研究论文

基于多策略的群聊话题检测技术*

吴　旭[1,2,3]　陈春旭[1,2]

[1]（北京邮电大学可信分布式计算与服务教育部重点实验室 北京 100876）

[2]（北京邮电大学网络空间安全学院 北京 100876）

[3]（北京邮电大学图书馆 北京 100876）

摘要：【**目的**】更好地解决群聊话题纠缠的问题，减少稀疏文本特征对聚类的影响，实现对多类型消息混合的连续群聊信息的话题检测。【**方法**】提出一种基于多策略的群聊话题检测技术，通过构建话题序列解决话题交叉，利用消息的用户、时间、类型等属性提升聚类效果。【**结果**】本方法处理三份群聊记录样本的纯文本数据时的F值较对比算法分别提升2.9%、6.1%和3.0%，速度分别提高约27.6%、32.1%和47.1%。本方法还能处理传统算法无法应对的混合类型数据，且比处理对应的纯文本数据时的性能分别提升约29.4%、27.1%和22.5%。【**局限**】对群聊消息文本特征的利用率不足，算法所设阈值过多。【**结论**】本文方法能够在一定程度上提高群聊话题检测效果，并扩大了话题检测所能应对的消息类型的广度，提升了舆情分析效率。

关键词：群聊消息　话题检测　短文本

分类号：TP391

DOI：10.11925/infotech.2096-3467.2020.0718

引用本文：吴旭，陈春旭．基于多策略的群聊话题检测技术[J]．数据分析与知识发现，2021，5(5)：1-9.(Wu Xu, Chen Chunxu. Detecting Topics of Group Chats with Multiple Strategies[J]. Data Analysis and Knowledge Discovery, 2021, 5(5): 1-9.)

1　引　言

在互联网及其配套设备普及率不断提高的今天，即时消息系统的用户群也在持续壮大，中国互联网络信息中心发布的第44次《中国互联网络发展状况统计报告》[1]显示，截至2019年6月，中国网民各类互联网应用的用户规模中，即时通信以82 470万排名第一，网民使用率高达96.5%，半年增长率达4.2%；而在各类手机互联网应用的用户规模中，手机即时通信也以82 069万位居首位，其使用率高达96.9%，半年增长率达5.2%。由此可见，国内网民已逐渐离不开QQ、微信等即时通信工具，而群聊作为这些应用的主要功能，也愈加频繁地出现在人们的日常生活中。对普通用户来说，大量复杂的群聊消息很难被快速消化，同样，群聊消息的上述特性也使得对其监管需要消耗大量人力。要解决以上两个主要矛盾，就必须使用计算机对群聊进行舆情分析，话

通讯作者（Corresponding author）：吴旭（Wu Xu），ORCID：0000-0002-1297-2726，E-mail：wux@bupt.edu.cn。

*本文系国家重点研发计划基金项目（项目编号：2017YFC0820603）、国家自然科学基金项目（项目编号：62072488）和北京市自然科学基金项目（项目编号：4202064）的研究成果之一。

The work is supported by the National Key Research and Development Plan (Grant No. 2017YFC0820603), the National Natural Science Foundation of China (Grant No. 62072488), the Natural Science Foundation of Beijing,China (Grant No. 4202064).

题检测是群聊文本分析的一个重要研究方向[2]。

本研究通过构建、维护话题序列的方法解决话题纠缠，并利用群聊消息的用户、时间、类型等属性强化属于同话题的消息之间的联系，弥补单纯依靠稀疏的短文本特征展开聚类的不足，同时也实现了将非文本类型或只含标点、停用词等无实义内容的文本类型消息加入聚类过程，处理完整、连续的群聊记录，更加准确地检测群聊话题内容，确定参与话题的用户群体，提升群聊舆情分析效果。

