基于候选主题词与话题分类的人物行为研究

刘晓芳[1],欧荣安[2],罗　欢[3],刘芳婷[4],张辉极[1],韩　冰[2],赵建强[1,5*]
(1. 厦门市美亚柏科信息股份有限公司,福建 厦门 361008;
2. 广州市刑事科学技术研究所,广东 广州 510030;
3. 福州大学,福建 福州 350108;
4. 厦门市人民检察院检察技术信息部,福建 厦门 361008;
5. 西安电子科技大学,陕西 西安 710071)

摘　要:如何从海量聊天数据获取聊天主题和聊天人物行为是案件智能化分析的热点问题之一。传统词嵌入方法,将文本中的所有词汇映射到向量空间,存在词汇特征冗余的问题。为了缓解这一问题,该文提出一种基于候选主题词的话题分类算法—CTW (candidate topic words)。该算法使用 LDA 主题模型抽取聊天文本中的关键词,使用预训练词向量得到显著的语义特征,同时为增强特征,将字符特征与获取的词汇特征进行融合。传统方法同时还存在只关注话题无法更精确地刻画人物行为的问题。针对该问题,该文提出了同时获取聊天话题和人物行为的方案:针对已归类的话题,该方案使用群成员互动强度、群成员活跃度作为人物行为网络中的权值,构建话题参与人的行为网络图;最后通过成员在群中的备注给人物赋予不同的社会标签,以丰富人物行为。实验表明,提出的话题分类算法,在实际搜集的数据集上比基线模型拥有更佳的性能,在获取群聊话题的同时得到了更丰富的人物行为描述。

关键词:聊天主题;候选主题词;话题分类;人物行为;互动强度;群成员活跃度;社会标签
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2022)04-0044-07
doi:10.3969/j.issn.1673-629X.2022.04.008

Research on Character Behavior Based on Candidate Keywords and Topics Classification

LIU Xiao-fang[1], OU Rong-an[2], LUO Huan[3], LIU Fang-ting[4], ZHANG Hui-ji[1],
HAN Bing[2], ZHAO Jian-qiang[1,5*]
(1. Xiamen Meiya Pico Information Co., Ltd., Xiamen 361008, China;
2. Guangzhou Institute of Criminal Science and Technology, Guangzhou 510030, China;
3. Fuzhou University, Fuzhou 350108, China;
4. Procuratorial Technology Department of the People's Procuratorate of Xiamen, Xiamen 361008, China;
5. Xidian University, Xi'an 710071, China)

Abstract: For the high incidence of new types of cybercrime, criminals use various chat groups to induce victims and commit crimes. The effective description and portrayal of the chat content and the character relationships will provide strong support for the case investigation. Generally traditional methods only focus on one aspect of chat topics or character behavior, which is unable to extract behaviors and chat topics at the same time quickly and effectively in forensics. Therefore, we propose a method- CTW (candidate topic words) to portray character relationships based on candidate keywords and group chat topics. Firstly, LDA model is used to extract key words, using the embedding of char and extracted key words for TextCNN to categorize topics. For categorized topics, we treat interaction intensity and group member activity as the weight value on the directed edges, not only create a network diagram of the character behavior of the participants in the group chat and get the members' remarks in the group as social relationship labels which are put on the edges of the relationship diagrams. Experiments show that the method proposed shows good results on actual data sets, which can effectively portray character relationships based on the topics of group chats.

收稿日期:2021-03-16　　**修回日期**:2021-07-20
基金项目:广州科技攻关重大专项(201903007);国家自然科学基金青年基金项目(G61801121)
作者简介:刘晓芳(1991-),女,工程师,研究方向为自然语言处理;通信作者:赵建强,讲师,博士,CCF 会员(12330M),研究方向为人工智能、自然语言处理、电子数据取证。

Key words: group chat topic; candidate keywords; topic classification; personalities relationship; interaction intensity; group member activity; social label

0 引　言

近年来移动互联网发展迅猛,截止 2020 年 12 月,中国网民的规模已达到 9.89 亿,其中手机网民占比为 99.7%,即时通信类 APP 占手机网民的 99.2%[1]。以 QQ、微信为代表的通讯软件已成为多数人日常工作和生活中不可或缺的一部分,其群组聊天与微博、论坛等瀑布型弱关系为主的社交网络不同,以交友和表达为主,展现出多对多的复杂社交关系和情感依赖特征。在电子数据取证分析领域,聊天文本中的关键话题和人物行为的挖掘能为案件侦破带来关键进展。

