浙江科技学院学报,第 31 卷第 5 期,2019 年 10 月
Journal of Zhejiang University of Science and Technology
Vol. 31 No. 5, Oct. 2019

doi: 10.3969/j.issn.1671-8798.2019.05.007

基于评论的热点新闻事件识别方法研究

郑飘飘[1],万　健[1],司华友[2]
(1.浙江科技学院 信息与电子工程学院,杭州 310023;2.杭州电子科技大学 计算机学院,杭州 310018)

摘　要:随着互联网的普及,非结构化文本数据的规模不断扩大且越来越多地用于大众传播。因此,从海量数据抽取热点信息已成为一个重要的研究课题。针对新闻的热点挖掘进行方法改进及分析,结合新闻及事件模型,使用 TextRank 算法提取关键词,运用相似度计算方法,提出了一种基于评论的热点新闻事件识别方法。研究结果表明该方法具有一定的可行性。

关键词:新闻;评论;事件识别;信息抽取
中图分类号:TP391.43　　**文献标志码**:A　　**文章编号**:1671-8798(2019)05-0392-08

Research on methodology of identifying hot news event based on comments

ZHENG Piaopiao[1], WAN Jian[1], SI Huayou[2]
(1. School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China; 2. School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)

Abstract: With the popularity of the Internet, the scale of unstructured text data is drastically expanding and increasingly used for mass communication. Therefore, extracting hotspot information from massive data has become an important research topic. Aiming at method improvement and analysis of hot news mining, a method for identifying hot news events based on comments was proposed by virtue of similarity calculation method, in combination with the news and event model, and the TextRank algorithm for extracting keywords. Experimental results show that this method has certain feasibility.

Keywords: news; comment; event identification; information extraction

随着互联网的发展,网络信息规模急速扩大。网民逐渐习惯使用网络获取和浏览信息,其中网络新闻亦是网络信息的一个重要传播途径。新闻具有重要、新鲜、时效性强、真实等特点,且能够在一定篇幅

收稿日期:2019-05-02
基金项目:国家自然科学基金项目(61572163)
通信作者:万　健(1969—　),男,福建省泉州人,教授,博士,主要从事云计算大数据研究。E-mail:wanjian@zust.edu.cn。

内为读者提供大量信息。但同时又因为互联网的开放属性，新闻网页常常具有多变、冗余、异构等特点，报道相同内容的新闻会被发布在不同的网页上，表现形式亦各有不同。为实现在繁杂的新闻报道中捕获热点新闻事件信息，新闻事件识别的重要性变得不言而喻。

近年来，由于使用主流商业搜索引擎进行信息检索，信息抽取引起了研究者的关注。Kluegl 等[1]设计了一种工具以帮助计算客观评价，以符合平台严格要求的方式从 Web 可访问的半结构化资源中提取相关数据。基于信息提取的机器学习由于其有效性和高效性，特别是在许多共享任务中取得的成功而受到越来越多的关注。一些机器学习方法直接用于自然语言处理任务，如肿瘤信息提取[2]。信息提取是自然语言处理领域的一个热点问题，其中事件提取是三个主要任务之一。多样的自然语言处理系统已经被开发出来，并被用来从文本中提取事件和概念，应用这些工具的几个成功案例已经被广泛报道[3-4]。Yazdani 等[5]提出了一种基于心房和心室活动分析的方法，该方法基于最近提出的一种从生物医学信号中提取短期事件的新型非线性滤波技术。虽然事件的定义可能存在很大差异，但是事件的概念在自然语言处理的许多相关研究领域中得到了广泛的应用。Deve 等[6]认为，新闻事件是在特定的时间和地点发生的特定的事情，且可以在一段时间内被各种媒体连续报道。基于 Allan 等[7]提出的方法，目前最流行的新闻事件检测方法主要是 Single Pass 和凝聚层次聚类算法的变体。*K*-means 算法及其变体[8-9]是目前最流行的层次聚类算法，但其缺陷在于：有效性依赖于随机初始化，且随着数据量的增加而降低[10]。一些与本体论、机器学习和自然语言处理相关的技术被应用于帮助文献工作者对新闻进行分类和标记[11]。许多研究人员将事件检测应用于特定领域[12-13]，例如检测类似并购的可能影响市场的经济事件。Yang 等[14]提出了一个事件检测框架，用于发现来自多个数据域的真实事件，包括在线新闻媒体和社交媒体。从大量的新闻报道中提取特定的信息并非易事，许多研究都面临着尚未解决的问题。通过对大量热点事件的分析，发现热点事件的重要组成部分即评论往往被忽略，然而具有大量评论的新闻形成事件的可能性更大。基于此，笔者提出的算法综合评论和时间因素，由高评论的新闻构建初始事件库，设计包含时间因素的相似度计算方法，新闻归入事件后再进行事件特征项权重调整，完成热点事件识别，并通过试验结果分析验证算法的有效性。