2 研究现状

在文本特征提取方面，研究者提出了许多方法，例如：词、短语、N-grams[3-4]；分类树或本体[5-8]；意见、感情等领域特征[9-11]；嵌入特征，如 Word2Vec[12-13]、GloVe 等；主题特征[14-17]；相异空间[18]等。由于单条群聊信息的内容往往过于简短，且句式、语法多不规范，因此在提取文本特征时，使用关注词语特征的 Word2Vec 等方式能够取得更佳效果，使用适于提取长篇文档主题特征或其他特征的方法则效果不佳。在话题检测方面，早期的研究多采用监督学习的方法，Elnahrawy[19]发现朴素贝叶斯分类器对话题检测的效果最好，Özyurt 等[20]引入了模式匹配思想，使用支持向量机分类法使效果达到最优，但监督学习的方法在处理数量庞大的群聊信息时会耗费大量时间，效率极其低下。在无监督学习方法中，Adams 等[21]针对短文本词向量稀疏的特性，以时间为惩罚项，同时利用 WordNet 语义网扩充短文本内容，取得了一定成效，但算法所依赖的语义特征的准确度与语义网规模和更新的及时性有关，因此很难应对多领域的文本内容以及新生词汇；Wang 等[22]利用用户上下文与“@”关系判定消息的话题归属，但在实际场景中，“@”在大量群聊记录中的出现次数并不多，也不一定标志着两条消息间的“回复”与“被回复”关系，因此很难带来话题检测效果的明显提升；李天彩等[23]则结合消息的内容、用户、时间属性来提高会话抽取效果，但该算法只能处理去除了无文本消息的间断群聊记录，所以在根据相邻消息数计算用户亲密度与根据时间间隔划分会话片段时容易产生较大误差，导致算法性能不稳定。

由于群聊文本有特征稀疏性、奇异性、动态性和交错性等特点[24]，尽管针对群聊的话题检测已历经多年的发展，但仍存在很大提升空间。此外，现有研究往往忽略了对非文本类型的消息和只含停用词或标点符号消息的处理，而在实际应用场景中，这样会遗漏大量消息，导致无法准确划分会话片段和定位参与话题的用户群体，不能很好地根据聊天记录进行用户画像[25]、分析用户性别[26]等，也容易丧失对部分关键的非文本信息内容的关注，导致舆情分析结果的准确度与可信度不足。

3 研究思路与框架

本文综合考虑群聊消息的内容及用户、时间、类型等辅助信息，研究了它们在话题检测中的作用，结合实际聊天场景，提出一种基于多策略的群聊话题检测技术。通过构建话题序列解决话题交叉的问题，利用辅助信息减少短文本特征稀疏对聚类效果的影响，实现对多种类消息混合的连续群聊记录的话题检测。

3.1 话题序列

群聊和私聊有所不同的根本原因在参与人数。在点对点聊天的情境下，一个话题的产生必然伴随着上一个话题的消亡，而群成员数量众多，两个话题的参与群体若不完全重合，则有交叉并行的可能，即一段连续出现的消息中，属于不同话题的消息会交替出现。群聊话题提取必须要进行话题分割，分割的关键在于正确判断一条新消息是否属于某个已有的话题以及具体属于哪个话题。本研究参考真实群聊情境，通过话题序列对当前群聊可能讨论的话题进行识别与排序，使新消息能够以较高的概率匹配到其所属的话题中，达到解决话题交叉并行问题的目的。

群聊消息具有文本内容、发送用户、发送时间、消息类型等属性，而话题序列中的话题具有话题的持续时间、最近一段时间内的消息频率和话题热度三个重要属性，这三个属性的计算方法如公式(1)-公式(3)所示。

$$t.duration = m_{recent}.time - \min_{m_i \in t}(m_i.time) \quad (1)$$

$$t.frequency = \frac{\mathrm{count}_{m_j \in t}(m_j)}{H_t},$$

$$m_{recent}.time - m_j.time < H_t \quad (2)$$

$$t.popularity = \frac{t.frequency}{t.duration} \quad (3)$$

其中，$t.duration$ 表示话题 t 的持续时间，$m_{recent}.time$ 是最新一条群聊消息的发送时间，$\min_{m_i \in t}(m_i.time)$ 表示话题 t 中最早一条消息的发送时间；$t.frequency$ 表示话题 t 最近一段时间的消息频率，这段时间为 H_t，实验中取经验值；$\text{count}_{m_j \in t}(m_j)$ 表示属于话题 t 的最近 H_t 时间内的消息数；$t.popularity$ 表示话题 t 的热度，它与话题持续时间成反比，与话题最近一段时间的消息频率成正比。

话题序列的结构如图 1 所示。

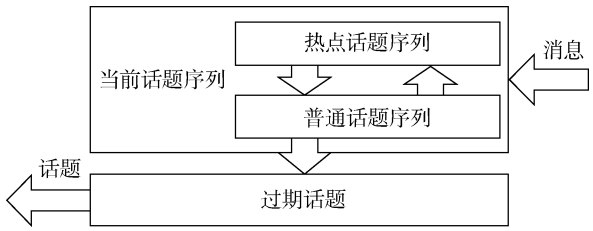

图 1　话题序列

Fig.1　Topic Sequence

话题序列中的话题有当前话题和过期话题两类。过期话题是从当前话题序列淘汰的不再更新的话题，新消息不会加入这些话题，它们将作为历史数据被储存起来；当前话题序列分为普通话题序列和热点话题序列，新消息将会加入这两个序列中的话题，或在这两个序列中开启新话题。话题序列的更新机制如算法 1 所示。

