微博网络中用户主题兴趣相关性及主题信息扩散研究

罗春海，刘红丽，胡海波

(华东理工大学商学院 上海 徐汇区 200237)

【摘要】运用Twitter-LDA主题模型对新浪微博数据进行了主题分析，基于用户主题兴趣相关性的研究表明用户间的主题兴趣具有三度相关性，同一主题兴趣下三度以内粉丝的发文数随用户发文数增加而波动式增加，各度粉丝与用户主题兴趣的相似度随粉丝度数的增加而下降。通过分析比较不同主题类别微博的扩散差异，发现生活情感类的信息最受用户欢迎，不同主题类别微博被转发的概率存在显著差异，平均转发数相差可达10倍，微博信息扩散树中各类主题在微博信息扩散深度、扩散时间间隔和用户的扩散能力方面都表现出不同的特征。

关 键 词 信息扩散； 微博网络； 主题分析； 用户行为

中图分类号 N949 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.02.022

Research on Correlation of Users′ Topic Interests and Topic Information Diffusion in Microblog Networks

LUO Chun-hai, LIU Hong-li, and HU Hai-bo

(School of Business, East China University of Science and Technology Xuhui Shanghai 200237)

Abstract The topic analysis on the Sina microblog data is studied by using the Twitter-LDA topic model. The analysis based on correlation of users' topic interests shows that topic interests between users follow the three degrees of correlation. Within the same topic interest when the number of microblogs that users publish increases, the number of microblogs that their fans within three degrees publish also increases in fluctuation, and the similarity of topic interests between users and their multi-degree fans decreases with the increase of degree. Through the analysis and comparison of the diffusion difference of diverse topic categories, we find that users prefer the information with lifestyle topic, reposting probability is significantly different among microblogs within different topic categories, and the average reposting count can be 10 times in difference. In microblog information diffusion trees, diffusion depth, diffusion time interval and users' diffusion ability all show different characteristics for microblogs with different topic categories.

Key words information diffusion; microblog network; topic analysis; user behavior

微博集媒体性和社交性于一身[1]，以其多途径接入、多类型信息传播的特点，吸引了大量的用户。《2014年新媒体蓝皮书》显示在中国提供微博服务的网站有103家，注册用户数达13亿之多，2013年仅新浪微博每天就产生超过一亿条微博，对这些信息及其扩散特征的研究具有重大的社会经济意义，如对微博信息的研究，可以用于预测电影的票房[2]、股市走势[3]，对信息扩散特征的研究则可以用于精准广告营销[4]等。利用用户历史数据研究用户兴趣和用户与其粉丝间兴趣的相关关系可以更好的理解用户偏好，有助于理解网上信息扩散的机理，帮助政府部门有效引导、控制网络舆情。

对微博网络信息扩散的研究可以从微观和宏观两个角度进行，微观角度主要从单个用户或用户间关系的视角研究影响信息扩散的各种因素，宏观角度则主要研究信息扩散的整体特征[5]。在微观角度上，文献[6]分析了博文是否包含网址、标签，是否提及他人以及用户粉丝数、朋友数、帐号使用时间等因素对微博转发概率的影响。文献[7]研究了用户发表的微博数量分布的幂指数和用户间的互动指数之间的关系，发现两者呈反向变动趋势。也有学者通过对用户历史数据的分析，根据各种因素对信息扩散的影响预测未来的信息扩散[8-11]。近年来，学者们利用微博网络历史数据研究了信息内容对信息扩散的影响，如文献[12]发现信息包含更多的消极情绪、行为和复杂的认识过程会加快信息的消亡，文

收稿日期：2015－05－17；修回日期：2016－01－23
基金项目：国家自然科学基金(61473119, 61104139)；中央高校基本科研业务费(WN1524301)
作者简介：罗春海(1989－)，男，主要从事在线社会网络方面的研究.

献[13]的研究表明不同的信息不仅在用户和用户之间扩散的概率不同，不同信息重复暴露对其被采用的边际贡献率也不同。有的学者在研究信息内容对信息扩散的影响时同时考虑用户的主题兴趣，如文献[14]将用户兴趣和信息内容结合起来提出了一种基于信息亲和机制的SKIR扩散模型，研究表明信息亲和阈值影响了信息的最终扩散范围，文献[15]则根据用户之间的主题兴趣相似度发现Twitter中用户和她/他的直接粉丝之间存在着同质现象，用户间主题兴趣越相似，信息越容易在两个用户间扩散，文献[16]利用用户主题兴趣和间接影响力，提高了预测Twitter用户转发行为的准确度。不仅文本信息，情绪、行为等也可以在社会网络上扩散，并且遵循“三度影响力原则”[17]。文献[18]的研究表明新浪微博中用户间不同的情绪尤其是愤怒具有较高的相关性，这种相关性同样限于三度粉丝以内。在宏观角度上，文献[19]系统分析了Twitter信息扩散树深度、扩散时间间隔等特征。此外，学者们在不同主题类别的信息扩散差异上也做了一些研究，如文献[20]对Twitter内容分析时发现在Twitter中不同主题类别微博的转发率存在差异，文献[21]则发现新浪微博的热门话题大多是关于休闲娱乐的话题。

用户的主题兴趣是影响信息扩散的一个重要因素，研究它能否像情绪、行为一样在微博网络中具有相关性以及这种相关性遵循的规律，有助于理解某一类主题信息的扩散过程和微博网络的形成，引导用户兴趣的培养和微博网络中的信息扩散，然而目前对相关方面的研究仍不够深入。直观上看，信息扩散的整体特征是大量用户转发行为构成的，不同主题的信息在用户之间的扩散概率存在差异，因此不同主题类别的信息扩散整体上可能会表现出不同的特征。虽然可以像文献[21]根据微博实时提供的热门话题关键词对不同主题类别的信息扩散展开研究，但是热门话题只包含了少量的主题和微博，隐藏在热门话题外的大量微博仍有待于进一步分析，而文献[20]的研究主要针对Twitter和传统媒体在内容上的区别，没有针对不同主题类别的微博在扩散上的差异，对不同主题类别的信息扩散特征的研究仍相对较少。为此，本文在主题分析的基础上，探讨用户与其各度粉丝之间主题兴趣的相关性，并对各类主题微博的扩散差异展开研究。为方便叙述，下文将用户的发表和转发行为统称为发表行为，用户的粉丝称为一度粉丝，粉丝的粉丝称为二度粉丝，并依此类推。

1　数据描述

1.1　数据收集

本研究利用新浪微博提供的API接口[22]，从一个粉丝数和微博数较多的用户开始，将该用户加入爬取队列，根据研究需要爬取该用户最新发布的100条微博，对其中的每条微博，再爬取该微博的原创微博和转发微博以及原创微博和转发微博的用户信息，并将这些用户加入爬取队列。一个用户处理完后，再提取爬取队列中的第一个用户进行相同处理，并不断重复上述操作。从2014年10月15日至10月20日共收集了21 992个用户信息和这些用户发布的2 076 564条微博的详细信息，随后本文收集了这些用户的转发关系，排除陌生人(即非本用户粉丝)转发，共得到258 116条关注关系。本文收集每个用户最新发表的100条微博和这些微博间有转发关系的粉丝，因此得到每个用户粉丝列表和关注列表的一部分。