1　模型及定义

1.1　问题描述及概念定义

新闻事件通常指在特定地点和时间发生的特定事情，通常多篇相似的新闻文章在一段时间内会被连续报道。如图 1 所示，新闻事件由四部分构成，包括时间、地点、人物和内容。就本质而言，事件识别依据新闻描述的不同事件将新闻报道聚类，把描述同一事件的新闻报道聚为同一簇。跟往常的文本聚类相比较，其特殊性表现为两方面：1）事件识别的对象针对按时间顺序依次出现的新闻报道数据流，它不是封闭静态的，是随时间动态改变的文本集合；2）事件识别依据报道描述的事件进行聚类，而不是依据话题类别进行聚类，区别于话题类别，事件依据的信息粒度更小，因此通过事件识别能够得到更多类。即使如此，事件识别的基础仍然是文本聚类分析。

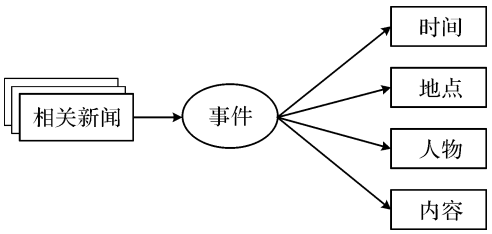

图 1　新闻事件的基本构成

Fig. 1　Basic composition of news event

我们所提出的新闻事件识别方法的流程可概括为四部分：用文本表示模型表示事件和新闻；用相似度算法计算事件和新闻间的相似度；对比计算所得相似度值和给定的相似度阈值；根据比较结果对新闻能否归属于某一事件进行判断。在识别新闻事件的过程中，可能会遇到一篇新闻报道与多个事件均相关的情形，但此时相关程度并不会完全相同[15]，可以把新闻归类到与之相关度最高的事件中。

1.2 事件与新闻模型

事件识别常用的表示方法是向量空间模型，其中使用向量表示文本。向量空间模型在起始时通过遍历所有文档完成词袋构建，文档篇幅越长、篇数越多意味着用于表示文本的词向量维度越高，并且当有新的文档加入时，词袋需要重新构建。由此，我们提出一种新的文本表示方法，可以有效解决上述模型的两点不足：相似度计算量较大且有新文本加入时，需要重新计算标引词权重；不适用于处理长度较大的文本，因其近似值不理想（过高的次元及过小的标量积）。

1.2.1 事件模型

热点事件模型的具体数学表达为 $E\{e_1:w_1,e_2:w_2,\cdots,e_n:w_n\}$，其中，$e_i$ 指事件特征项，即事件关键词；w_i 指对应特征项 e_i 的权重。初始事件关键词及权重可以在各事件中用户关注度最高的一篇新闻报道中继承得到，事件关键词及权重再随着后续新闻报道归入同一事件时进行动态调整，从而使得事件模型更具代表性。

1.2.2 新闻模型

新闻主体由新闻标题和新闻正文组成，新闻模型构建从每篇新闻主体中提取特征项开始，其中，特征项指从新闻主体中抽取的关键词。key-value 向量数据类型 $\mathbf{N}\{n_1:w_1,n_2:w_2,\cdots,n_k:w_k\}$ 用于表示具体新闻模型，其中，$n_i(1\leqslant i\leqslant k)$ 表示新闻特征项，即从新闻主体中抽取的新闻关键词；$w_i(1\leqslant i\leqslant k)$ 指对应特征项 n_i 的权重。新闻关键词及权重可以使用关键词抽取算法得到。