目前,基于词汇特征的话题分类方法已经具备一定的特征捕捉和语义理解能力,但仍存在特征冗余问题,因此需要选择性地关注话题中的关键词汇,同时使用丰富的语义特征提升话题分类的性能。可是,单纯得到聊天话题,无法多方位地刻画人物,无法对案件快速有效地定位。为解决该问题,该文在分析聊天话题的同时从多个维度对人物行为进行刻画,为案件侦办提供更有力的支撑。

1 相关研究现状

会话分析理论(Conversation Analysis Theory)[2]起源于 20 世纪 60 年代美国的一种社会学研究理论:基于真实会话揭示人的社会行为和互动交际行为的内在组织结构,进而发现人类言语交际模式和规律[3]。文献[3-5]对该理论进行进一步分析,将会话结构划分为整体结构与局部结构两部分。整体结构指一个完整会话活动在其展开过程中依照交际要求所形成的功能模式,包括会话的开始、维持及结束整个过程;局部结构指话题参与者的局部发言,包括发言者联系、话轮交替及构成连贯话语的方式等。文献[6-8]基于文本语义统计信息和聚类算法,实现对微博等社交网络话题的挖掘。话题分类是文本分类的重要应用,目前,基于神经网络的文本分类方法是文本分类的主流方向。Kim 最早提出使用 TextCNN 对文本进行分类,利用不同尺度的卷积核很好地捕捉到文本的局部关键信息[9]。Kalchbrenner 提出 DCNN 模型,使用宽卷积和 K-max 池化采样解决了 TextCNN 无法捕获长距离信息的问题,取得了不错的分类效果[10]。

因 CNN 快速的计算能力和可并行性,不仅被广泛应用于工业界,同时也吸引了研究者对其网络结构的不断改进,通过使用注意力机制[11]、增加网络深度[12]、使用层级网络[13]等实现对 CNN 分类效果的优化。另外,也有研究者通过不同的嵌入方式,取得了不错的分类效果。目前主流的文本嵌入方法有词嵌入[14]、字符嵌入[15]、句子嵌入[16]及不同嵌入方式的组合。词嵌入虽然可以捕获到文本的句法和语义特征,但是无法处理未登录词,存在 OOV(Out-of Vocabulary)问题,字嵌入可以解决这一问题。同时,为丰富词汇和字符的表征能力,研究者提出不同的预训练模型:Word2vec[17]、GloVe[18]、CoVe[19]和 ELMo[20]、ULMFiT[21]、OpenAI-GPT[22]和 Bert[23],以获取文本中词和字更丰富的表征,提高分类性能。

基于预训练模型的词嵌入方式虽然能够获取很好的文本分类效果,但是每个词汇对文本分类的贡献度并不等价。因此,该文使用 LDA 主题模型,在获取关键话题的同时提取到影响话题的关键词,并通过训练好的词向量获得关键词的语义表示,同时增加字符特征,以提高话题的分类性能。为多方位、多维度地刻画人员行为,除了获取聊天话题,还应该包含该人员在聊天中的互动强度、活跃程度及其他附属标签等描述信息。

群组的各个成员在聊天中扮演着不同的角色,因此每个对话并不是简单的文字输入,而是一个成员在网络空间行为的表达[24]。近年来,人物行为的定量分析领域取得了一定进展,文献[25]使用聚类分析和关联分析的方法获取人物行为。文献[26]提出基于同义词词林的关系抽取技术,扩展了传统二元关系,并实现了人际网络的定量分析及结果的可视化,在一定程度上促进了人物行为刻画技术的发展。文献[27]研究微信群内的会话数据,通过构造成员活跃度和基于成员相互回应行为与共同回应行为设计成员关联强度算法,进而构建会话交流网络来分析群聊天的静态结构特征及动态结构演化过程。文献[28]以聊天时间为对象,利用聚类分析技术发现通信时间与人物行为之间的关联关系,将人物行为按照“亲疏远近”分类。

该文提出一种基于候选主题词的话题分类算法-CTW 和人物行为刻画方案,主要创新为:

(1)CTW 不仅充分考虑聊天文本中字符级别信息,同时关注影响话题的关键词信息,缓解了词汇特征的冗余问题。

(2)从话题类别、成员互动强度、活跃程度和社会标签等多个维度对人物行为进行刻画,丰富了人物行为的描述。

(3)还考虑到了聊天中成员的社会属性标签,通过构建的人物行为库,建立人物行为拓扑图。进而帮助取证人员快速形成嫌疑人员的画像,提高取证效率。

2 人物行为刻画

2.1 LDA 主题模型

LDA[29]模型是一种经典主题模型算法,可以计算出文档集中每个文档的主题概率分布。具体地,一个语料库由 M 篇文档的文档集合 $D = < d_1, d_2, \cdots, d_M >$,N 个词的词汇表 $W = < w_1, w_2, \cdots, w_N >$ 和 K 个主题 $Z = < z_1, z_2, \cdots, z_K >$ 组成。文档 d 的主题分布 θ_d 满足狄利克雷分布 $\mathrm{Dir}(\alpha)$,其中 θ_d 是通过采样得到的多项式分布。词汇表中的词语定义为 φ_k ,与主题有关且同样满足另一个狄利克雷分布 $\mathrm{Dir}(\beta)$ 。因此,每个词 w_i 由 θ_d 分布的主题 z_i 和 φ_k 分布的文档所决定。因此单词、主题、文档三者的联合分布由主题和文档的联合分布以及单词和主题的联合分布共同决定。

$$\theta_d \sim \mathrm{Dir}(\alpha), \varphi_k \sim \mathrm{Dir}(\beta), z_i \sim \theta_d, w_i \sim \varphi_{z_i} \quad (1)$$

$$P(\theta, z, w \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \beta) \quad (2)$$

根据已知的文档和单词分布,给出两个超参数 α 和 β 来求解该模型,得到文档主题分布 θ 和单词主题分布 φ 。

2.2 CTW 模型设计

2.2.1 数据预处理

聊天数据中存在大量的噪声,会影响模型分类效果。为解决该问题,该文对数据进行预处理操作,具体预处理流程包括:

(1)数据清洗:去除各类系统通知消息、各类表情符号、各类型网页、图片、视频连接及其他乱码和非中文字符。

(2)去停用词:构建停用词表,去除文本中多次出现但对文本语义价值不大的词。

(3)分词:采用 jieba 对数据进行分词。

(4)分段:因聊天话题不时的变化,需对聊天文本进行分段处理,以确保话题的唯一性。通过对聊天时间、文本长度、句子数、聊天话题的分析,设计了如图 1 的分段策略。

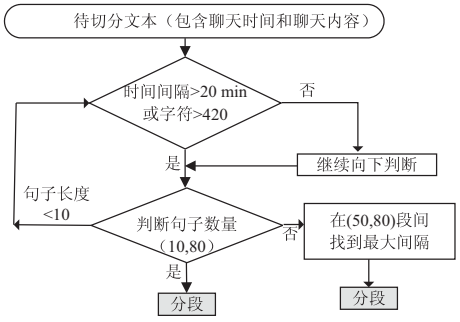

图 1 聊天数据分段策略

针对每个聊天文本,按时间间隔和句子数量分段。如果上下句间间隔大于 20 分钟,并且句子数量在 10 到 80 句之间或字符长度大于 420,则分段,字符长度大于 420 的长文本一般为广告、通知等消息;如果上下句间隔大于 20 分钟,但是句子数量小于 10(通过分析得到小于 10 句的聊天文本不包含关键话题),则继续向下寻找时间间隔大于 20 分钟的句子;如果句子数量大于 80,则从 50 至 80 句之间寻找一个最大的时间间隔分段,并依次向下按该规则分段。该文以预处理后的子片段为单位,作为 CTW 模型的输入。提出的基于候选主题词的话题分类算法分类的模型结构如图 2 所示,CTW 的模型主要包括特征输入层、嵌入层、卷积层、K-max 池化层、全连接层和输出层。

