



用于网络新闻热点识别的热点新词发现

王煜*, 徐建民

(河北大学 网络空间安全与计算机学院, 河北 保定 071000)

(* 通信作者电子邮箱 wy@mail.hbu.edu.cn)

摘要:通过分析网络新闻热点词的特点,提出了一种用于网络新闻热点识别的热点新词发现方法。首先,用改进FP-tree算法提取频繁出现的词串作为热点新词候选,删除新闻数据中非频繁1-词串,并利用1、2-非频繁词串切割新闻数据,从而删除新闻数据中的大量无用信息,大幅降低FP-tree复杂度;其次,根据二元逐点互信息(PMI)扩展成多元PMI,并引入热点词的时间特征形成时间逐点互信息(TPMI),用TPMI判定热点新词候选的内部结合度和时间性,剔除不合格的候选词;最后,采用邻接熵确定候选新词边界,从而筛选出热点新词。采集百度网络新闻的7 222条新闻标题作为数据集进行实验验证。在将半月内报道次数不低于8次的事件作为热点新闻且时间特征的调节系数为2时,采用TPMI可以正确识别51个热点词,丢失识别2个长时间热点词和2个低热度词,而采用不加入时间特征的多元PMI可正确识别全部热点词55个,但错误识别97个非热点词。分析可知所提的算法降低了FP-tree复杂度,从而减少了时间空间代价,实验结果表明判定热点新词时加入时间特征提高了热点新词识别率。

关键词:热点新词;FP-tree;逐点互信息(PMI);邻接熵;时间特征

中图分类号:TP391 **文献标志码:**A

Hot new word discovery applied for detection of network hot news

WANG Yu*, XU Jianmin

(School of Cyber Security and Computer, Hebei University, Baoding Hebei 071000 China)

Abstract: By analyzing the characteristics of hot words in network news, a hot new word discovery method was proposed for detection of network hot news. Firstly, the Frequent Pattern tree (FP-tree) algorithm was improved to extract the frequent word strings as the hot new word candidates. A lot of useless information in the news data was reduced by deleting the infrequent 1-word strings from news data and cutting news data based on infrequent 1, 2-infrequent word strings, so as to greatly decrease the complexity of FP-tree. Secondly, the multivariate Pointwise Mutual Information (PMI) was formed by expanding the binary PMI, and the Time PMI (TPMI) was formed by introducing the time features of hot words. TPMI was used to judge the internal cohesion degree and timeliness of hot new word candidates, so as to remove the unqualified candidates. Finally, the branch entropy was used to determine the boundary of new words for selecting new hot words. The dataset formed by 7 222 news headlines collected from Baidu network news was used for the experiments. When the events reported at least 8 times in half a month were selected as hot news, and the adjustment coefficient of time feature was set 2, TPMI correctly recognized 51 hot words, missed 2 hot words because they were hot for a long time and 2 less-hot words because they occurred insufficiently; the multivariate PMI without time features correctly recognized all 55 hot words, but incorrectly recognized 97 non-hot words. It can be seen from the analysis that the time and space cost is reduced by decreasing the complexity of FP-tree, and experimental results show that the recognition rate of hot new words is improved by introducing time feature during the hot new word judgement.

Key words: hot new word; Frequent Pattern tree (FP-tree); Pointwise Mutual Information (PMI); branch entropy; time feature

0 引言

网络信息具有传播速度快、影响范围广的特点。网络热点话题的识别与追踪通过整合互联网信息采集技术及信息智能处理技术对互联网海量信息进行处理,解决人们在海量信息中甄选话题的难题。热点词直接反映了热点话题的中心思想。因此,热点词的识别对于热点话题识别与追踪非常重要。

识别热点词首先需要分词系统可以将其识别为“词”。在自然语言处理中,中文处理技术比西文处理技术复杂。其中一个重要原因就是中文只有句和段能通过明显的分界符来简单划界,但作为句子基本单元的词却没有形式上的分界符。因此分词,也就是识别句中的词,成为了中文信息处理的基础。分词技术需要将已经存在的词存于词典中,分词依赖于

收稿日期:2020-04-28;修回日期:2020-06-27;录用日期:2020-07-03。

基金项目:国家社会科学基金资助项目(17FTQ002);河北省社会科学基金资助项目(HB15SH064)。

作者简介:王煜(1971—),女,河北保定人,教授,博士,主要研究方向:文本挖掘、信息检索;徐建民(1966—),男,河北保定人,教授,博士,主要研究方向:个性化信息检索、Web社区发现、话题识别与追踪、社会网络建模。