1.3 关键词提取

TextRank 关键词提取算法把词视作万维网上的节点，根据词与词之间的共现关系来计算每个词的重要性，并将 PageRank 算法（Google 公司）中的有向边转换成无向边。相较于 PageRank 算法，TextRank算法核心公式多了一个权重项 w_{ji}，用于呈现两个节点间的边连接具有不同的重要程度，其核心公式如下。

$$W_s(V_i)=(1-d)+d\times\sum_{V_j\in \mathrm{In}(V_i)}\frac{w_{ji}}{\sum\limits_{V_k\in \mathrm{out}(V_j)}w_{jk}}W_s(V_j)。\tag{1}$$

运用 TextRank 算法针对每篇新闻报道按照下列步骤对关键词进行提取：

1）将整个句子分割成给定文本，即文本 $T=[S_1,S_2,\cdots,S_m]$；

2）针对每个句子 $S_i\in T$ 进行分词、标注词性、剔除停用词和替换同义词，保留特定词性的词，例如名词、动词、形容词等，即 $S_i=[t_{i1},t_{i2},\cdots,t_{in}]$，其中 t_{ij} 表示句子 S_i 中保留下的词；

3）构造词图 $G=(V,E)$，V 表示上一步骤里生成的词所构成的节点集合，再用共现关系构造任意两个节点之间的边 E（两个节点之间有边，当且仅当它们的对应词在长度 k 的窗口中共现，k 表示窗口大小，k 默认为 5）；

4）依据式（1）迭代计算每个节点权重，直至收敛；

5）倒序排序节点权重，将最重要的 t 个词作为 top-t 关键词；

6）针对 top-t 关键词在原始文本中进行标记，若它们之间构成相邻词组，就提取为关键词组。

1.4 相似度计算

利用新闻相似性比较方法查找描述同一事件的新闻集合。针对事件与新闻间的相似性，我们提出一种基于关键词相关分析的新方法用于计算新闻事件相似性。相关定义如下：N，一篇新闻报道；E，一个事件；t_0，当前时间；P_1，集合 $P_1=\{k_1,k_2,\cdots,k_m\}$，其中 k_i 指同时出现在新闻 N 和事件 E 中的关键词；P_2，集合 $P_2=\{w_1,w_2,\cdots,w_m\}$，其中 w_i 指 P_1 中每个关键词的权重；P_3，集合 $P_3=\{t_1,t_2,\cdots,t_m\}$，其中 t_i 指 P_1 中每个关键词最近更新的时间；P_4，集合 $P_4=\{e_1,e_2,\cdots,e_n\}$，包含事件的所有关键词，其中 e_j 表示事件的一个关键词；P_5，集合 $P_5=\{s_1,s_2,\cdots,s_n\}$，包含 P_4 中每个关键词的权重，其中 s_j 对应 e_j 的权重；P_6，集合 $P_6=\{q_1,q_2,\cdots,q_n\}$，其中 q_j 指 P_4 中每个关键词最近更新的时间。

事件 E 与新闻 N 的相似度计算公式：

$$\mathrm{Sim}(N,E)=\frac{\sum_{i=1}^{m}\frac{1}{t_0-t_i}\times w_i}{\sum_{j=1}^{n}\frac{1}{t_0-q_j}\times s_j}。\tag{2}$$

式(2)表明，事件与新闻相同的关键词越少，相同关键词更新时间相距越远，则相似度值越低。由式(2)计算得到的相似度值处于 0～1 之间，相似度值越接近 0，表示新闻 N 描述的内容是事件 E 的可能性越小。将该相似度值与给定的相似度阈值比较，如果大于给定阈值，就把新闻 N 归类到事件 E 中；反之，将新闻 N 作为新事件存入事件库中或直接舍弃。

2　基于评论的新闻事件提取方法

计算新闻报道与所有已存在事件的相似度值，比较得到其中最大相似度值。如果最大相似度值大于给定相似度阈值，就将该新闻归类到对应事件；反之，先判断该新闻在前置聚类比例的范围内与否，再裁定是创建新事件还是直接丢弃。另外，当新闻归入之前存在的事件中时，需要调整对应事件的关键词权重。算法流程如图 2 所示。