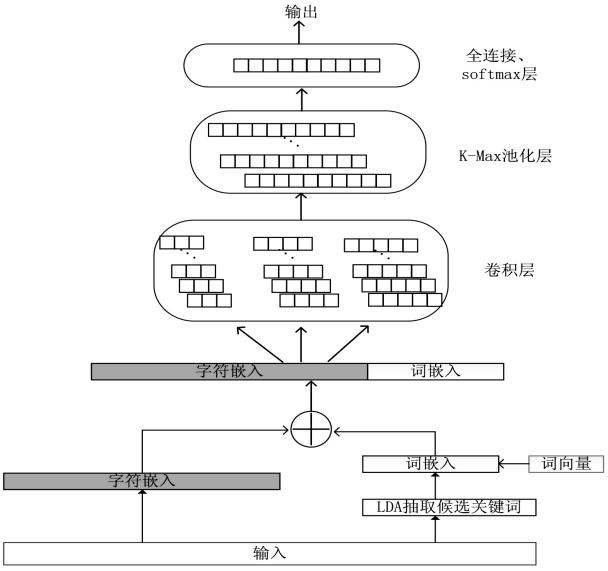

图 2 CTW 网络模型结构

2.2.2 嵌入层

嵌入层包含关键词嵌入和字符嵌入两个模块。首先,使用 LDA 主题模型获取聊天话题中的关键词,假定一个分段后的子片段仅包含一个主要关键主题。因此,使用 LDA 主题模型获取每个片段的关键话题,同时得到主话题前 k 个主题词 $x_{\mathrm{Top-k-words}} = \{x_1, x_2, \cdots, x_k\}$;然后,使用腾讯训练好的词向量,得到 k 个主题词的词嵌入矩阵 $k \times n$, n 为每个词语对应的词向量的维度;其次,针对输入文本 $x_{\mathrm{char}} = \{x_{\mathrm{char}-1}, x_{\mathrm{char}-2}, \cdots, x_{\mathrm{char}-m}\}$,得到维度为 $m \times n$ 的字符嵌入矩阵, m 为一个段落中的字符数, n 为每个字符对应的字向量的维度, $x_{\mathrm{char}-i} \in \mathbb{R}^n$ 表示段落中第 i 个字符的 n 维字嵌入。最后,拼接得到整个 CTW 网络的输入:

$$\boldsymbol{x} = \{x_{\mathrm{char}}, x_{\mathrm{Top-k-words}}\} \quad (3)$$

其中, $\boldsymbol{x}$ 为 $(m + k) \times n$ 维矩阵。

2.2.3 卷积层

在嵌入层矩阵上使用卷积操作得到输入文本的特征:

$$c_i = f(w \bullet x_{i:i+h-1} + b) \quad (4)$$

其中, $x_{i:i+h-1}$ 表示输入矩阵第 i 到第 $i + h - 1$ 行所组成

的 $h \times n$ 的窗口，h 表示所包含的字符或词语个数，$\boldsymbol{w}$ 为权重矩阵，b 为偏置参数，f 为非线性函数。通过定义不同尺度卷积核，提取不同尺度的特征向量，得到聊天内容的上下文语义特征。

$$c_i = c_1 \oplus c_2 \oplus \cdots \oplus c_{i+h-1} \tag{5}$$

2.2.4　K-max 池化层

将语义特征 c_i 输入到 K-max 池化层，保留强化特征：

$$o = \mathrm{kmax}\{c_i\} \tag{6}$$

最后对全连接层的输出，采用 SoftMax 对输出特征分类得到输入片段的主题类别。

2.3　人物行为

2.3.1　群成员互动强度

在群聊天过程中，群成员间存在某些互动模式来支撑话题的起承转合，这些互动模式中隐藏着群成员间的关系信息。因此，本算法考虑引入成员的互动强度指标来刻画群成员的关系强弱。互动强度分为两种：显性互动强度和隐性互动强度[30]。显性互动强度为成员通过点对点@功能建立的强关联关系；隐性互动强度为成员围绕某一话题展开多对多交流时产生的弱关联关系。该文认为在即时通讯软件的群聊天中显性互动强度对于成员关系刻画的贡献度高于隐性互动强度。在显性互动强度中，成员间相互@的互动行为是对成员间关系的最直观体现。并且，由于@行为具有指向性，因此显性互动强度指标具有不对称性，如公式(7)所示。

$$\mathrm{Ove\_intimacy}_{ij} = \mathrm{At\_Num}_{ij} \tag{7}$$

其中，$\mathrm{At\_Num}_{ij}$ 表示在某个话题内成员 i 对成员 j 发出@行为的次数。隐性互动强度考虑成员在每一个话题片段内的共现关系，因此针对每个话题片段借助科学计量学中的合作强度 Salton 指标来计算基于话题片段的隐性互动强度。计算公式如下：

$$\mathrm{Imp\_intimacy}_{ij} = \frac{\mathrm{T\_occur}_{ij}}{\sqrt{T_i * T_j}} \tag{8}$$

其中，$\mathrm{T\_occur}_{ij}$ 表示成员 i 和成员 j 的互动频次，即两个成员共同参与话题讨论的次数，文中表现为成员 i 和成员 j 在同一个话题片段中共同出现的次数，T_i 和 T_j 分别表示两位成员各自参与的话题总数量。Salton 指数越高，表示两者间的隐性互动强度越强。

