斯(NB, native Bayes)3 种分类器. 其中,支持向量机采用径向基核函数(RBF, radial basis function),并通过十折交叉验证的方式确定核函数中涉及的参数;随机森林中设定树的个数为 500,候选分裂属性数为所有预测特征个数的二次方根. 在每种分类器下,都对全部数据进行了十折交叉验证,即随机将数据分成 10 份,利用其中 9 份作为训练集,1 份作为测试集,使用训练集建立分类器后,在测试集上计算评价指标. 将 10 份数据轮流作为测试集,并将计算的评价指标结果取平均来衡量该分类器的预测精度. 通过不同方法下曲线下面积(AUC, area under the curve)、准确率、召回率和 F1 值的情况来展示预测效果的准确性.

3 实验评价

3.1 用户兴趣展示

表 1 展示了兴趣主题数 $K=50$ 时,基于实时数据的 LDA 模型挖掘出的用户前 10 个热门兴趣主题. 通过概括每个热门兴趣主题下出现概率最高词的含义来归纳总结该兴趣主题的含义. 例如,第一个热门兴趣主题中出现概率较高的词为“累/吃/上班/睡觉/抓狂/晚上/回家/明天/回家/休息”,这些词多是反映上班族的生活情况,所以概括该主题为“上班族”;在第 2 热门兴趣主题中,出现概率较高的词多是和旅行以及旅游景点相关的,则概括该兴趣主题为“旅行”.

表 1 用户的前 10 个热门兴趣主题

兴趣主题	出现概率较高的词语
主题 1:上班族	累/吃/上班/睡觉/抓狂/晚上/回家/明天/回家/休息
主题 2:旅行	旅行/走/旅游/丽江/出发/风景/路上/云南/公里/美丽
主题 3:青春时光	生活/青春/时光/人生/美好/阳光/岁月/回忆/幸福/梦
主题 4:生日祝福	爱/生日快乐/开心/蛋糕/谢谢/快乐/幸福/祝福/记忆/温暖
主题 5:爱情婚姻	女人/爱/男人/爱情/喜欢/结婚/幸福/老公/老婆/一辈子
主题 6:美食	吃/馋嘴/好吃/美食/味道/吃货/店/不错/咖啡/美味
主题 7:情感强烈	偷笑/嘻嘻/泪/鼓掌/酷/汗/抓狂/开心/晕/衰
主题 8:搞笑	哈哈哈/笑/偷笑/爆笑/节操/笑哈哈/挖鼻屎/尼玛/屌丝/女神
主题 9:写作生活	说/写/发现/故事/工作/发表/文章/生活/书/思考
主题 10:公司企业	公司/工作/企业/市场/产品/员工/客户/管理/服务/行业

3.2 模型结果

在评价预测结果之前,为研究各特征指标对于用户兴趣是否有影响,采用逻辑回归分类器进行训练,得到相应的回归系数及显著性指标. 由于兴趣较多,回归结果多样化,对于所有主题的回归系数计算平均值以展示其平均趋势. 所有系数的平均结果通过表 2 列出. 其中对于前 10 个主题,动量特征、同质性特征、同兴趣特征、同因素特征以及用户相似

性特征 3 系数均在 0.05 显著性水平下显著.

表 2　平均回归系数结果

预测特征	系数值	预测特征	系数值
动量特征	17.125	兴趣广度	−8.650
同质性特征	1.680	用户相似性特征 1	−0.105
同兴趣特征	0.854	用户相似性特征 2	0.006
同因素特征	2.266	用户相似性特征 3	0.327

下面对各个显著的预测特征加以说明. 首先,动量特征系数显著为正,表明用户历史兴趣对于将来兴趣有非常好的预测作用. 同质性特征、同兴趣特征、同因素特征系数显著为正,表明与用户存在双向链接的人、用户关注的人、用户的粉丝在上一期的兴趣均对于用户当期的兴趣有显著的正向影响作用. 用户相似性特征 3 的系数都是正的,说明上一期用户和所有关注他的人在主题分布上越相似(这往往是由于上一期在热门主题上有相同的偏好),用户就越有可能在当期关注某个热门的兴趣主题.

3.3　预测精度

如上文所述,采用十折交叉验证的方法对于模型预测精度进行评价. 这里只展示 $K=50$ 时的模型结果. 选取最热门的 10 个兴趣主题,利用往期用户兴趣信息预测当期用户是否存在该兴趣主题. 在选择预测模型时,除逻辑回归分类器之外,还将提出的用户动态兴趣预测方法在支撑向量机、随机森林和朴素贝叶斯 3 个分类器上进行测试.