算法 1 话题序列更新

输入：当前话题序列 $S=[t_1, t_2, \cdots]$，话题的热点判定频率 H_f，普通话题序列大小 $seqsize$。

输出：热点话题序列 $hTopics$，普通话题序列 $cTopics$，过期话题集合 $eTopics$。

for t_i in S
　if $t_i.frequency > H_f$
　　将 t_i 加入 $hTopics$
　else
　　将 t_i 加入 $cTopics$
将 $cTopics$ 中所有 t 按属性 $popularity$ 升序排列
$n = cTopics$ 中比 $seqsize$ 多出的话题数
if $n > 0$
　for t_j in $cTopics$
　　将 t_j 加入 $eTopics$
　　将 t_j 从 $cTopics$ 中删除
　　$n = n - 1$
　　if $n < 1$
　　　终止循环
输出 $hTopics$，$cTopics$，$eTopics$

3.2　群聊消息的辅助信息

群聊会产生大量由纯符号或停用词组成的文本信息，以及图片、视频等非文本信息，本文将其称为无义信息，其他信息则为有义信息。由于群聊消息文本长度较短，具有特征稀疏的缺点，单纯依靠语义信息难以提高话题检测性能，而且，保证群聊记录的连续性和话题参与群体的完整性也对群聊舆情分析至关重要，因此本文提出一种利用辅助信息进行话题检测的方法。辅助信息主要有两种，一种是群聊消息的类型，另一种是消息的发信时间与用户，通过研究二者在真实群聊情境中与话题分割之间的关系，本文将总结出的规律应用到聚类中以提高话题检测性能。

（1）消息的类型属性。以微信为例，群聊消息的类型主要有“文本”“图片”“视频”“链接”等，其中只有“文本”类型直接包含语义信息，支持采用传统的计算词向量相似度等方式进行话题聚类，但其余没有文本内容的消息也并非不能对聚类过程提供帮助。提取群聊话题必须要进行话题分割，本文分割的依据之一是各话题的第一条消息，即起始消息。

与当前话题序列中任何一个话题的语义相似度都很低的文本类型消息会开启一个新的话题，此时它便是一条起始消息，同样，非文本类型的消息也有可能引起一个新的话题。通过研究大量群聊记录样本，发现一个话题除了被一段文字引起，还有可能由一张图片、一段视频、一个链接引起，特别是当发出该条消息的用户未参与最近一段时间的聊天的情况下，因此，在处理群聊数据时，为每条消息添加了“是否为起始消息”这一属性，当消息为非文本类型时，其计算方法如公式（4）所示。

$$m.start = \begin{cases} \text{True}, m.type = \text{"picture", "video"or"link"} \\ \text{False}, \text{otherwise} \end{cases}, \quad m.type \neq \text{"text"} \quad (4)$$

其中，*m.start*表示消息*m*是否可能为起始消息的标志，*m.type*表示消息*m*的类型。对于文本类型消息，通过研究大量群聊内容，结合汉语语法习惯，本文总结出以下几种可用来判定非起始消息的文本特征：句首为连词或副词；句尾是“吧”等特定语气词；句中使用了包含“你”“他”“这”等字的指示代词；不含代词与名词。因此在聚类过程中，文本内容具备以上特征之一的消息*m*应有*m.start* = False。

（2）消息的用户与时间属性。群聊消息具有发送时间戳与发送用户ID两个附加信息。按照“一定时间内，一个用户所发送的消息极可能属于同一话题”的原则，取话题热度检测时间H_t，当新消息为无义消息或文本内容不与任何话题足够相似时，若寻找到H_t时间内同一用户发送的消息，则新消息将加入满足上述条件的最新一条消息所属的话题。

3.3 使用话题序列与群聊辅助信息的Single-Pass聚类算法

本文所提出的话题检测算法SP_{TSAI}（Single-Pass Using Topic Sequence and Auxiliary Information）通过构建话题序列解决话题交叉问题，综合利用文本特征和时间、用户、消息类型等辅助信息提高聚类效果。聚类过程中，语义相似度计算主要发生在消息与话题间、话题与话题间，其中后者是为了应对Single-Pass聚类算法易形成小簇的问题，即避免短文本特征稀疏导致话题分割粒度过小。两种相似度计算方法如公式（5）所示。

$$sim(x,t_i)=\begin{cases}\cos sim\left(\boldsymbol{x.vector},\dfrac{\sum_{k=1}^{V_2}\boldsymbol{m_k.vector}}{V_2}\right), x \text{ is a message}\\ \cos sim\left(\dfrac{\sum_{n=1}^{V_1}\boldsymbol{m'_n.vector}}{V_1},\dfrac{\sum_{k=1}^{V_2}\boldsymbol{m_k.vector}}{V_2}\right), x \text{ is a topic}\end{cases},\quad t_l=\{m_1,m_2,\cdots,m_{V_2}\}\{m'_1,m'_2,\cdots,m'_{V_1}\} \tag{5}$$