为平衡两种互动强度的贡献程度，引入权重控制因子 λ，统一后的互动强度公式为：

$$\mathrm{Inti}_{ij} = \lambda\ \mathrm{Ove\_intimacy}_{ij} + \theta\ \mathrm{Imp\_intimacy}_{ij} \tag{9}$$

其中，λ 和 θ 表示控制两种互动强度的权重因子，$\lambda + \theta = 1$。该文考虑显性和隐性互动强度的重要程度，将 λ 设置为 0.7，θ 设置为 0.3。

2.3.2　群成员活跃程度

在群聊天的会话过程中，不同群成员在群中的参与地位存在很大差异。少数处于中枢位置的成员在群会话中扮演重要角色，这种成员发言量大，常作为话题片段的发起者和积极参与者。另外有部分成员则恰恰相反，对群会话交流更多持围观态度而非积极参与。此外，还存在部分成员只和固定几个成员互动的情况。因此，为了刻画群聊天内成员的人物行为，有必要先对群成员的活跃程度进行分析。该文认为发言数量是最直观反映群成员活跃程度的指标，而发言天数则反映了成员对该群的黏度，因此综合考虑发言数量和发言天数，群成员活跃程度计算公式如(10)所示[31]：

$$\mathrm{Coreness}_i = \alpha \frac{\mathrm{Mes\_Mum}_i}{\mathrm{Total\_Mes\_Mum}_i} + \beta \frac{\mathrm{Mes\_Day}_i}{\mathrm{Total\_Mes\_Day}_i} \tag{10}$$

其中，α 和 β 表示发言数量和黏度的影响因子；$\mathrm{Mes\_Mum}_i$ 表示成员 i 在某个话题内的发言数量；$\mathrm{Total\_Mes\_Mum}_i$ 表示该话题内所有成员的总发言数量；$\mathrm{Mes\_Day}_i$ 表示成员 i 在某话题内的发言天数；$\mathrm{Total\_Mes\_Day}_i$ 表示该话题的总发言天数。该文假定发言量和发言天数对描述成员互动强度具有同等重要性，因此设定 $\alpha = \beta = 0.5$。针对全体成员的 Coreness 指标，排序后计算相邻数值间的差值，取差值前三的节点作为活跃程度的分隔点，将全体成员分为：核心成员、活跃成员、围观成员和疏离成员四种类型[32]。

2.3.3　群成员社会标签

对于每个话题片段，成员间的人物行为刻画分为：群成员间互动强度，群成员活跃程度和群成员间社会关系标签。社会关系标签考虑聊天中发言人昵称和成员的备注。该文发现通过群昵称和设置的备注信息可以捕获部分成员的社会关系。例如：若出现“爸”、“老婆”等备注信息则能够发现家庭关系的标签；出现“教练”、“学员”等备注时则可能是健身交流。

2.4　人物行为库

考虑到聊天中的人物行为是基于话题构建的，因此将聊天建模成(群聊-话题-成员)的层次模型。整个聊天被按照分段策略分割若干个段落，并对聊天话题类别归类，每个片段中由若干个参与话题讨论的群成员组成。图 3 展示了基于层次模型建模的人物行为库。

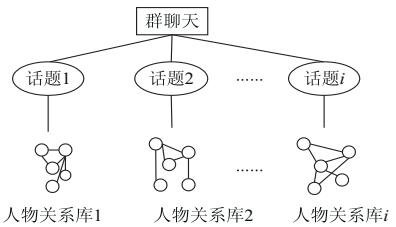

图 3　人物行为库

通过上述的人物行为刻画规则,构建了基于话题片段的人物行为图库,每个成员代表图库中的一个节点,图库的边可添加成员间的互动强度、活跃程度和社会关系标签的属性。增加了社会关系标签的基于话题的成员人物行为图库如图 4 所示。