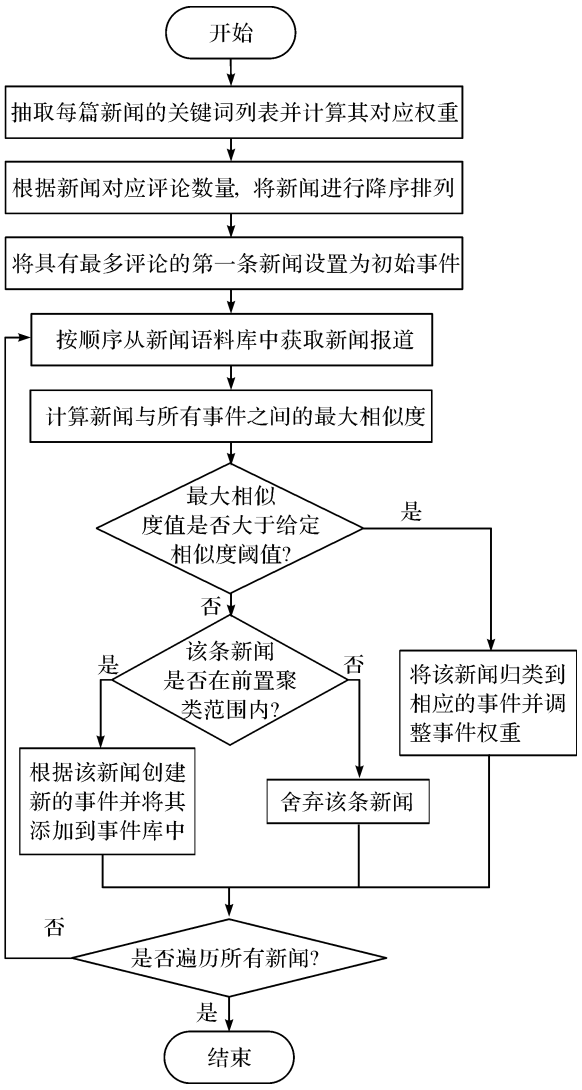

图 2　算法流程图

Fig. 2　Algorithm flow chart

此算法中，前置新闻聚类比例(pre-news clustering proportion，POPC)、相似度阈值(similarity threshold，ST)和表示新闻的关键词数量(number of keyworks，NOK)是有可能影响热点事件检测结果的 3 个因素。此算法的关键问题：其一，利用初始局部聚类自动生成一个热事件库，即初始热点事件数量及内容会受到前置聚类比例的影响，若前置聚类比例较高，那么更多的新闻会被聚集在一起，可以提取的事件也就越多；其二，此算法将计算所得到的最大相似度值与相似度阈值进行比较，用以确认是否将新闻归类于事件库中，如果相似度阈值太高，将难以对新闻进行归类，由此，相似度阈值会对新闻聚类结果有显著影响；其三，基于表示新闻的关键词序列的相似度计算，由此可知，表示新闻的关键词数量增减会直接影响相似度计算结果，进而影响最大相似度值。

3　新闻事件识别因子实证研究

对 3 项影响因子展开实证研究以评估本文方法，并与经典聚类算法作对比。采用 Excel 绘制图表以展示试验结果。

3.1　试验数据

试验所涉及的新闻数据来自新浪、搜狐、凤凰、网易、腾讯五个新闻门户网站，其时间范围由 2017-07-01 至 2017-08-09，共 19 681 篇新闻报道。采用人工统计法标注新闻数据集，筛选出 10 个热点事件形成标准事件集，标明每篇新闻报道与每个标准事件相关与否，且一篇新闻报道只与一个事件相关，与其他事件均不相关。表 1 即标准事件集，第三列表示与每个热点事件相关的新闻数量，由人工标注每篇新闻报道与表中 10 个热点事件是否相关，若相关，则为该篇报道标注对应事件号，进而统计每个事件号总数得出每个事件的相关新闻数量。