其中，$sim(x,t_i)$表示对象x与话题t_i之间的语义相似度；文本特征向量方面，如$\boldsymbol{m_k.vector}$用于表示消息m_k的文本特征向量；cos*sim*()则是计算两个向量间余弦相似度的函数，计算方法如公式（6）所示。

$$\cos sim(\boldsymbol{v}_1,\boldsymbol{v}_2)=\frac{\boldsymbol{v}_1\times\boldsymbol{v}_2}{|\boldsymbol{v}_1||\boldsymbol{v}_2|} \tag{6}$$

SP_{TSAI}的主要流程为：对每一条新的群聊消息，将它们按一定算法加入当前话题序列中，之后将语义相似度高于阈值的一些话题合并，最后按照算法1更新话题序列，得到当前的热点话题、普通话题和过期话题。新消息加入话题序列的具体流程如算法2所示。

算法2 消息加入话题序列

输入：待加入的消息*m*，当前话题序列$S=[t_1,t_2,\cdots,t_V]$，消息与话题的相似度阈值T_t，话题热度检测周期H_t。

输出：更新后的当前话题序列*S*。

```
if S不为空
    if m的内容包含能够被向量化的有义文本
        maxvalue = 0
        for t_i in S
            simvalue = sim( m , t_i )
            if simvalue > maxvalue
                maxvalue = simvalue
                t_max = t_i
            if maxvalue > T_t
                将m加入t_max
                return S
    minvl = H_t
    for t_j in S
        for m_k in t_j
            interval = m.time - m_k.time
            if m.user == m_k.user且interval < minvl
                minvl = interval
                t_recent = t_j
    if minvl == H_t
        if m.start == True
            将t_{V+1}加入S
            将m加入t_{V+1}
            return S
        recentime = 0
        for t_i in S
            maxtime = max_{m_n ∈ t_i}(m_n.time)
```

if $maxtime > recentime$
$recentime = maxtime$
$t_{recent} = t_l$
将 m 加入 t_{recent}
else
将 t_l 加入 S
将 m 加入 t_l
return S

算法 2 的思路为：若当前话题序列为空，则以新消息开始一个新话题，若序列中已存在其他话题，则在相应指标满足阈值要求的情况下加入语义相似度最高或该用户最近参与的话题，如果未能加入任何话题，判断该消息是否为起始消息，是则开始一个新话题，否则加入上一条消息所在的话题。

4　实验过程

4.1　实验数据

本文实验使用的文本向量是在网络上搜集的约 26 GB 百科数据、13 GB 新闻数据和 229 GB 小说数据合并得到的语料上使用 Word2Vec 训练并处理得到的。选用 Word2Vec 的原因是：相较长文本，短文本特征对词序、句式的依赖小得多，使用 Word2Vec 能在提取较好文本特征的前提下获得更高效率。作为数据集的群聊信息则是自行收集所得，数据集 D 涵盖多个群，时间范围为 2019 年 11 月 25 日至 2020 年 2 月 15 日，共有消息 102 083 条，参与用户 823 人。本实验随机选取其中三个群的消息，并将它们分别记为 D_1、D_2、D_3，D_1 来自一个互联网职业交流群，D_2 来自一个电影讨论群，D_3 来自一个教师通知群，具体信息如表 1 所示。

表 1　群聊数据集信息

Table 1　Information of Dataset

数据集	消息数	参与的用户数	话题数	平均汉字数	无义消息占比
D	102 083	823	NA	NA	NA
D_1	9 024	117	1 413	8.28	21.80%
D_2	3 690	148	855	7.35	24.44%
D_3	636	68	303	26.49	29.40%

4.2　实验环境

本实验的系统环境为 Windows 10，版本号 1909（OS 内部版本 18363.959）；所用语言为 Python 3.7.6，集成开发环境为 PyCharm 2019.3.1（Community Edition）。