1.2　数据预处理

爬取的数据集中原创微博占36.3%，除去空文本微博共得到1 919 406条博文。删除博文中系统自动产生的文本以及@用户名、表情符、所有非中文字符，同时将繁体中文转换成简体中文。之后利用ICTCLAS&NLPIR[23]对博文进行分词，删除停止词、高频词、低频词后得到表1统计信息。除去文本容量少于2 kB的用户后共得到21 750个有效用户。

表1　分词后博文统计信息

用户数/个	微博数/条	单词种类/种	单词总数/个
21 750	1 919 406	78 736	66 972 397

2　微博主题分析和主题分类

2.1　主题分析

主题是指所说或所写的内容[24]。文献[20]将主题分为事件型、实体型和长期型，并认为主题类别是属于共同主题领域的一组主题。在主题分析方法中LDA(latent dirichlet allocation)作为强有力工具被广泛运用到微博文本分析中，学者们根据微博文本的特点，提出了许多适用于微博环境的主题分析模型，例如文献[20]在对Twitter和传统新闻媒体纽约时报进行内容比较时提出了Twitter-LDA模型。文献[25]对各种主题分析模型进行了研究，发现UserLDA、AuthorLDA和Twitter-LDA运用到微博环境时各有自己的优点。根据研究需要，本文运用Twitter-LDA进行微博主题分析。

利用Twitter-LDA对收集的博文进行主题分析，得到：1) 用户主题分布矩阵**DT**，**DT**为$D\times T$维矩阵，D表示用户数量，T表示主题数量，T=120，**DT**(i, j)表示节点i对主题j的感兴趣程度，其值越大表明节点i对主题j越感兴趣；2) 各个主题单词的概率分布；3) 每条微博所属主题。

2.2 主题识别和分类

对主题分析得到的120个主题进行人工识别，舍弃其中不能识别的34个主题、3个杂乱主题和2个有关微博本身的主题，剩余的81个参考新浪微博的分类方法，将它们分成社会、体育、娱乐、旅游、美食、医疗保健、财经、科技、生活情感、政治、教育、文化、天气、时尚共14个类别。不同于新浪微博主题分类方法，本研究增加了政治、教育、天气、文化、时尚主题类别，将公益主题合并为社会类别，将综艺、娱乐八卦、电视节目、电视剧、电影、动漫、音乐归为娱乐类。教育类包含校园生活、读书，政治类包含国际历史和国际社会，表2列出了每个主题类别相关信息，其中相关词汇是每个主题词汇分布中出现频率最高的前3个词语。

表2 主题类别、主题数和相关词汇表

主题类别	包含主题数	相关词汇
社会	7	公益 活动 爱心 媒体 微博 新闻 村民 冲突 死 寻 联系 女孩 动物 车 野生 粮食 浪费 亿 环卫工 垃圾 扔
体育	3	北京 跑 马拉松 比赛 对 球 瘦 动作 腿
娱乐	8	深夜 预告 终极 座 星座 羊 吸毒 警方 柯震东 卫视 节目 明星 活动 届 现场 电影 剧片 歌 音乐 唱 刘诗诗 风 缘
旅游	7	山 旅行 拍 旅行 景区 丽江 签证 美国 签 卡 台湾 车 攻 安全 司机 路 站 车 飞机 航空 架
美食	2	美食 做法 菜 放 水 锅
医疗保健	11	医院 钱 希望 水 皮肤 洗 洗手 正确 洗 头发 脱发 牙膏 衰老 身体 器官 粥 寒露 春喝 病毒 茶 健康 药 女孩 使用 健康 熬夜 医生 中医 健康 埃博拉 死亡 医院
财经	6	亿 美元 经济 价格 柴油 汽油 重要 成功 工作 公积金 住房 贷款 公司 企业 互联网 车 汽车 辆
科技	5	奖 诺贝尔 科学家 转基因 食品 安全 手机 卡 银行 苹果 电脑 系统 小米 送 台
生活情感	9	心 人生 累 妈妈 宝宝 父母 麻烦 回家 看到 老人 生命 人生 奶奶 花心 风 男 钱 结婚 人生 幸福 努力 女人 男人 爱情 快乐 生日 幸福
政治	8	香港 美国 朝鲜 日本 历史 演讲 改革 党 法制 法院 案 律师 调查 书记 工作 人民 烈士 历史 韩国 船 韩 国家 社会 政治
教育	6	工作 学生 厦 英语 单词 背 考 考试 报名 档案 毕业生 年级 大学 学生 老师 读书 阅读 周末
文化	5	心 佛 一点 书 读 文化 文艺 习近平 作品 画 艺术 座谈会 诺贝尔 文学奖 作家
天气	2	雾霾 霾 空气 地震 级 气温
时尚	2	穿 拍 街 设计 款 时尚

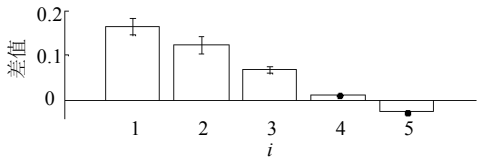
3 用户主题兴趣相关性研究

2.1节主题分析得到了用户主题分布**DT**矩阵，**DT**矩阵可以用来量化用户的主题兴趣。本节利用**DT**矩阵，对用户主题兴趣相关性展开研究。

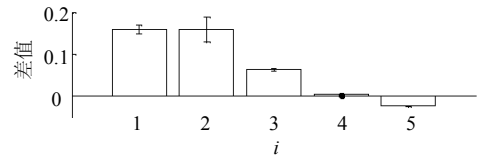
3.1 用户与其各度粉丝之间主题兴趣的相关性

本文选取每个主题下发表关于该主题的微博最多的前100个用户，探讨这些用户和他们的各度粉丝之间主题兴趣的相关性。

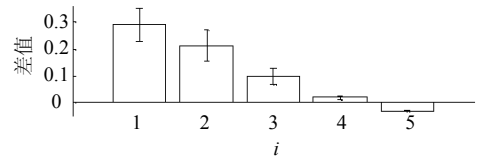
设 p_1 为用户各度粉丝发表相同主题微博的概率，$\overline{p}$ 为全网络的平均概率，i表示粉丝度数，图1展示了各主题类别内 $p_1-\overline{p}$ 在i取不同值时的均值和标准差。图1表明，$p_1-\overline{p}$ 随i的增加而逐渐下降(只有在政治、文化主题类别下 $p_2-\overline{p}$ 大于 $p_1-\overline{p}$，但是总体还是呈下降趋势)，当$i\leqslant 3$时，各类主题普遍存在 $p_1-\overline{p}>0$，当i=4时，$p_1-\overline{p}$ 已经很小甚至为负值，当i=5时，所有主题类别的 $p_1-\overline{p}$ 值都为负。这说明用户发表关于某一主题微博的行为和她/他各度粉丝的相同行为具有相关性，这种相关性随着粉丝度数的增加而下降，超出三度分隔这种相关性就消失了，用户主题兴趣在微博网络中具有三度相关性。造成用户主题兴趣相关的可能原因一是社会影响，即被关注者的主题兴趣爱好影响了她/他的粉丝，使得这些粉丝跟被关注者具有相同的兴趣爱好；二是趋同性，人们往往倾向于跟与自己相似的人建立好友关系，因而用户会倾向于关注跟自己的兴趣爱好类似的其他用户。而用户主题兴趣限于三度相关的原因可能包括用户的相互影响力本身固有的衰减性和微博网络的不稳定性。