表1 标准事件集

Table 1 Standard event set

事件号	事件名称	相关新闻数量
1	战狼2上映	103
2	崔永元辞去商城职务	8
3	徐峥被曝光殴打女记者	10
4	天津静海传销组织蝶贝蕾	129
5	九寨沟地震	258
6	邹市明惜败WBO世界拳王金腰带卫冕赛	54
7	中国福建教师在日本失踪	32
8	京东和苏宁的物流战	37
9	保定容大平局后威胁退出中超联赛	110
10	法院冻贾跃亭资产	504

3.2 评估指标

评价指标体系包括四部分：召回率(R_r)、精确率(R_p)、生成率(R_g)和F_1值(F_1)，用以评估新闻事件识别准确程度。标准事件集见表1中，事件总计为10，标记为常数n；标准事件集$E=\{E_1,E_2,\cdots,E_n\}$，其中，每个标准事件$E_i=\{e_1,e_2,\cdots,e_k\}$，$k\leqslant 10$；提取出的事件集合$T=\{T_1,T_2,\cdots,T_r\}$，其中，$T_i=\{t_1,t_2,\cdots,t_g\}$，$g$由试验参数动态决定；若事件$T_j$与事件$E_i$拥有的相同关键词比例大于30%，则认为$E_i\in E\cap T$。由此，生成率公式表示为：

$$R_g=\frac{E\cap T}{n}。$$

生成率等于1表示事件识别效果最佳；生成率等于0，则表示事件识别效果最差。

基于我们所提出的事件识别方法，契合标准事件集和生成事件集，精确率及召回率可表述为：

$$R_p=\frac{\sum_{i=1}^{n}\frac{\max|E_i\cap T_k|}{|E_i|}}{n},\tag{3}$$

$$R_r=\frac{\sum_{i=1}^{n}\frac{\max|E_i\cap T_k|}{|T_k|}}{n}。\tag{4}$$

当存在事件T_j使得$E_i\cap(\forall T_j\in T)$取得最大值，将$T_j$定义为$T_k$。由式(3)～(4)可知，精确率为正确聚类到相关标准事件的新闻占标准事件中新闻总数的比率均值，召回率为正确聚类到相关标准事件的新闻占聚类得到事件中新闻总数的比率均值。

用F_1值综合上述2项指标，赋予精确率与召回率相同权重。

$$F_1=\frac{2R_pR_r}{R_p+R_r}。$$

其中，若F_1值越接近1，则表明识别新闻事件的效果越好。

3.3 实证试验

试验1：利用评价指标体系对本文识别方法进行验证评估。试验总共包含140组并行测试，小组间的差异来自前置新闻聚类比例、相似度阈值和表示新闻的关键词数量3个参数。其中，前置新闻聚类比例分别是1%、3%、5%和7%；相似度阈值取值10%～70%，间隔为10%；表示新闻的关键词数量4～12个，间隔为2。总计140($4\times7\times5$)个不同的组合。

试验2：对比经典事件聚类算法变体与新闻事件识别算法的性能。选择常用聚类算法作为比对基准。其中，K-means算法把新闻数据集合划分为K类簇，初始化聚类中心采用Forgy方法与随机划分方法，Forgy方法从新闻数据集合中随机选择K个聚类中心当作初始均值，随机划分方法先为每个聚类中心随机分配类簇，然后进入更新步骤，计算类簇均值当作中心点，逐步迭代直至满足终止条件；过滤算法

运用 kd-tree 加速每个 K-means 步骤[16]。另外，mean shift 聚类，affinity propagation 聚类，agglomerative 聚类，spectral 聚类，birch 聚类，dbscan 聚类和小批量 K-means 聚类均处于对比列表中。

3.4　结果分析

如图 3 所示，从形状上看，数据的线性相关性比较相似。由图 3 可知，使用 10 个关键词表示一篇新闻报道促使聚类结果在 4 个评价指标上的表现较为均衡，同时也略好于用其他数量关键词表示一篇新闻，这表明用 10 个关键词使事件识别效果更好。