4.3　评价指标

本文所采用的对照算法 SP_{TSWKV} 来自文献[23]，其性能已被证明优于经典的 SP_B、SP_{WC}、SP_{NN} 和 SP_{WNN} 等算法[27]，是近年在群聊话题检测领域性能较好的算法之一，其基本原理是：综合计算相邻消息间的文本特征、发信用户亲密度数值及时间距离，将达到设定阈值的消息分至同一话题，并对时间距离较小的话题再进行合并。评价指标选用准确率、召回率与 F 值，计算方法如公式(7)-公式(10)所示。

$$P(i,j) = \frac{N_{ij}}{N_j} \tag{7}$$

$$R(i,j) = \frac{N_{ij}}{N_i} \tag{8}$$

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j) + R(i,j)} \tag{9}$$

$$F = \sum_i \frac{N_j}{N} \max_j (F(i,j)) \tag{10}$$

其中，$P(i,j)$、$R(i,j)$ 和 $F(i,j)$ 分别是人工标注话题 i 在算法检测出的话题 j 中的准确率、召回率和 F 值；N 为总消息数，N_i 为话题 i 中的消息数，N_j 则为话题 j 中的消息数，N_{ij} 为话题 i 与 j 中相同消息的条数；F 表示该群所有话题 F 值的均值，是算法性能的主要评价指标。

4.4　实施过程

本实验首先通过逆向技术采集针对 PC 版微信客户端接收的群聊消息，再从主题不同的群中选取几段连续的群聊记录作为样本数据集。考虑 SP_{TSWKV} 算法不能处理无义消息，因此在开展第一组对照实验时，选取在样本数据集中去除无义消息所得的群聊记录作为输入数据，对两种方法进行调优后评价话题检测效果。第二组实验再次使用 SP_{TSAI} 算法处理完整的样本数据集，对比在处理去除与保留无义消息的群聊记录时 SP_{TSAI} 算法的性能，并对 SP_{TSAI} 算法的几个重要参数进行分析。

5　实验结果

对本文所提出的群聊话题检测方法 SP_{TSAI}，阈值

方面，话题热点判定频率H_f和普通话题序列大小*seqsize*分别取5条/分钟和3。由于作为对照的SP_{TSWKV}算法不能处理无义消息，故对照实验所采用的数据集是在D_1、D_2、D_3的基础上过滤掉无义消息得到的。主要评价指标为平均F值，对SP_{TSWKV}算法的两个阈值T_{time}和T_c以及本文SP_{TSAI}算法的三个阈值T_t、T_f、H_t进行遍历，取各方法的最优结果进行对比，实验结果如表2所示。

表2　在去除无义消息的数据集上的实验结果

Table 2　Experimental Results on Data Sets Without Nonsense Messages

算法	指标	D_1	D_2	D_3
SP_{TSWKV}	T_{time}	490	530	950
	T_c	0.10	0.10	0.15
	F	0.645	0.594	0.698
SP_{TSAI}	T_t	0.20	0.70	0.65
	T_f	0.75	0.30	0.05
	H_t	5	4	7
	F	0.664	0.630	0.719

可以看出，本文方法比SP_{TSWKV}算法在三个数据集上的F值分别高约2.9%、6.1%和3.0%，总体提升幅度并不大，这是因为SP_{TSWKV}算法采用TF-IDF与Word2Vec相结合的方式，更充分地利用了消息的文本特征，而且删除无义消息导致聊天记录残缺不连续，从而无法充分发挥本文方法借助辅助信息改善聚类效果的优势。但评价算法综合性能还需考察运行速度，两种方法的运行时间如图2所示。

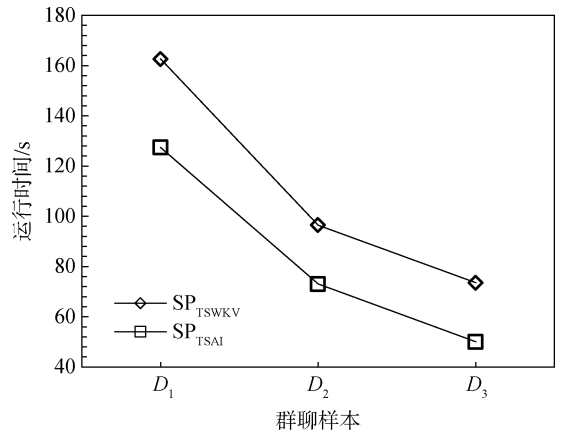

图2　两种方法的运行时间

Fig.2　Running Time of Two Methods

可以看出，本文方法运行速度比SP_{TSWKV}算法在三个群上分别提高约27.6%、32.1%和47.1%，这是因为SP_{TSWKV}算法虽然也综合考虑时间、用户关系，并在一定程度上降低了短文本特征稀疏对聚类的影响以优化聚类效果，但它对每一属性均进行量化表示，算法时间复杂度明显高于本文方法。综上，在处理不含无义消息的群聊数据时，SP_{TSAI}算法性能更优。另外，本文方法处理无义消息实验结果如表3所示。