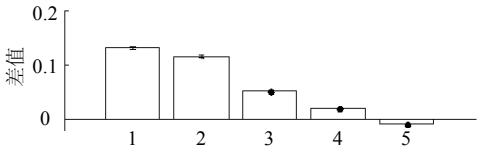
a. 社会



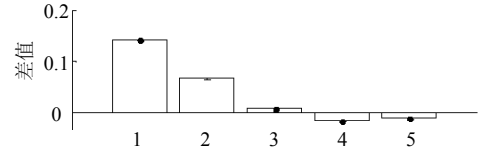
b. 体育



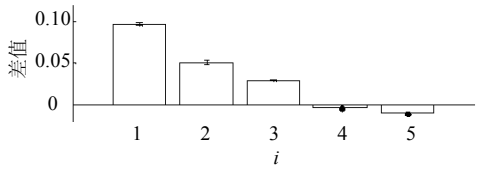
c. 娱乐



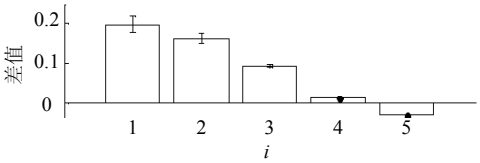
d. 旅游



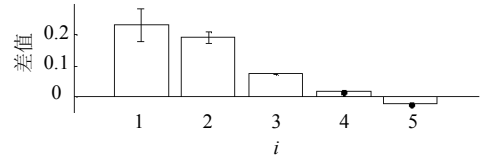
e. 美食



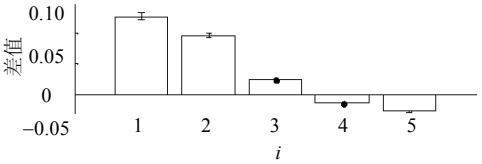
f. 医疗保键



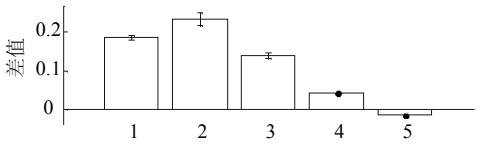
g. 财经



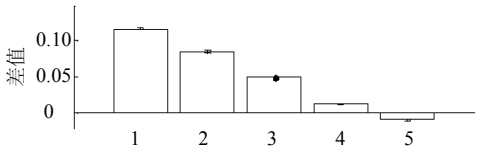
h. 科技



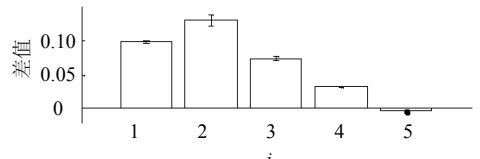
i. 生活情感



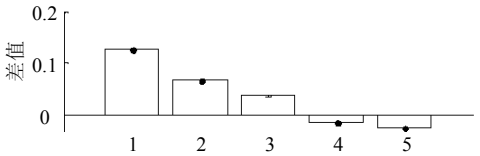
j. 政治



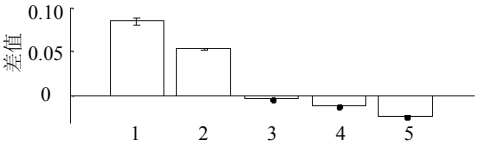
k. 教育



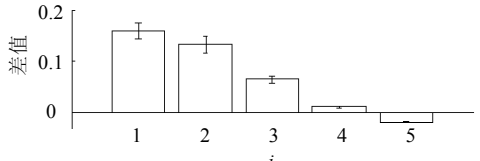
l. 文化



m. 天气



n. 时尚



o. 所有主题

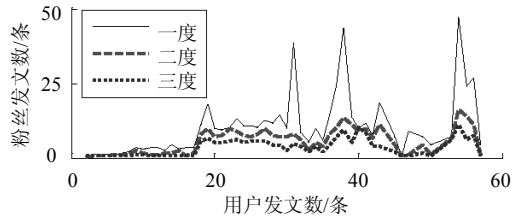
图1　各度粉丝的$p_i-\bar{p}$差值

3.2　发文数量相关性

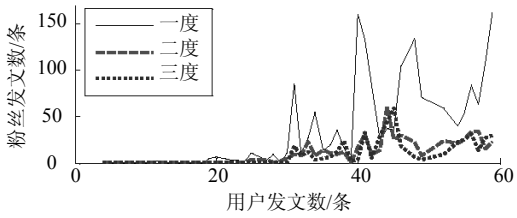
用户与其粉丝的主题兴趣具有三度相关性，本节分析在同一主题下用户三度以内粉丝发文数和用户发文数的相关性。

选取每个主题下发表关于该主题的微博数在1～60之间的用户，分析用户和她/他三度以内粉丝在同一主题下发文数的相关性。图2展示了用户发文数和她/他各度粉丝发文数的变化趋势，横坐标表示用户发表关于某一主题微博的数量，纵坐标表示用户各度粉丝发表相同主题微博的平均值。从图2可以

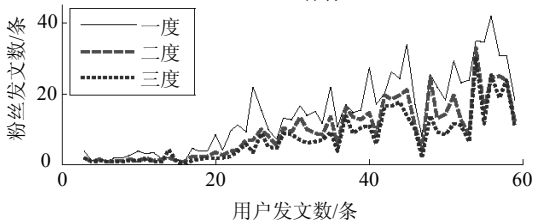
看出各曲线总体呈波浪式上升趋势，当横坐标小于20时，曲线比较平稳，当横坐标大于20时，曲线开始出现剧烈的波动并呈上升趋势，表明用户与其三度以内粉丝发表同一主题微博的数量具有相同增长趋势。造成图2曲线剧烈波动的原因其一是用户各度粉丝可能拥有多个主题兴趣，对不同主题感兴趣程度不同，发文数就会不同，其二是用户及其三度以内粉丝发表行为可能会存在社会影响，不同用户的影响力不同，致使发文数变化趋势出现不稳定现象。此外，从图2也可以看出一度粉丝发表相同主题微博的平均值总体上比二度粉丝高，二度粉丝要比三度粉丝高，说明不仅相同行为相关性随粉丝度数增加而下降，同一主题兴趣下三度以内粉丝发文数随用户发文数增加的幅度也随粉丝度数增加而下降。