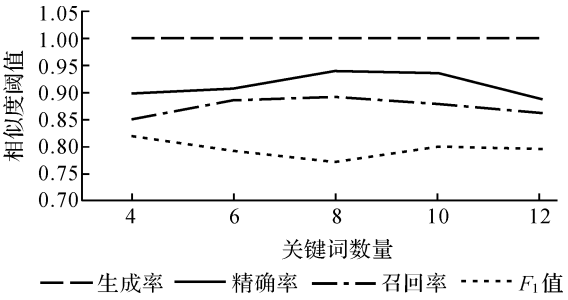

图 3　表示新闻的关键词数量测评

Fig. 3　Evaluation on number of keywords

图 4～7 表明前置新闻聚类比例和相似度阈值对我们所提出事件识别方法的影响。通过分析可知，针对前置新闻聚类比例，取值 7%时优于其他选项，且前置新闻聚类 7%搭配相似度阈值 40%的组合使得各评估指标达到最优，事件识别效果最佳。在此条件下，本文算法所抽取出的事件关键词在一系列演化后结果见表 2，事件号及名称与表 1 一致。

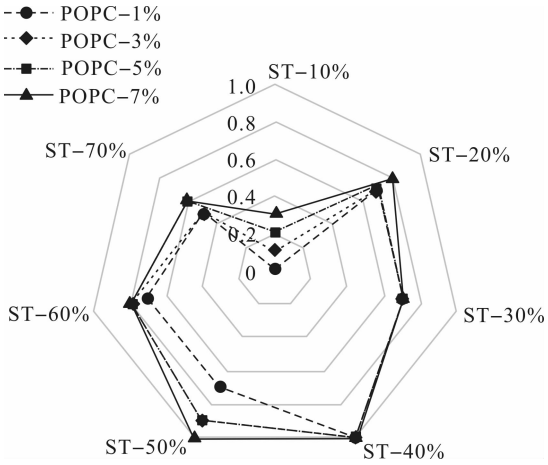

图 4　POPC 和 ST 不同组合的生成率(NOK＝10)

Fig. 4　Generation rates of various combinations of POPC and ST(NOK＝10)

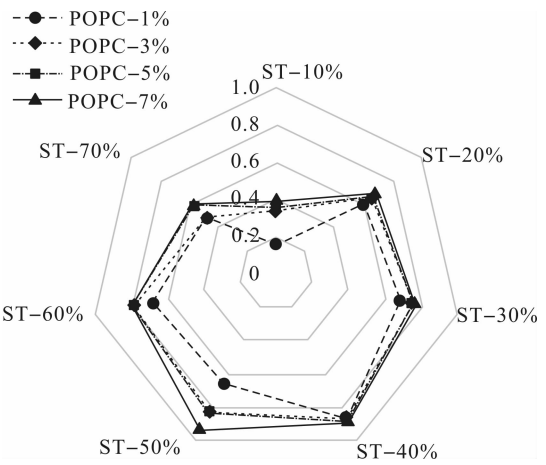

图 5　POPC 和 ST 不同组合的精确率(NOK＝10)

Fig. 5　Accuracy rates of various combinations of POPC and ST(NOK＝10)

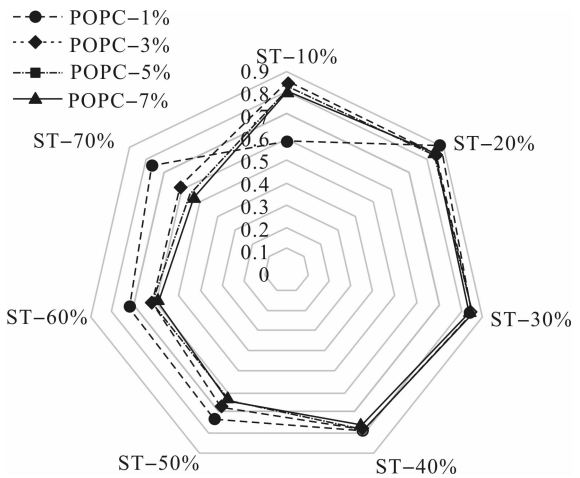

图 6　POPC 和 ST 不同组合的召回率(NOK＝10)

Fig. 6　Recall rates of various combinations of POPC and ST(NOK＝10)