表3　在保留无义消息的数据集上的实验结果

Table 3　Experimental Results on Data Sets with Nonsense Messages

算法	指标	D_1	D_2	D_3
SP_{TSAI}	T_t	0.65	0.75	0.20
	T_f	0.15	0.15	0.70
	H_t	2	4	4
	F	0.859	0.801	0.881

表3中三个数据集的平均F值比表2中的分别高出约29.4%、27.1%和22.5%，说明SP_{TSAI}算法在处理完整数据集时优势更大，这是因为使用完整数据集能够避免连续而大量的无义消息将一个话题分割为多个，并能利用无义消息的时间、用户属性，提高其邻近消息的聚类效果。

以表3中D_1的数据为例，分析三个阈值T_t、T_f和H_t。在H_t=2时，在不同的T_t下对T_f进行遍历，其中T_t分别取0.25、0.45、0.65和0.85，T_f则在0.10~0.80之间以0.05的步长进行变化，结果如图3所示。可以看出，当T_f取值较大时，T_t取较小值才能获得较高的平均F值，这是因为在话题合并条件较为严苛的情况下，只有增大消息聚类所得话题的粒度才能得到更接近真实情况的话题检测结果；同样地，当T_f取值较小时，T_t取较大值时平均F值较高，这是因为消息聚类使用较高相似度阈值会导致类簇过小，只有降低话题合并要求才能有效减少一个话题被分为多个话题的情况出现。

分析话题热度检测时间H_t，在T_t取0.65，T_f取0.15的情况下，H_t在1~20分钟范围内以步长为1分钟进行遍历，结果如图4所示。可以看出，当H_t向小于或大于2分钟的方向取值时，平均F值都呈下降趋势，这表明对数据集D_1来说，话题状态更新的最合适周期应为2分钟。

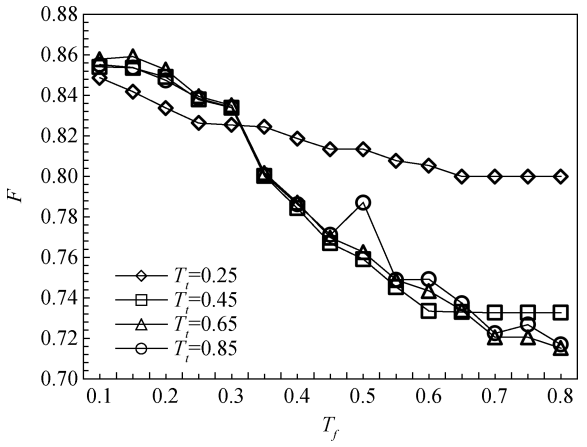

图 3　在不同的 T_t 下对 T_f 进行遍历

Fig.3　Traverse T_f with Different T_t

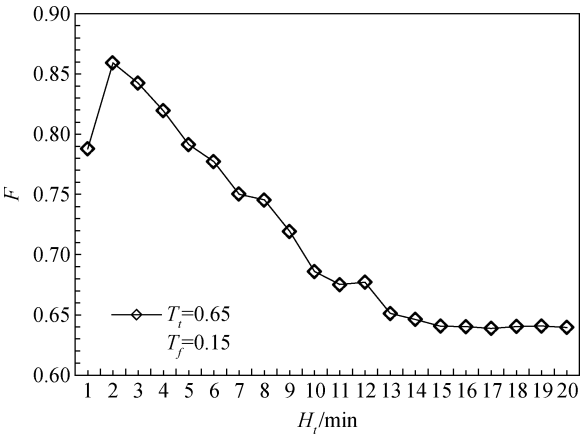

图 4　遍历 H_t

Fig.4　Traverse H_t

6　结　语

本文研究了群聊消息的内容、时间、用户和类型在话题检测中的作用，并针对群聊话题交叉并行的问题设计了话题序列，提出一种基于多策略的群聊话题检测技术。在三个来源于真实群聊记录的数据集上的实验结果表明，本文方法相较于传统方法在处理过滤了无义消息的数据时具备更好的性能，而且在处理包含无义消息的完整数据集时能取得更好的效果。

本文方法还存在一些不足，比如未能充分挖掘消息的深层语义特征，所设阈值较多，下一步需要针对这些方面展开工作。另外，通过对群聊记录的研究发现，群聊人数越多，用户间的亲密度就越能影响用户对某些话题的参与度，在未来的工作中，会将这一规律应用到话题检测中。

参考文献：

[1] 中国互联网络信息中心 . 第 44 次中国互联网络发展状况统计报告 [R/OL]. (2019-08-30). [2020-04-10]. https://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/t20190830_70800.htm. (China Internet Network Information Center. The 44th China Statistical Report on Internet Development[R/OL]. (2019-08-30). [2020-04-10]. https://www. cnnic. net. cn/hlwfzyj/hlwxzbg/hlwtjbg/201908/t20190830_70800.htm.)