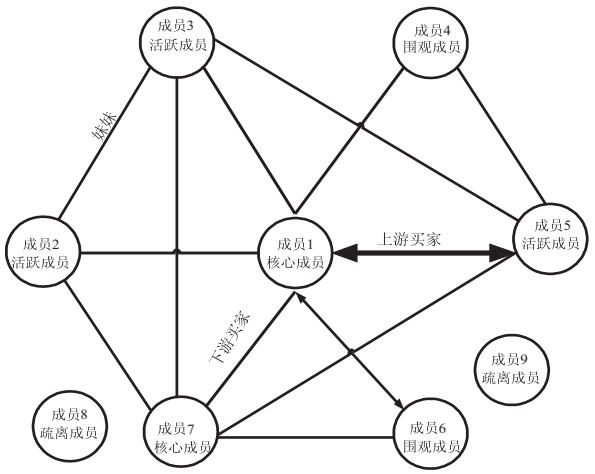

图 4 话题 i 的成员人物行为图库

3 实验设计

实验所用操作系统 Ubuntu18. 0, CPU Inter® Xeon® Sliver 4116 CPUC 2. 1 GHz GPU 为 RTX8000, 显存 48G, 开发环境为 python 3. 7. 4 和 tensorflow 1. 14. 0。模型参数初始化设置为:文本最大长度为 500;候选词个数是 10;字嵌入维度为 200;词嵌入维度为 200;学习率为 0. 01;正则化学习率为 0. 1;Dropout 为 0. 5;Batch_size 为 512;词向量总数为 100 000,字符总数是 5 000。针对不同数目的候选关键词进行实验,选取关键词个数分别为 5,10,15,30,综合考虑精度、召回率、F1 值发现,关键词数量为 10 时,主题分类效果最好。

3.1 实验数据说明

从微信、QQ 等社交 APP 上收集不同种类的聊天文本数据,通过分段、人工标注最终形成 21 473 段中文社交短文本样本,共包含 8 个话题类别。采用分层抽样将数据划分为训练集 17 167 段和测试集 4 306 段,各个类别数据样本分布如表 1 所示。

表 1 数据集样本分布

类号	类别	训练集	测试集
1	游戏	1 915	456
2	拼车	935	230
3	物业	1 789	464
4	美食	1 770	238
5	学校	974	236
6	健身	886	238
7	宠物	920	238
8	其他	7 978	2 309

3.2 话题关键词提取

使用 GibbsLDA++软件包对聊天文本中的话题关键词进行抽取。给出 2 个文本聊天对话中前 4 个主要关键主题词的抽取结果,文本 1:王者、荣耀、钻石、排位;文本 2:教练、训练、瘦身、出汗。因此,LDA 能够得到某些聊天话题的主题词集合。从文本 1 的关键词看,该文本属于游戏主题,而文本 2 属于健身主题,与人工标注结果一致。

3.3 CTW 模型话题分类结果分析

使用 CTW 和基线模型 TextCNN 做对比实验,使用精度、召回率和 F1 值(文本分类通用指标)作为该文的评价指标,表 2 展示了 7 类对取证有重要帮助话题的分类结果。

表 2 聊天话题分类结果

类别	Precision		Recall		F1	
	TextCNN	CTW	TextCNN	CTW	TextCNN	CTW
游戏	0. 92	0. 94	0. 38	0. 48	0. 54	0. 64
拼车	1	1	0. 98	0. 99	0. 99	0. 99
物业	0. 98	0. 99	0. 83	0. 85	0. 9	0. 92
美食	0. 81	0. 79	0. 42	0. 63	0. 56	0. 7
学校	0. 96	0. 92	0. 81	0. 87	0. 88	0. 89
健身	1	0. 95	0. 12	0. 57	0. 21	0. 71
宠物	0. 92	0. 93	0. 45	0. 63	0. 6	0. 75

从表 2 中可以看出,与基线模型相比,除美食、学校、健身三类的精度低于 TextCNN 外,其他话题分类精度均高于 TextCNN 模型;召回率均高于基线模型;提出的 CTW 算法的 F1 值均高于 TextCNN 模型。结果表明,增加话题关键词作为特征,可以增强文本的语义特征,得到更丰富的语义表征,同时兼具字符特征可以有效捕获文本不同层次的特征表示,提升话题识别的性能。

3.4 人物行为结果分析

3.4.1 群成员互动强度

亲密程度的判定基于不同话题中成员间的互动强度,文本中采用启发式的固定阈值进行判定,将亲密度程度判定为:亲密、一般亲密、不亲密三种程度,不同话题设置不同的阈值,在一定程度上反映了同一话题中不同成员间互动强度的相对亲密程度。表 3 展示了不同话题中部分成员间的互动强度和亲密程度。

(1)在不同的话题中,因互动强度的计算方式导致互动强度的绝对数值范围往往不一致,占据聊天更多篇幅的话题会使互动强度的数值偏大,因此在亲密度判定时要根据不同的话题设置不同阈值。