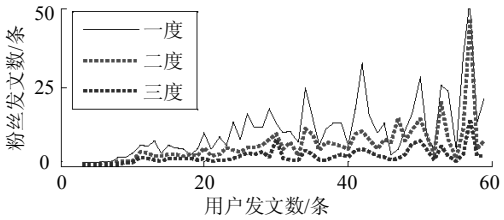
a. 社会



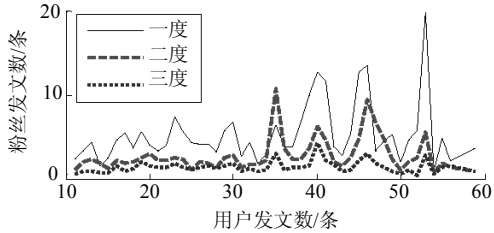
b. 体育



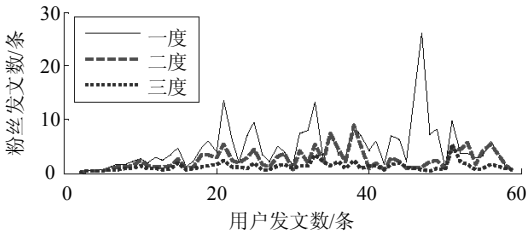
c. 娱乐



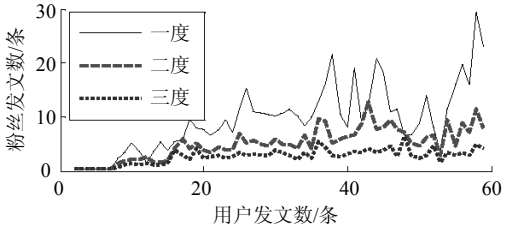
d. 旅游



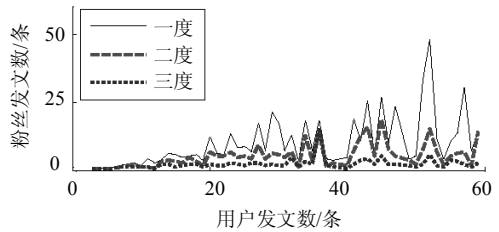
e. 美食



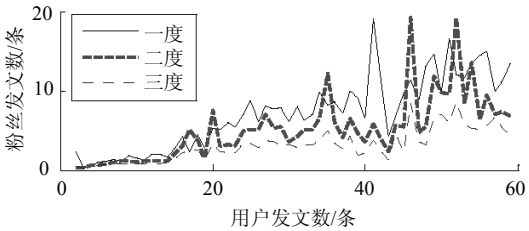
f. 医疗保键



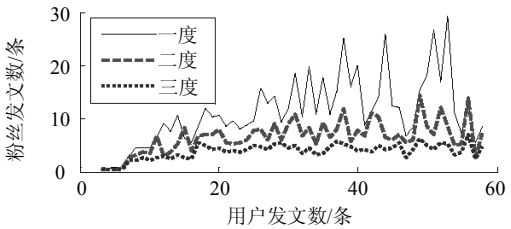
g. 财经



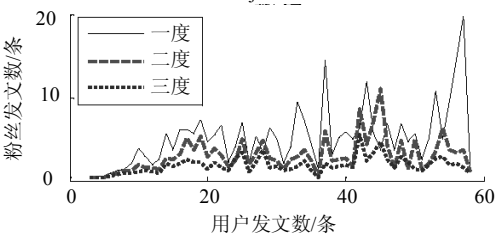
h. 科技



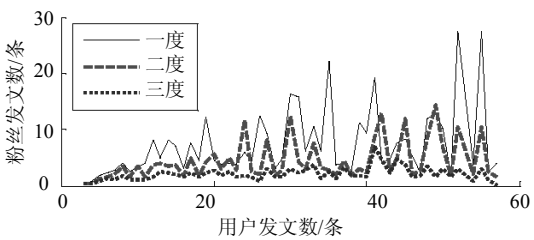
i. 生活情感



j. 政治



k. 教育



l. 文化

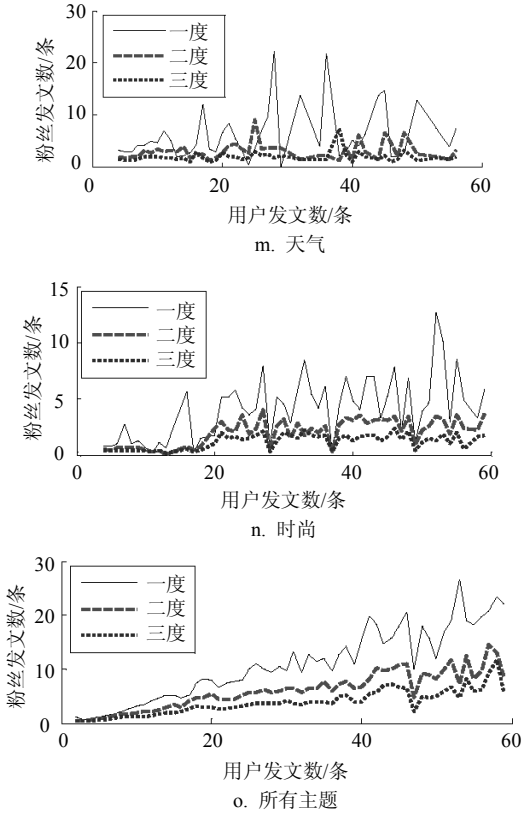


图2　同一主题下用户发文数与其各度粉丝发文数的相关性

3.3　主题兴趣相似性差异检验

如果3.1节、3.2节分析的现象长期存在，用户三度以内粉丝和用户主题兴趣分布就会表现出相似现象，并且相似性随粉丝度数增加而下降，本节检验用户及其三度以内粉丝主题兴趣相似性差异。

用$u_{\text{fl_dr1}}$表示一度粉丝和用户主题兴趣的差异均值，其中fl_dr1表示一度粉丝，同样$u_{\text{fl_dr2}}$、$u_{\text{fl_dr3}}$表示二度、三度粉丝和用户主题兴趣的差异均值。提出以下假设：

假设 1　H_0：$u_{\text{fl\_dr2}}-u_{\text{fl\_dr1}}=0$，$H_1$：$u_{\text{fl\_dr2}}-u_{\text{fl\_dr1}}>0$

假设 2　H_0：$u_{\text{fl\_dr3}}-u_{\text{fl\_dr2}}=0$，$H_1$：$u_{\text{fl\_dr3}}-u_{\text{fl\_dr2}}>0$