[2] Uthus D C, Aha D W. Multiparticipant Chat Analysis: A Survey [J]. Artificial Intelligence, 2013, 199-200: 106-121.

[3] Onan A, Korukoğlu S, Bulut H. Ensemble of Keyword Extraction Methods and Classifiers in Text Classification[J]. Expert Systems with Applications: An International Journal, 2016, 57(C): 232-247.

[4] Xie F, Wu X D , Zhu X Q. Efficient Sequential Pattern Mining with Wildcards for Keyphrase Extraction[J]. Knowledge Based Systems, 2017, 115: 27-39.

[5] Kang Y B, Haghigh P D, Burstein F. TaxoFinder: A Graph-based Approach for Taxonomy Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(2): 524-536.

[6] Sanchez-Pi N, Martí L, Garcia A C B. Improving Ontology-based Text Classification: An Occupational Health and Security Application[J]. Journal of Applied Logic, 2016, 17(C): 48-58.

[7] Saleh A I, Al Rahmawy M F, Abulwafa A E. A Semantic Based Web Page Classification Strategy Using Multi-layered Domain Ontology[J]. World Wide Web, 2017, 20(5): 939-993.

[8] Wu D Z , Zhu H, Li G L , et al. An Efficient Wikipedia Semantic Matching Approach to Text Document Classification[J]. Information Sciences: An International Journal, 2017, 393(C): 15-28.

[9] Agathangelou P, Katakis I, Koutoulakis I, et al. Learning Patterns for Discovering Domain-oriented Opinion Words[J]. Knowledge and Information Systems, 2018, 55(1): 45-77.

[10] Bandhakavi A, Wiratunga N, Padmanabhan D, et al. Lexicon Based Feature Extraction for Emotion Text Classification[J]. Pattern Recognition Letters, 2017, 93: 133-142.

[11] Manek A S, Shenoy P D, Mohan M C, et al. Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews Using Gini Index Feature Selection Method and SVM Classifier[J]. World Wide Web, 2017, 20(2): 135-154.

[12] Chaturvedi I, Ong Y S, Tsang I W, et al. Learning Word Dependencies in Text by Means of a Deep Recurrent Belief Network[J]. Knowledge-Based Systems, 2016, 108(C): 144-154.

[13] Tommasel A, Godoy D. Short-text Feature Construction and

Selection in Social Media Data: A Survey[J]. Artificial Intelligence Review, 2018, 49(3): 301-338.

[14] Pavlinek M, Podgorelec V. Text Classification Method Based on Self-training and LDA Topic Models[J]. Expert Systems with Applications: An International Journal, 2017, 80(C): 83-93.

[15] Qin Z C, Cong Y H, Wan T. Topic Modeling of Chinese Language Beyond a Bag-of-words[J]. Computer Speech and Language, 2016, 40(C): 60-78.

[16] Zhang H, Zhong G Q . Improving Short Text Classification by Learning Vector Representations of Both Words and Hidden Topics[J]. Knowledge-Based Systems, 2016, 102(C): 76-86.

[17] Zuo Y, Zhao J C, Xu K. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts[J]. Knowledge and Information Systems, 2016, 48(2): 379-398.

[18] Pinheiro R H W, Cavalcanti G D C, Tsang I R. Combining Dissimilarity Spaces for Text Categorization[J]. Information Sciences: An International Journal, 2017, 406-407: 87-101.

[19] Elnahrawy E. Log-based Chat Room Monitoring Using Text Categorization: A Comparative Study[C]//Proceedings of the 2002 International Conference on Information and Knowledge Sharing. 2002.

[20] Özyurt Ö, Köse C. Chat Mining: Automatically Determination of Chat Conversations' Topic in Turkish Text Based Chat Mediums [J]. Expert Systems with Applications, 2010, 37(12): 8705-8710.

[21] Adams P H, Martell C H. Topic Detection and Extraction in Chat [C]//Proceedings of the 6th International Conference on Semantic Computing, 2008: 581-588.

[22] Wang L D, Oard D W. Context-based Message Expansion for Disentanglement of Interleaved Text Conversations[C]//Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009: 200-208.