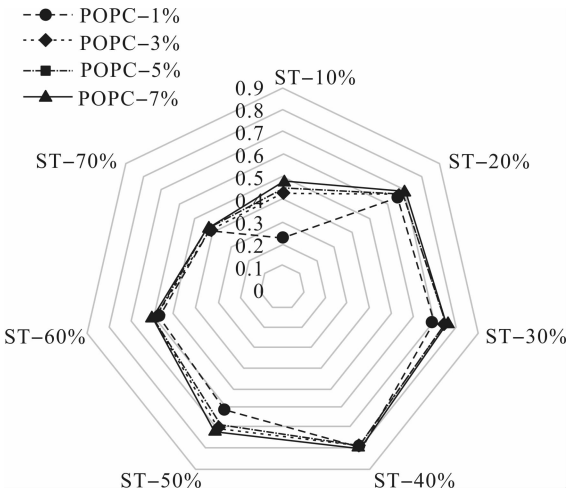

图 7　POPC 和 ST 不同组合的 F_1 值(NOK＝10)

Fig. 7　F_1 values of various combinations of POPC and ST(NOK＝10)

表 2 事件关键字列表

Table 2 Keyword list for events

事件号	事件名称	事件关键词
1	战狼 2 上映	公司,战狼 2,票房,谢世,电影,主旋律,市场,中国,院线
2	崔永元辞去商城职务	谷塘,辞去,崔永元,职务,转基因,食品,商城,利益集团,横遭,报复
3	徐峥被曝光殴打女记者	曝出,徐峥,跟踪,女子,和解,否认,殴打,气愤,微博,偷拍
4	天津静海传销组织蝶贝蕾	传销,活动,犯罪,善心汇,公安部,策划,涉嫌,公安机关,21
5	九寨沟地震	阿坝州,救援,地震,九寨沟,四川,应急,游客,九寨沟县,供电机场
6	邹市明惜败 WBO 世界拳王金腰带卫冕赛	邹市明,职业,对手,木村,回合,世界,拳击,进攻,体能
7	中国福建教师在日本失踪	危秋洁,福建,失联,日本,发现,札幌,行程,酒店,记者,教师
8	京东和苏宁的物流战	快递,孙为民,物流,中国,京东,服务,电商,桐庐,苏宁,速度
9	保定容大平局后威胁退出中超联赛	俱乐部,希望,球迷,足球,比赛,处罚,保定,事件,退出,卓尔
10	法院冻贾跃亭资产	媒体,贾跃亭,乐视,司法,冻结,金融机构,公司,控股,供应商

在相似性阈值单维度上,验证了精确度与召回率的反比关系,即一个指标增强的代价为另一个指标的减弱。表 3 列出表示新闻的关键词数量 10、前置新闻聚类 7%时,单维度相似性阈值上的试验具体数值。通常,精确度和召回率不会被单独解析和讨论,而是相互结合构成 F_1 值(精确率和召回率的加权调和平均值)[17]。

表 3 单维度相似性阈值上的试验数据

Table 3 Experimental data on single-dimensional similarity threshold

相似性阈值/%	精确率	召回率
−10	0.385 612 102	0.812 311 218
−20	0.680 086 392	0.840 866 789
−30	0.768 732 519	0.837 736 526
−40	0.883 824 456	0.759 592 619
−50	0.936 733 762	0.626 990 142

经典聚类算法变体和所提出事件识别算法的比较结果见表 4。其中,经典聚类算法中存在 5 种算法,在相同硬件条件下无法成功聚类,包括 affinity propagation 聚类、mean shift 聚类、agglomerative 聚类、spectral 聚类和 birch 聚类,其报错原因在于内存不足。

表 4 与经典聚类算法变体的对比试验结果

Table 4 Comparison of experimental results with classical clustering algorithm variants

算法	生成率	精确率	召回率	F_1 值
multi-feature model	0	0	0	0
K-means	1	0.294 310 455 054 21	0.663 352 283 163 753	0.279 574 848 325 917
mini batch k-means	0	0.192 740 686 011 107	0.695 701 836 791 106	0.181 561 033 920 5
dbscan clustering	0	0.006 325 898 074 284 84	1	0.012 466 969 175 432 7