用2.1节得到的**DT**矩阵可以很容易判别用户间主题兴趣的差异。利用文献[15]对用户u和v主题兴趣差异的定义：$\text{dist}(u,v)=\sqrt{2D_{\text{JS}}(u,v)}$，可以测量两个用户间主题兴趣的差异，其中$D_{\text{JS}}(u,v)$是两个用户主题分布的Jensen-Shannon散度，$D_{\text{JS}}(u,v)=\frac{1}{2}(D_{\text{KL}}(\mathbf{DT}'_{u\cdot}\|M)+D_{\text{KL}}(\mathbf{DT}'_{v\cdot}\|M))$，$M$是两个概率分布的平均值，$M=\frac{1}{2}(\mathbf{DT}'_{u\cdot}+\mathbf{DT}'_{v\cdot})$，$D_{\text{KL}}$是Kullback-Leibler散度，对于$\mathbf{DT}'_{u\cdot}$和$M$有：

$$D_{\text{KL}}(\mathbf{DT}'_{u\cdot}\|M)=\sum_i \mathbf{DT}'_u(i)\log\frac{\mathbf{DT}'_u(i)}{M(i)}$$

本研究从收集的用户(记为S_u)中随机选取101个一度粉丝数大于30的用户，记为S_u'，设任意用户$s_i\in S_u'$，从S_u选取s_i的一度、二度和三度粉丝，记为S_{i1}、S_{i2}、S_{i3}。计算用户s_i和她/他的某一一度粉丝$s_{i1j}\in S_{i1}$的主题兴趣差异值$\text{dist}(s_i, s_{i1j})$，可得到用户$s_i$ 和她/他的所有一度粉丝主题兴趣的差异均值：$u_{\text{fl\_dr1}}=\frac{1}{|s_{i1}|}\sum_{j=1}^{|s_{i1}|}\text{dist}(s_i,s_{i1j})$。同理可以计算用户$s_i$和她/他二度、三度粉丝主题兴趣的差异均值$u_{\text{fl_dr2}}$和$u_{\text{fl_dr3}}$。

除去一个二度粉丝列表小于30的用户和一个计算无效的用户共得到99个实验用户，对每个用户进行假设检验，假设1和假设2分别有78和71个用户在0.05显著水平下拒绝原假设，用户的一度粉丝与用户的主题兴趣相似性比二度粉丝的大，二度粉丝与用户的主题兴趣相似性比三度粉丝的大。实验结果显示假设1比假设2稳定，这是因为主题兴趣的相似性随着粉丝度数的增加而衰减，达到一定程度后主题兴趣相似现象就会随之消失，随着粉丝度数的增加，高一度粉丝和低一度粉丝与用户的主题兴趣相似性差异会呈现出不稳定性。

4　各类主题微博信息扩散差异分析

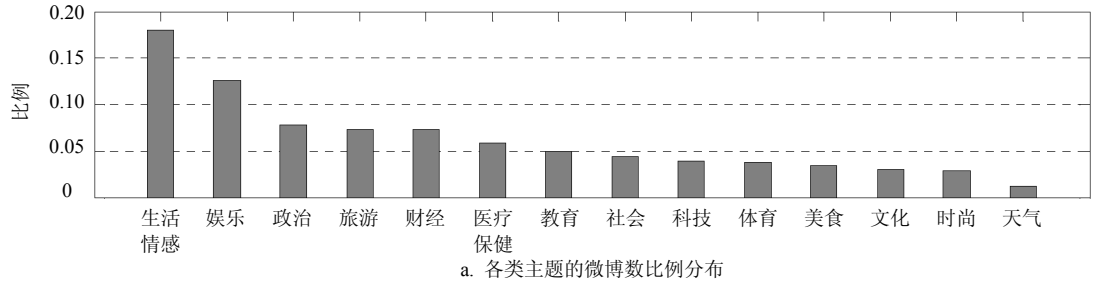
微博扩散树是大量用户转发行为构成的，不同主题类别的微博在用户之间的扩散概率是不同的，本节探讨不同主题类别微博的扩散差异。

4.1　各类主题的热门程度分析

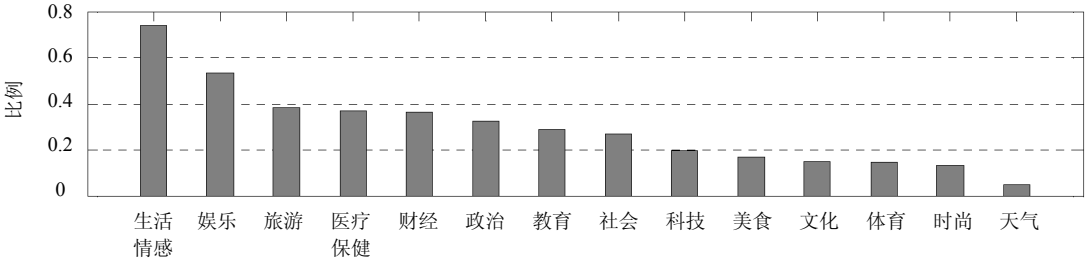
主题类别的热门程度是指主题类别受欢迎程度，本研究从微博数、参与用户数分析各类主题的热门程度，如图3所示。微博数是指属于某主题类别的微博总数，参与用户数是至少发表了5条相关主题类别的微博的用户数。

从图3a可以看出，各类主题的微博数百分比差别显著，数量最多的是生活情感类主题，占所有微博的18.0%，最少的是天气类主题，只有1.2%。图3a表明微博开放环境也为政治、经济类主题的讨论提供了一个很好的平台，两类主题微博数分别占7.8%、7.3%。图3b显示各类主题中参与讨论的用户数百分比最高和最低分别为73.9%和4.7%，分别是生活情感类和天气类，参与各类主题讨论的用户数量差异显著。无论是从微博数还是从参与用户数来看，生活情感类都远多于其他主题类别，微博数和参与用户数排第二位的是娱乐类主题，这与文献[21]的结论不同，其中原因一方面是微博方便即时短消息的发布

传播，另一方面除了组织机构和有特殊目的的用户外，生活情感类主题是每个用户在生活中都会遇到的主题，而这种有关个人生活情感的主题被微博网站列入到热门话题的却不多。另外，天气、时尚、体育、文化、美食等主题类别的微博数较少，只有少部分用户参与讨论。