[23] 李天彩, 王波, 席耀一 . 基于多策略的短文本信息流会话抽取 [J]. 计算机应用研究, 2016, 33(4): 997-1002.(Li Tiancai, Wang Bo, Xi Yaoyi. Conversation Extraction in Short Text Message Streams Based on Multiple Strategies[J]. Application Research of Computers, 2016, 33(4): 997-1002.)

[24] 黄九鸣, 吴泉源, 刘春阳, 等 . 短文本信息流的无监督会话抽取技术 [J]. 软件学报, 2012, 23(4): 735-747.(Huang Jiuming, Wu Quanyuan, Liu Chunyang, et al. Unsupervised Conversation Extraction in Short Text Message Streams[J]. Journal of Software, 2012, 23(4): 735-747.)

[25] Ding Y X, Meng X J, Chai G R, et al. User Identification for Instant Messages[C]//Proceedings of the 18th International Conference on Neural Information Processing. 2011: 11-13.

[26] Köse C, Özyurt Ö, İkibaş C. A Comparison of Textual Data Mining Methods for Sex Identification in Chat Conversations[C]// Proceedings of the 4th Asia Conference on Information Retrieval Technology. 2008: 638-643.

[27] Shen D, Yang Q, Sun J T , et al. Thread Detection in Dynamic Text Message Streams[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 35-42.

作者贡献声明：

吴旭：提出研究思路，论文最终版本修订；

陈春旭：设计研究方案，采集、清洗和分析数据，进行实验，论文起草。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail: sinkingsand@qq.com。

[1] 陈春旭 . GroupChatAnalysis.zip. 实验所用 Python 代码 .

[2] 陈春旭 . wechat.xlsx. 微信群聊记录 .

[3] 陈春旭 . Word2Vec.zip. 实验所用 Word2Vec 模型 .

收稿日期：2020-07-22

收修改稿日期：2020-11-21

Detecting Topics of Group Chats with Multiple Strategies

Wu Xu[1, 2, 3]　Chen Chunxu[1, 2]

[1](Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China)

[2](School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

[3](Beijing University of Posts and Telecommunications Library, Beijing 100876, China)

Abstract: [**Objective**] This paper tries to detect topics of continuous group chats with variou types of message, aiming to address the topic entanglement issue of group chats, and reduce the influence of sparse text features on clustering. [**Methods**] We proposed a detection model for group chat topics based on multi-strategies. This model solves topic crossover issue with topic sequences, and improves clustering results with data on users, time, and types of messages. [**Results**] We examined our model with plain texts of three group chats. The new method's F value was 2.9%, 6.1% and 3.0% higher than those of the existing algorithms. The speed of our model is about 27.6%, 32.1% and 47.1% faster. This method also processed mixed types of data that cannot be handled by traditional algorithms, and the speed was improved by about 29.4%, 27.1%, and 22.5% respectively. [**Limitations**] We do not fully utilize the text features of group chat message and set too many thresholds for the algorithm. [**Conclusions**] The proposed method could identify group chat topics, and improve the efficiency of public opinion analysis.

Keywords: Group Chat Message　Topic Detection　Short Text

利用兄弟姐妹的信息捕捉人脸3D形状

捕捉人脸3D形状，以及具有不同遗传学的个体之间的3D形状变化的能力可以为各种应用提供信息，包括研究人类进化、制定手术计划以及法医科学。但是，现有的将遗传学与身体特征联系起来的工具需要输入简单的测量值，例如眼睛之间的距离，而这些测量值无法充分捕捉面部形状的复杂性。

来自比利时的研究人员在开放存取期刊*PLOS Genetics*中介绍了一项新发现，他们开发出一种能捕获人脸3D形状的新策略，可利用兄弟姐妹的信息以识别出面部形状特征与人类基因组中特定位置之间的新联系。研究人员最初是通过从273对欧洲血统的兄弟姐妹的3D面部数据进行学习，发现了1 048个在兄弟姐妹之间共享的面部特征，因此可以假设这是具有遗传学基础的，然后研究人员开发了该策略。进一步，研究人员将他们的新方法用于捕捉面部形状，并将其应用于8 246名具有欧洲血统的人，产生了关于兄弟姐妹之间面部形状相似性的数据，结合这些人的遗传信息，并使用现有的工具进行分析，以将遗传学特征与身体特征进行关联。

这项研究可以作为未来其他研究的基础，例如在更大的人群中去复制本研究，以便更好地了解面部发育所涉及的生物学过程。研究人员还补充说："由于存在血缘关系，兄弟姐妹很可能会共享一些面部特征，因此可以从表型相似的兄弟姐妹对中提取与生物学相关的性状。"

(编译自：https://www.sciencedaily.com/releases/2021/05/210513142511.htm)

（本刊讯）