综合 4 个评价指标,本文识别方法优于其他聚类算法且对内存要求不高。

4 结 语

本研究提出了一种基于评论的热点新闻事件识别方法,采用 key-value 向量数据类型的方式构建新闻及事件模型,使用 TextRank 算法提取关键词,并利用结合时间因素在内的相似度计算方法,进行热点事件识别。进而通过对试验结果的分析发现,使用特定的 POPC、ST 和 NOK 能使热点事件识别效果更佳。该方法对从海量网络新闻中检测热点事件的实际应用有一定的参考价值。然而,笔者将相似度阈值设为固定值有一定的局限性;另外,许多热点事件可能首先出现在微博、论坛中,为提供更全面的社交热点事件,信息来源应由多个数据源集成。因此,今后的研究将围绕这两点展开。

参考文献:

[1] KLUEGL P, TOEPFER M, BECK P D, et al. UIMA Ruta: rapid development of rule-based information extraction applications[J]. Natural Language Engineering, 2016, 22(1): 1.

[2] YIM W, DENMAN T, KWAN S W, et al. Tumor information extraction in radiology reports for hepatocellular

carcinoma patients[C]//2016 Joint Summits on Translational Science. San Franciso：AMIA，2016：455.

[3] WI C I, SOHN S, ROLFES M C, et al. Application of a natural language processing algorithm to asthma ascertainment. An automated chart review[J]. American Journal of Respiratory and Critical Care Medicine，2017，196(4)：430.

[4] AFZAL N, MALLIPEDDI V P, SOHN S, et al. Natural language processing of clinical notes for identification of critical limb ischemia[J]. International Journal of Medical Informatics，2018，111：83.

[5] YAZDANI S, FALLET S, VESIN J M. A novel short-term event extraction algorithm for biomedical signals[J]. IEEE Transactions on Biomedical Engineering，2018，65(4)：754.

[6] DAVE K, LAWRENCE S, PENNOCK D M. Mining the peanut gallery：Opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th international conference on World Wide Web. Budapest：ACM，2003：519.

[7] ALLAN J, CARBONELL J G, DODDINGTON G, et al. Topic detection and tracking pilot study final report[C]//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco：Morgan Kaufmann Publishers，1998：194.

[8] CAPO M, PÉREZ A, LOZANO J A. An efficient approximation to the *K*-means clustering for massive data[J]. Knowledge-Based Systems，2017，117：56.

[9] CUI X, ZHU P, YANG X, et al. Optimized big data K-means clustering using MapReduce[J]. The Journal of Supercomputing, 2014，70(3)：1249.

[10] STEINBACH M, KARYPIS G, KUMAR V. A comparison of document clustering techniques[C]//KDD-2000 Workshop on Text Mining. Boston：SIGKDD，2000：1.

[11] GARRIDO A L, BUEY M G, ESCUDERO S, et al. The genie project：a semantic pipeline for automatic document categorisation[C]//The 10th International Conference on Web Information Systems and Technologies. Barcelona，Spain：WEBIST，2014：161.

[12] LEFEVER E, HOSTE V. A classification-based approach to economic event detection in dutch news text[C]//Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia：European Language Resources Association (ELRA)，2016：330.

[13] JACOBS G, LEFEVER E, HOSTE V. Economic event detection in company-specific news text[C]//The 56th Annual Meeting of the Association for Computational Linguistics. Melbourne，Australia：Association for Computational Linguistics，2018：1.

[14] YANG Z, LI Q, WENYIN L, et al. Shared multi-view data representation for multi-domain event detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，99：1.

[15] RUPNIK J, MUHIČA, LEBAN G, et al. News across languages：cross-lingual document similarity and event tracking [J]. Journal of Artificial Intelligence Research，2016，55(1)：283.

[16] CUI X, ZHU P, YANG X, et al. Optimized big data K-means clustering using MapReduce[J]. The Journal of Supercomputing，2014，70(3)：1249.

[17] POWERS D MW. Evaluation：from precision，recall and F-measure to ROC，informedness，markedness and correlation [J]. Journal of Machine Learning Technologies，2011，2(1)：37.