a. 各类主题的微博数比例分布

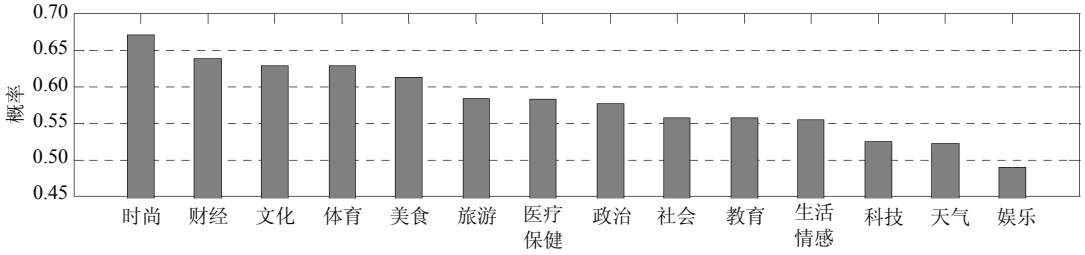


b. 各类主题中参与讨论的用户数比例分布

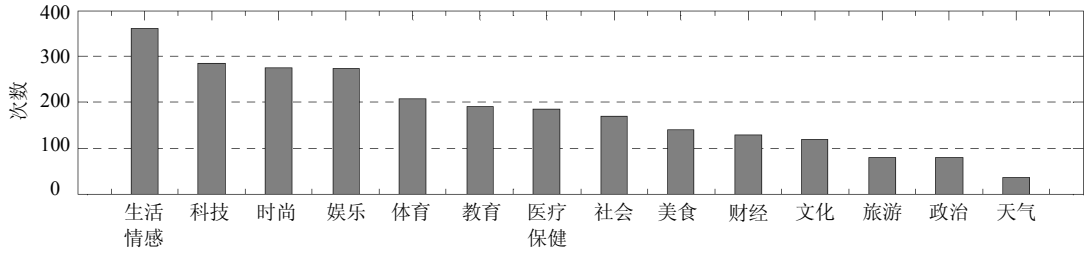
图3 各类主题的热门程度

4.2 各类主题微博的转发率和平均转发数

转发率是指每类主题所有原创微博被转发的比例，平均转发数是指每类主题所有被转发微博的平均转发次数，两者衡量各类主题的微博在微博网络中扩散的可能性和扩散的范围。图4a显示各类主题微博的转发率存在差异，最高的是时尚类，为0.67，最低的则是娱乐类，为0.49。图4b显示各类主题微博的平均转发数最高的是生活情感类，达361.1，最低的是天气类，只有36.5，两者相差近10倍，说明不同主题类别微博的平均扩散范围差别显著。



a. 各类主题的微博被转发概率分布



b. 各类主题的微博被转发次数分布

图4 各类主题的微博被转发概率和平均转发数分布

4.3 各类主题的微博信息扩散树实证分析

微博在微博网络中通过转发机制形成微博信息扩散树，不同主题类别的微博在信息扩散树中表现出不同的特征。不同主题类别微博的平均转发数量差别很大，为了消除转发数量对扩散树深度和扩散时间间隔的影响，本研究从各类主题中选择100～110条转发量在1 000～2 000的微博，追踪这些微博完整的信息扩散树，共获得2 685 154条转发微博。

4.3.1　信息扩散树深度

用信息扩散树的分支深度和平均深度对扩散树深度进行分析。分支深度指从微博原创节点到最后转发节点的路径长度，信息扩散树的平均深度是指所有分支深度的平均值，表3展示了不同主题类别的微博扩散树的平均深度、最大分支深度和深度为1的分支所占比例。各类主题的微博信息扩散树的绝大部分分支深度都比较小，除了政治类，深度为1的分支所占比例均在80%以上，美食类最高，达到96%，深度在2以上的转发节点对信息扩散的贡献较小。各类主题的微博平均扩散深度也比较小，分布在1.10～1.55之间，平均扩散深度相对较大的是政治类、财经类和社会类主题，这三类主题分支深度为1的转发量所占比重相对较小，最大分支深度比较大。

为进一步了解各类主题微博扩散树的结构，本研究分析了不同分支深度的分布，如图5展示了不同主题类别的微博信息扩散树各分支深度的补累积概率分布。图5表明，相对其他主题类别，政治、财经和社会类主题的曲线下降较缓，且政治、财经类曲线较长，而美食、旅游和教育类主题的曲线下降较急。表3和图5说明政治、财经和社会类主题的微博在微博网络上纵向影响力较强，美食、旅游和教育类则较弱。

表3　各主题类别的微博信息扩散树分支深度统计信息

主题类别	平均深度	最大分支深度	深度为1所占比例	主题类别	平均深度	最大分支深度	深度为1所占比例
社会	1.38	14	0.80	科技	1.29	13	0.86
体育	1.16	10	0.91	生活情感	1.13	12	0.94
娱乐	1.22	11	0.89	政治	1.55	16	0.72
旅游	1.11	13	0.93	教育	1.14	11	0.91
美食	1.10	15	0.96	文化	1.24	16	0.89
医疗保健	1.18	13	0.91	天气	1.16	15	0.92
财经	1.34	20	0.82	时尚	1.17	13	0.93

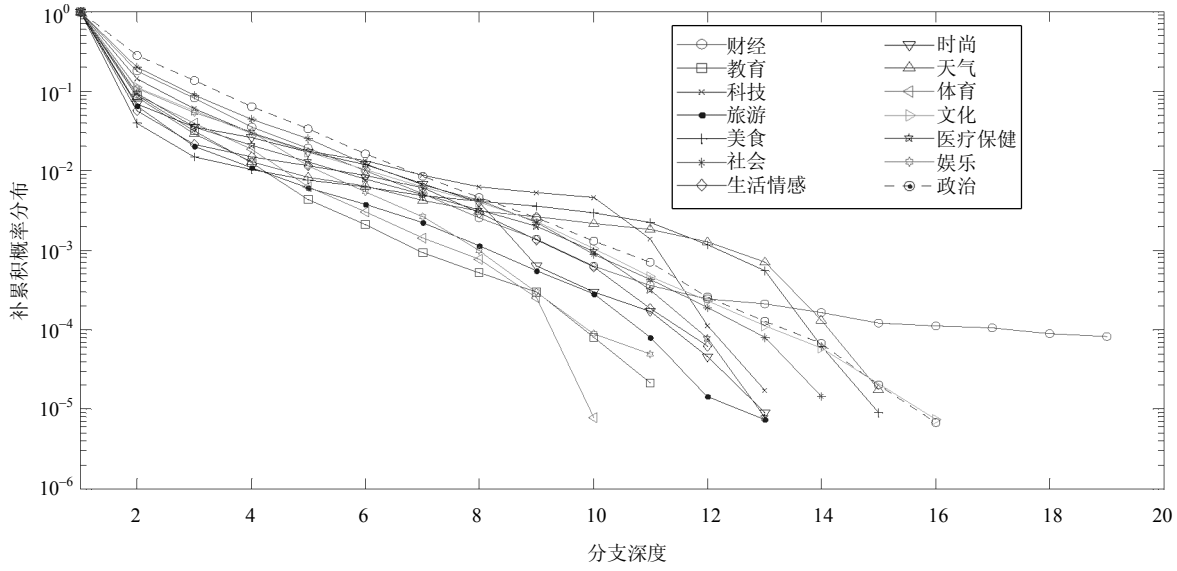


图5　信息扩散树分支深度的补累积概率分布

4.3.2　扩散时间间隔

扩散时间间隔是指原创微博的时间和转发微博的时间差。对不同主题类别微博的扩散时间间隔进行分析，发现各主题类别微博的扩散时间间隔在1小时内的占20%～40%，在1天内的占78%～91%，少量时间间隔是在1个月以上，表明微博信息扩散具有很强的时效性。

表4给出了转发时间间隔均值、中位数和第3分位数，可以看出三者最高和最低分别都是政治类和天气类，说明天气类微博时效性远比政治类强。其中原因是天气类微博多为预报信息，这种信息逾期后很少人再关注，政治类主题包含国际历史、国际社会和国家长期的方针政策，这类主题大部分是长期型的，讨论持续的时间较长。