(2)成员在不同的话题中亲密程度存在差异,这也证明提出的针对不同的话题刻画成员行为是合理的。

表 3　不同聊天话题中部分成员间的互动强度和亲密程度

所属话题	账号 1	账号 2	互动强度	亲密程度
话题 1	新 A40	创 20	90.859	亲密
话题 1	阿敢	新 A40	81.584	亲密
话题 1	阿敢	创 20	77.424	一般亲密
…	…	…	…	…
话题 1	宇宙第一	创 20	40.556	不亲密
话题 1	粉红小猪	创 6	40.488	不亲密
…	…	…	…	…
话题 2	云数	A 豪 * 杨总	15.242	不亲密
话题 3	新 A40	创 20	121.750	亲密
话题 3	粉红小猪	8 落花 111	58.731	不亲密

3.4.2　群成员活跃程度

从搜集的数据中选取美食、物业、宠物三个话题,统计其在 0 ~ 24 时的活跃度,图 5 展示了不同时间下发言数量的分布情况。其中横轴表示发言时间,纵轴表示发言数量。从图 5 中可以发现:

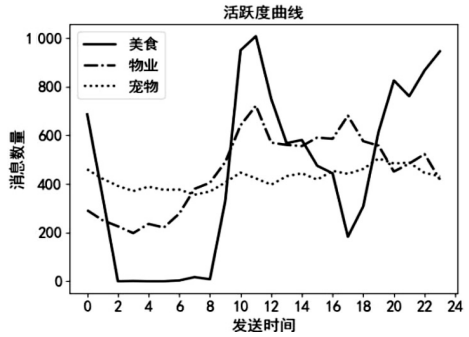

图 5　群聊天发言数量分布

(1)不同群聊天的活跃程度存在较大差别,活跃度曲线也反映出群聊天的性质。如物业在办公时间段内活跃度较高;宠物群在每个时间段均保持较高的活跃度;美食群则在中午、晚上及夜宵活跃度较高。

(2)群聊天的活跃程度是分析群成员活跃程度的一个重要因素,群聊天的活跃程度反映了群成员在该时间段内的活跃程度总和。

基于 2.3.2 中的描述,表 4 展示了不同话题中成员活跃程度的数值及成员类型判定结果。

表 4　不同话题中成员类型

所属话题	账号	活跃程度	成员类型
话题 1	创 14	0.446	核心成员
话题 1	创 20	0.283	活跃成员
…	…	…	…
话题 1	表秀越	0.102	围观成员
话题 1	风轻云淡	0.042	疏离成员
话题 2	创 14	0.415	核心成员
…	…	…	…

续表 4

所属话题	账号	活跃程度	成员类型
话题 2	奔跑	0.042	疏离成员
话题 3	创 14	0.414	核心成员
话题 3	此号已停	0.278	活跃成员
话题 3	爱恨情仇	0.042	疏离成员

从表 4 中可以发现:

(1)同一群成员在不同的话题中参与程度不同,成员更偏向于参与自己感兴趣的话题。

(2)部分群中会出现某些成员在所有话题中均非常活跃,属于核心成员,作为话题的发起者和积极参与者,在群中具有很强的号召力和影响力,在取证过程中应重点关注。

(3)在多成员构成的群聊中,疏离成员占据群聊天的大多数,许多人在群里都属于“潜水”的角色,一直在接收群信息但不发言。

结合群成员互动强度和活跃程度,即可构建出基于每个话题的三元联系表(成员-亲密程度-成员),两边的节点代表群成员,边表示亲密程度类型,成员的属性为活跃程度。在检索时,通过检索目标成员名称即可获取与该成员产生亲密关系的所有成员和成员的活跃程度。可用于进一步挖掘与目标成员关系密切的其他成员信息,刻画出人物行为画像,从而为梳理群成员之间的关系提供有力支撑。

4　结束语

为解决词汇特征的冗余问题,提出了一种基于候选主题词的话题分类算法。该算法首先使用 LDA 主题模型抽取群聊天话题中的关键词,获取文本的强语义表征,同时算法融合了词汇和字符不同维度的特征,提升了话题分类的性能。为了丰富人物行为的描述,提出的人物刻画方案,不仅关注话题类别,同时也关注到成员间的互动强度、成员活跃程度、成员社会关系等多个维度。实验结果和分析验证了该方法的有效性和可行性,为群聊天中的人物行为刻画提供了新的分析视角,对海量电子数据中关键信息和人员的识别和定位提供有力的支撑,同时,有助于进一步挖掘关键聊天内容、成员及成员关系。提出的话题分类方法针对单个话题,实现多话题标签的分类是本研究持续关注和优化的方向。

参考文献:

[1]　中国互联网信息中心. 第 47 次中国互联网络发展状况统计报告[R]. 北京:中国互联网信息中心,2021.