图 1 和图 2 所示为 4 种模型在 10 个兴趣主题上的 AUC 取值. 可以发现,在各个主题上,4 种模型的 AUC 都在 0.65 以上,预测效果较好,其中有关“强烈情感”和“公司企业”的预测精度可以达到 0.8 以上. 对比不同模型的预测效果可以发现,逻辑回归的预测精度在所有主题上都是最好的.

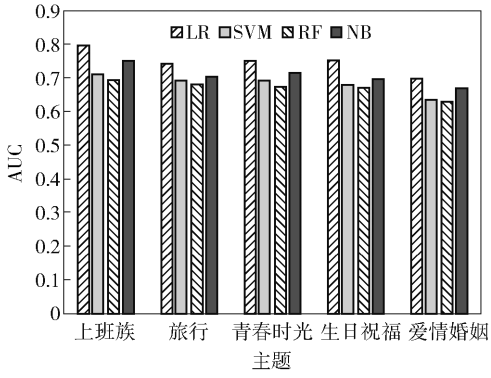

图 1　不同模型的 AUC 结果(前 5 个主题)

除 AUC 之外,还计算了不同分类器所取得的准确率、召回率和 F1 值. 表 3 给出了对前十大热门主题的预测上,各个分类器的平均准备率、召回率和 F1 值.

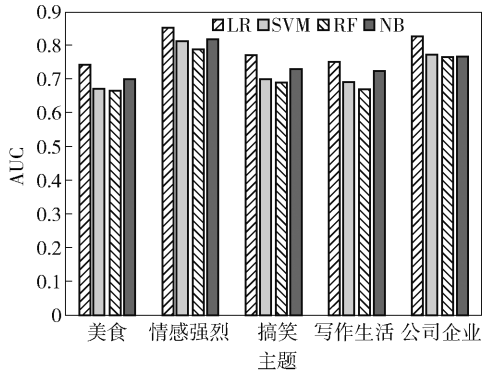

图 2　不同模型的 AUC 结果(后 5 个主题)

可以看到,所有分类器的准确率都在 0.9 以上,召回率相对较差,整体的 F1 取值可以维持在 0.64 ~ 0.71 的水平上,预测效果良好. 对比不同分类器发现,逻辑回归在准确率、召回率和 F1 取值上都明显高于其他分类器. 因此,在使用笔者提出的预测特征对用户的兴趣进行动态识别时,推荐使用逻辑回归分类器.

表 3　各分类器的平均准确率、召回率和 F1 值

分类器	准确率	召回率	F1 值
LR	0.972	0.632	0.710
SVM	0.926	0.570	0.640
RF	0.935	0.577	0.645
NB	0.962	0.605	0.679

以上实验结果表明,社交网络结构的动态兴趣识别方法对于用户的兴趣主题识别具有较好的预测效果,进而在识别用户兴趣的基础上,对于相关产品、服务、内容的个性化推荐具有非常重要的作用.

4　结束语

利用社交类服务中沉淀的两类极为重要的数据建立对于用户的动态兴趣预测模型:1)社交网络结构数据;2)用户所发布的文本内容数据. 在动态时间窗口的基础上,将用户在社交类服务中产生的文本内容信息划分为不同阶段,用户的兴趣随着时间发生的变化将被精细刻画. 在将每个用户的每个时间段定义为一个文档,并作为基础单位进行分析后,采用 LDA 模型对于用户的文本内容进行主题挖掘.

结合邻接矩阵的定义,网络微观结构信息与好友兴趣信息以及用户历史兴趣信息,对于用户的下一期兴趣进行预测. 随着时间不断推演,模型能够对于用户兴趣特征概率动态更新,预测用户的动态兴趣. 实证结果表明,此方法对于用户兴趣具有较好的预测效果,从而可应用于相关产品、服务、内容的个性化推荐.

参考文献:

[1] Lü L, Zhou T. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[2] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[3] Barbieri N, Bonchi F, Manco G. Who to follow and why: link prediction with explanations[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2014: 1266-1275.

[4] Ahmed C, Elkorany A. Enhancing link prediction in twitter using semantic user attributes[C]//IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. [S. l.]: ACM, 2015: 1155-1161.

[5] Wang Y, Zhang F, Liu Y. et al. Method for user's interest topic mining in micro-blog with repost feature[J]. Application Research of Computers, 2017, 34(7): 2068-2071.

[6] Chintagunta P K. Inertia and variety seeking in a model of brand-purchase timing[J]. Marketing Science, 1998, 17(3): 253-270.

[7] Shalizi C R, Thomas A C. Homophily and contagion are generically confounded in observational social network studies[J]. Sociological Methods & Research, 2011, 40(2): 211-239.

[8] Ma L, Krishnan R, Montgomery A L. Latent homophily or social influence? an empirical analysis of purchase within a social network[J]. Management Science, 2015, 61(2): 454-473.