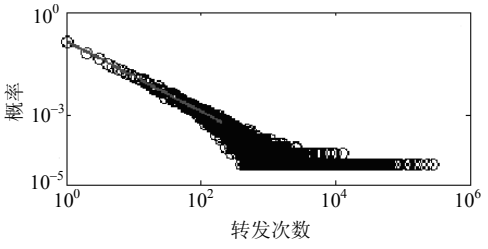
表4 各主题类别微博的扩散时间间隔统计信息

主题类别	均值/h	达50%/h	达75%/h	主题类别	均值/h	达50%/h	达75%/h
社会	87.99	2.92	13.36	科技	68.09	2.42	11.91
体育	93.32	1.76	8.72	生活情感	127.50	2.67	15.78
娱乐	97.38	2.78	13.56	政治	406.81	7.20	49.04
旅游	46.78	2.36	21.84	教育	196.08	2.24	17.50
美食	119.96	4.48	39.17	文化	401.55	2.93	19.58
医疗保健	233.36	2.26	15.60	天气	34.77	1.31	3.96
财经	270.08	3.45	15.62	时尚	111.54	3.54	31.54

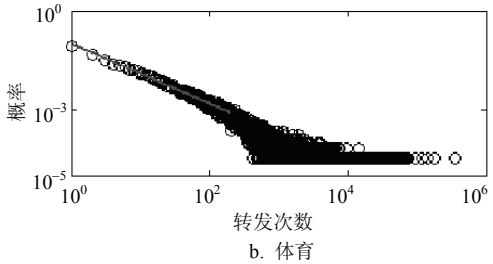
4.3.3 用户的信息扩散能力分析

扩散树深度纵向衡量了信息在微博网络的影响力，要测量用户在整个网络中的信息扩散能力，还需要考虑用户各度粉丝的扩散能力[16]。在微博网络中用户扩散能力表现为用户发表的微博在整个网络的扩散范围，本研究用用户发表的微博被其各度粉丝转发的次数来衡量用户的扩散能力。

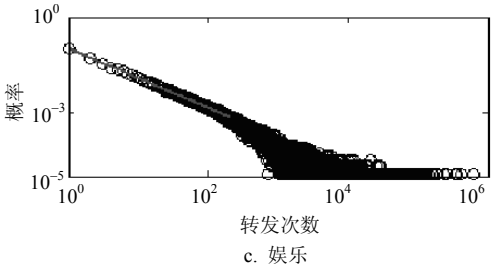
图6展示了不同主题类别的用户扩散能力分布，表明各类主题下用户扩散能力近似幂律分布，其幂指数如表5所示。一般来讲幂律分布指数越小，个体差异越大。分析发现不同主题用户扩散能力幂指数普遍偏小，在0.95左右浮动，表明用户的扩散能力差别很大，部分主题类别的用户扩散能力分布存在差异。



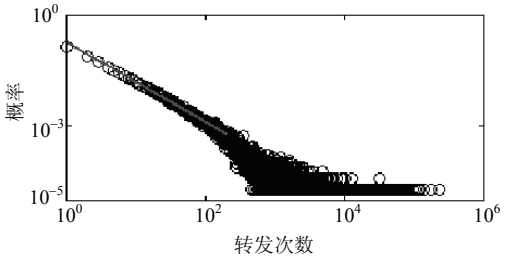
a. 社会



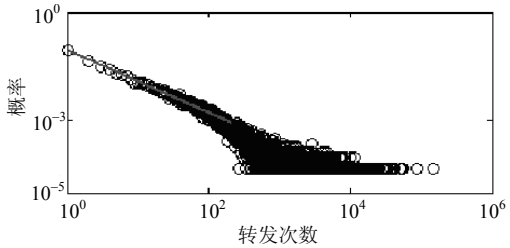
b. 体育



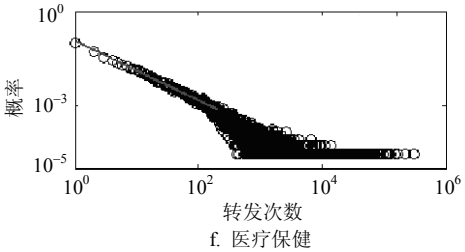
c. 娱乐



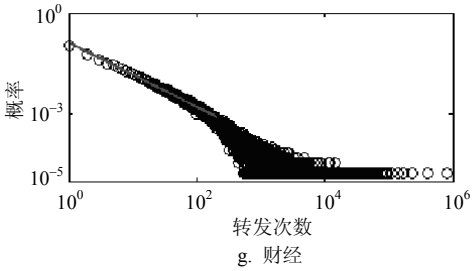
d. 旅游



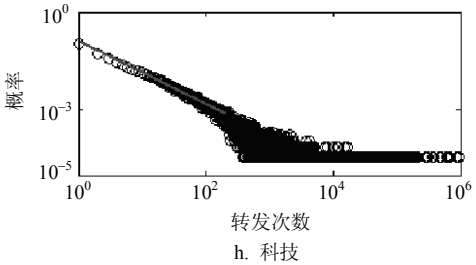
e. 美食



f. 医疗保健



g. 财经



h. 科技

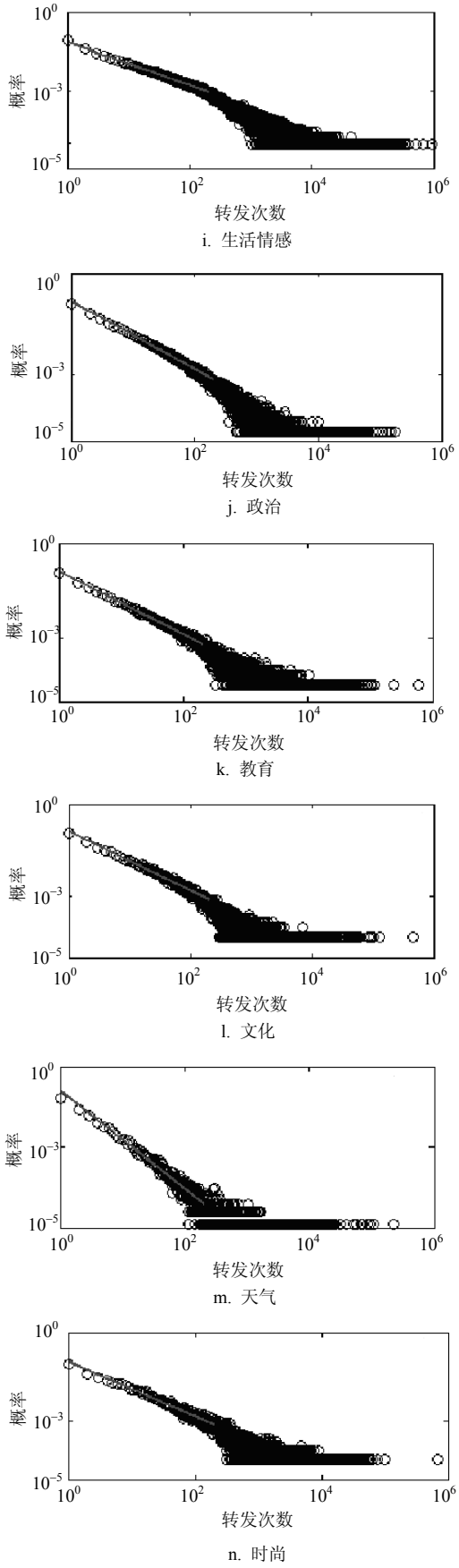


图6　各类主题的用户扩散能力分布

表5　各类主题的用户扩散能力分布的幂指数

主题类别	幂指数	主题类别	幂指数
社会	1.01	科技	0.96
体育	0.91	生活情感	0.82
娱乐	0.94	政治	1.01
旅游	1.08	教育	0.99
美食	0.87	文化	0.97
医疗保健	0.95	天气	1.22
财经	0.96	时尚	0.93