[2]　WETHERELL M,TAYLOR S,YATES S J. Discourse theory and practice[M]. London:SAGE,2001:47-56.

[3] 吴亚欣,于国栋. 为会话分析正名[J]. 山西大学学报:哲学社会科学版,2017,40(1):85-90.
[4] 梁 卉. 网络语言的会话结构分析[J]. 长沙大学学报:哲学社会科学版,2007,21(1):102-103.
[5] 王宏军. 会话结构的语用研究方法述评[J]. 天津外国语学院学报,2006,13(5):67-71.
[6] 迟呈英,李 红. 基于改进 TF * PDF 算法的网络新闻热点话题检测和跟踪[J]. 计算机应用与软件,2013,30(12):311-314.
[7] 李 慧,王丽婷. 基于词项热度的微博热点话题发现研究[J]. 情报科学,2018,36(4):45-50.
[8] 陈龙稳. 基于改进的 SingIe-Pass 算法微博话题发现[J]. 现代计算机,2016(29):22-25.
[9] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. 2014. https://arxiv. org/ pdf/1408. 5882. pdf.
[10] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences [C]//52nd annual meeting of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2014:655-665.
[11] KIM Y, LEE H, JUNG K. AttnConvnet at SemEval -2018 task 1: attention - based convolutional neural networks for multi - label emotion classification [C]//The 12th international workshop on semantic evaluation. Stroudsburg: Association for Computational Linguistics, 2018:141-145.
[12] JOHNSON R, ZHANG T. Deep pyramid convolutional neural networks for text categorization[C]//55th annual meeting of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2017:562-570.
[13] SHIMURA K, LI J, FUKUMOTO F. HFT-CNN: learning hierarchical category structure for multi-label short text categorization[C]//2018 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2018:811-816.
[14] NGUYEN H, NGUYEN M. A deep neural architecture for sentence level sentiment classification in twitter social networking[C]//15th international conference of the pacific association for computational linguistics. Singapore: Springer, 2017:15-27.
[15] ADAMS B, MCKENZIE G. Crowdsourcing the character of aplace: character-level convolutional networks for multilingual geographic text classification[J]. Trans. GIS, 2018, 22(2):394-408.
[16] CHEN Z, QIAN T. Transfer capsule network for aspect level sentiment classification[C]//57th conference of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2019:547-556.
[17] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. [s. l.]:[s. n.], 2013:3111-3119.
[18] RODRIQUEZ K J, BRYANT M, BLANKE T, et al. Glove: global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2014:1532-1543.
[19] MCCANN B, BRADBURY J. Learned in translation: contextualized word vectors[EB/OL]. 2017-08-01. https://arxiv. org/abs/1708. 00107. pdf.
[20] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//2018 conference of the north american chapter of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2018:2227-2237.
[21] HOWARD J, RUDER S. Universal language model fine-tuning for text classification[C]//56th annual meeting of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2018:328-339.
[22] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training. [EB/OL]. 2018. https://s3 - us - west - 2. amazonaws. com/openaiassets/research-covers/languageunsupervised/language understanding paper. pdf.
[23] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Conference of the North American chapter of the association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 2019:4171-4186.
[24] 蒋建国. 微信群:议题、身份与控制[J]. 探索与争鸣,2015(11):108-112.
[25] 钱景怡,张玉峰. 基于人际网络挖掘的企业竞争情报获取研究[J]. 图书情报知识,2009(3):90-93.
[26] 孙晓玲,林鸿飞. 人际网络关系抽取和结构挖掘[J]. 微电子学与计算机,2008,25(9):233-236.
[27] 李 纲,李显鑫,巴志超,等. 微信群会话网络结构及成员角色划分研究[J]. 现代情报,2018,38(7):3-11.
[28] 康艳荣,赵 露,范 玮,等. 基于微信聊天记录时间信息的人物行为刻画技术研究[J]. 刑事技术,2018,43(3):187-192.
[29] BLEID M, NGA Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[30] 李 纲,王 晓,郭 洋. 基于成员合作共现的微信群内部关系研究[J]. 数据分析与知识发现,2018,2(11):54-63
[31] 李 纲,王馨平,巴志超. 微信群中会话网络结构及用户交互行为分析[J]. 情报理论与实践,2018,41(10):124-130.
[32] PHAN Xuan-Hieu, NGUYEN Cam-Tu. GibbsLDA++: a C/C++ implementation of latent Dirichlet allocation[EB/OL]. 2007. http://gibbslda. sourceforge. net/.