5　结 束 语

本文在主题分析的基础上，探讨了用户主题兴趣的相关性，并分析了不同主题类别微博的信息扩散差异，发现：

1) 用户主题兴趣表现出三度相关性，同一主题下用户三度以内粉丝发文数随用户发文数的增加而增加，增加幅度随粉丝度数增加而下降；

2) 用户多级粉丝与用户主题兴趣的相似性随着粉丝度数的增加而递减；

3) 新浪微博的用户更喜欢和生活相关的信息；

4) 不同主题类别的微博被转发的概率和平均转发次数存在显著差异；

5) 不同主题类别的微博在扩散深度、用户扩散能力、扩散时间间隔上都表现出不同的特征。

本文为今后构建考虑用户兴趣和不同主题信息扩散的特征，且与微博网络环境接近的信息扩散模型提供了参考。

但仍存在需要改进的地方。除了文本信息，微博还包含图片、视频、超链接等其他信息，这给主题分析带来了更大的挑战。更科学合理的主题分类方法和主题分类粒度大小会得到更准确的结论。此外，用户主题兴趣三度相关性的形成过程中，社会影响与趋同性是否均在发挥作用？若是的话二者各自的权重有多少？这也是今后值得研究的问题。

参 考 文 献

[1] 陆豪放, 张千明, 周莹, 等. 微博中的信息传播: 媒体效应与社交影响[J]. 电子科技大学学报, 2014, 43(2): 167-173.
LU Hao-fang, ZHANG Qian-ming, ZHOU Ying, et al. Information spreading in microblogging systems: Media effect versus social impact[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(2): 167-173.

[2] ASUR S, HUBERMAN B A. Predicting the future with social media[C]//IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Washington, DC: IEEE Computer Society, 2010:

492-499.

[3] BOLLEN J, MAO H, ZENG X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[4] 苏萌, 柏林森, 周涛. 个性化: 商业的未来[M]. 北京: 机械工业出版社, 2011.
SU Meng, BO Lin-sen, ZHOU Tao. The future of personalized business[M]. Beijing: China Machine Press, 2011.

[5] 李栋, 徐志明, 李生, 等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1), 189-206.
LI Dong, XU Zhi-ming, LI Sheng, et al. A survey on information diffusion in online social networks[J]. Chinese Journal of Computers, 2014, 37(1), 189-206.

[6] SUH B, HONG L, PIROLLI P, et al. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network[C]//IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust. Washington, DC: IEEE Computer Society, 2010: 177-184.

[7] 闫强, 吴联仁, 郑兰. 微博社区中用户行为特征及其机理研究[J]. 电子科技大学学报, 2013, 42(3): 328-333.
YAN Qiang, WU Lian-ren, ZHENG Lan. Research on behavior characters and mechanism in microblog communities[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(3): 328-333.

[8] YANG Z, GUO J, CAI K, et al. Understanding retweeting behaviors in social networks[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010: 1633-1636.

[9] 张旸, 路荣, 杨青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4): 109-114.
ZHANG Yang, LU Rong, YANG Qing. Predicting retweeting in microblogs[J]. Journal of Chinese Information Processing, 2012, 26(4): 109-114.

[10] 张亚明, 唐朝生, 李伟钢. 微博机制和转发预测研究[J]. 情报学报, 2013, 32(8): 868-876.
ZHANG Ya-ming, TANG Chao-sheng, LI Wei-gang. Research on the micro-blog mechanism and re-posting prediction[J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(8): 868-876.

[11] 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790.
CAO Jiu-xin, WU Jiang-lin, SHI Wei, et al. Sina Microblog information diffusion and prediction[J]. Chinese Journal of Computers, 2014, 37(4): 779-790.

[12] WU S, TAN S, KLEINBERG J, et al. Dose bad news go away faster?[C]//Proceeding of the Fifth International AAAI Conference on Weblogs and Media. Menlo Park, CA: AAAI Press, 2011: 646-649.

[13] ROMERO M D, MEEDER B, KLEINBERG J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter[C]//Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 695-704.

[14] LIU H, XIE Y, HU H, et al. Affinity based information diffusion model in social networks[J]. International Journal of Modern Physics C, 2014, 25(5): 1440004.

[15] WENG J, LIM E P, JIANG J, et al. TwitterRank: Finding topic-sensitive influential twitterers[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261-270.

[16] LIU L, TANG J, HAN J, et al. Mining topic-level influence in heterogeneous network[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010: 199-208.

[17] CHRISTAKIS N A, FOWLER J H. Social contagion theory: Examining dynamic social networks and human behavior[J]. Statistics in Medicine, 2013, 32(4): 556-577.

[18] FAN R, ZHAO J, CHEN Y, et al. Anger is more influential than joy: Sentiment correlation in Weibo[J]. Plos One, 2014, 9(10): e110184.

[19] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 591-600.

[20] ZHAO W X, JIANG J, WENG J, et al. Comparing Twitter and traditional media using topic models[C]//Proceedings of the 33rd European Conference on Advances in Information Retrieval. New York: ACM, 2011: 338-349.

[21] YU L, ASUR S, HUBERMAN B A. Dynamics of trends and attention in Chinese social media[EB/OL]. (2013-12-02). http://arxiv.org/abs/1312.0649v1.

[22] 新浪网. 微博开放平台[EB/OL]. [2015-01-15]. http://open.weibo.com/.
Sina. Weibo open platform[EB/OL]. [2015-01-15]. http://open.weibo.com/.

[23] 张华平. NLPIR 汉语分词系统[EB/OL]. [2015-01-15]. http://ictclas.nlpir.org/.
ZHANG Hua-ping. ICTCLAS2013[EB/OL]. [2015-01-15]. http://ictclas.nlpir.org.

[24] BROWN G, YULE G. Discourse analysis[M]. Cambrige, UK: Cambrige University Press, 1983.

[25] 陈文涛, 张小明, 李军舟. 构建微博用户兴趣模型的主题模型的分析[J]. 计算机科学, 2013, 40(4): 127-130.
CHEN Wen-tao, ZHANG Xiao-ming, LI Jun-zhou. Analysis of topic models on modeling microblog user interestingness[J]. Computer Science, 2013, 40(4): 127-130.

编 辑 蒋 晓