词典。热点话题往往涉及人名、机构名、地名、产品名、商标名、简称、事件名称等。这些热点涉及的词不断增加,词典中却往往并无这些词。分词系统无法识别词典中没有的词,但是这些词对于新闻热点发现又至关重要。因此,热点新词识别成为网络热点话题识别与追踪要解决的关键问题之一。

网络媒体信息量巨大,新词不断出现。针对这些新词目前国内已有许多新词发现的研究。这些研究针对不同背景从不同角度出发识别网络媒体新词,其中互信息和信息熵是新词发现的重要方法之一。文献[1]针对社会媒体文本的领域分布广、口语化程度高等特征提出一种面向社会媒体的开放领域新词发现算法,采用了标注模型和语料库频繁模式挖掘相结合的方法。文献[2]提出一种非监督的新词识别方法,该方法利用互信息的改进算法与少量基本规则相结合,从大规模语料中自动识别网络新词。文献[3]提出一种融合内外部统计量的微博新词发现方法,该方法针对目前新词发现算法中的数据稀疏以及可移植性较差的缺点提出了融合内外部统计量的改进 N-Gram 算法。文献[4]提出的特定领域新词检测利用组合互信息技术解决用户发明新词和转换感伤词的疏忽问题。文献[5]提出了一种从左至右逐字在未切词的微博语料中发现新词的算法。这些研究均改进互信息和邻接熵信息作为新词识别标准之一。目前新词识别的研究多是根据应用背景不同分析新词特点,针对其特征提出识别方法。目前多为针对微博、贴吧等社交媒体进行新词识别研究,例如文献[1-3,5-8]均是以此背景展开研究。此外,还有一些针对其他应用背景的研究,文献[4]针对旅游领域研究新词识别,文献[9]针对食品安全研究新词,文献[10]针对金融知识自动问答研究新词识别,文献[11]则是研究古汉语中新词识别。因新闻热点词具有独有的特征,这些研究并不适应网络新闻的热点新词识别。文献[12]中新闻热点的新词发现中仅用改进 FP-tree(Frequent Pattern tree)算法识别新词,没有考虑热点新词的特性。

要识别网络新闻中的热点新词,首先分析新闻热点词特性:

1)新闻热点词具有时间特征。也就是说热点词会在短期内变得频繁出现在新闻中,之前或之后随着热点热度消失后可能很少出现或不出现。

2)新闻热点主体涉及人名、机构名、地名、场所、产品名、商标名、简称等名称。这些名称数量巨大,放入分词的词典会造成巨大成本。因其大多不适合存于词典中,造成分词系统无法识别。

3)新闻热点词有时在分词系统中并不成为一个词,如某些事件被冠以一些名称,但分词技术往往不将其作为一个词。例如,2018年发生的“杀妻骗保案”。“杀妻骗保案”五个字是一个整体,代表了天津男子给妻子买3000余万保险后在普吉岛杀妻骗保这件案件。

4)新闻热点中的这些新词以名称居多,不符合一般词构成规律。例如人名音译、事件简称、地名等均毫无规律可言。

5)由于新闻标题要表达出新闻关键,所以热点词必存在于新闻标题中。

针对热点新词特性,本文先给出改进的 FP-tree 来寻找频繁出现在新闻标题中的词串作为新词候选,而不是按照构词规律寻找候选词;根据加入时间特征值的多元时间逐点互信息(Time Pointwise Mutual Information, TPMI)判断词的内部结合强度,根据邻接熵判断词边界,从而识别出热点新词;网络

上的舆情监控需要处理数以亿计的网页,过长的识别时间严重影响实用性,因此本文仅使用新闻标题识别热点新词以提高识别速度。

1 基于改进 FP-tree 的热点新词候选集确定

FP-tree 算法是 Apriori 算法的改进,大幅度减少了扫描数据次数,并且 FP-tree 的树形结构保存了频繁集的完整信息,去除掉非频繁集的数据,删减了无关的内容。为了快速找到新闻中新的热点词,本文采用改进 FP-tree 算法利用新闻标题快速获得热点新词的候选集。

1.1 改进 FP-tree 的建树

本文采用 FP-tree 算法寻找频繁出现的词串,将其作为热点新词的候选集。词中字间的前后顺序不能改变且中间不能有其他词,因此必须改进 FP-tree 算法。由于新闻标题中含有大量和热点新词无关信息,必须进行删减方可降低 FP-tree 规模。因此,本文从如下两方面进行 FP-tree 算法改进:

1)为维持词顺序,频繁 1-词串的生成结果无需排序。

2)为了减少 FP-tree 的非频繁内容,降低其复杂度,利用非频繁 1-词串和非频繁 2-词串进行新闻标题的化简分割。

一个词若为非频繁词,则其不可能出现在频繁词串中,可判定该词不可能出现在热点新词中。新词构成是必须连续的,那么非频繁词去掉后并可将标题进行分割,如“周五国内油价‘五连跌’几成定局或下调幅度超 260 元/吨”。若按照出现 8 次以下为非频繁词串,则其中“几成”“下调”“260”为非频繁词,则该标题分割为“周五国内油价‘五连跌’”“定局或”“幅度超”“元/吨”四条数据(“元/吨”因符号分割开)。由于新闻标题必须反映新闻核心内容的特性,新闻标题中短期内频繁出现的词相对较少,因而删减非频繁词可以大幅削减数据量。非频繁 2-词串也不可能为热点新词的一部分,可据此分割数据。假如“周五国内”“国内油价”“幅度超”三个词串为非频繁 2-词串,则继续分割标题为“周五”“国内”“油价”“幅度”“超”。切割后的数据都变得非常短,新闻标题被切成比较小的数据可提高处理速度。由于是识别新词,因此还需去掉只包含一个词的数据。之后,数据中无用信息量大幅减少。此例中,该新闻标题被全部去除。据此建立的 FP-tree 不仅包含所有热点新词有用信息且删除了大部分无用信息。

本文根据上述分析,采用三次扫描数据建立改进 FP-tree,算法步骤如下:

1)用分词工具进行分词,若“”和《》内无标点的字串被分开则合并,作为一个新词候选,并计入集合 D (注意:集合 D 中的新词只需要根据时间特征值判定是否为热点词即可)。

2)第一次扫描新闻标题获得频繁 1-词串 word 列表(每项包括频繁 1-词串和头指针),根据频繁 1-词串生成频繁 2-词串候选集,删除集合 D 中非频繁词。

3)第二次扫描新闻标题,删除新闻标题中非频繁 1-词串中的词并分割新闻标题,被分割后若只剩下一个词则被删除,同时统计频繁 2-词串候选计数。

4)第三次扫描第 3)步处理后的新闻标题,两个连续词若不为频繁 2-词串,则从两个词间分割标题,被分割后若只剩下一个词则被删除,同时根据分割删除后的新闻标题建立 FP-tree:首先初始化根节点为 null;对每条数据的词从根节点出发,依次对比,若存在相同的词则计数加 1,若不存在则增加新的孩子节点,计数 1;相同词串成一条链,头指针存于 word 列表。



例如,“重庆 公交 坠 江”“重庆 公交 坠 江 事故 后”“重庆 公交 坠 江 悲剧”“公交 坠 江 悲剧”和“坠 江 事故 后 重庆 公交”新闻数据形成的 FP-tree 如图 1 所示(假定最小频繁计数为 2)。

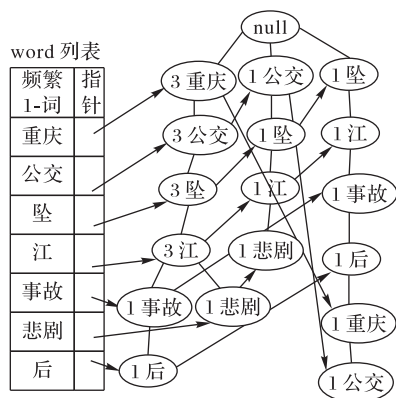


图 1 改进 FP-tree 示例

Fig. 1 Example of improved FP-tree

1.2 基于改进 FP-tree 的热点新词候选集生成

本文在改进 FP-tree 上挖掘新词候选的步骤如下:

1) 对集合 *word* 每个词在 FP-tree 上统计每个以该词为开头的所有词串的计数, 将频繁的词串 *x* 加入集合 *newword*, 如 *newword* 存在 *y*, 若 *y* 是 *x* 的子串且 *x* 和 *y* 的计数相同, 则删除 *y*; 若 *x* 是 *y* 的子串且 *x* 和 *y* 的计数相同, 则删除 *x*。

2) 根据图 1 挖掘的热点新词的候选集为 {“重庆 公交: 4”, “重庆 公交 坠 江: 3”, “公交 坠 江: 4”, “公交 坠 江 悲剧: 2”, “坠 江: 5”, “坠 江 事故 后: 2”}。

2 基于 TPMI 和邻接熵的热点新词判断

Pecina 等^[13]采用 55 种不同的统计量进行 2 元词汇识别实验, 结果表明逐点互信息 (Pointwise Mutual Information, PMI) 算法是最好的衡量词汇相关度的算法之一。通常情况下, PMI 方法能够很好地反映字串之间的结合强度, PMI 值越大表示结合字间程度越强。本文首先设计多元 PMI 的计算方法用来衡量候选新词的内部结合度, 并引入了时间特征。对于结合强度满足阈值的候选新词用邻接熵衡量其左邻接字和右邻接字符的不确定性, 解决新词左右边界问题。

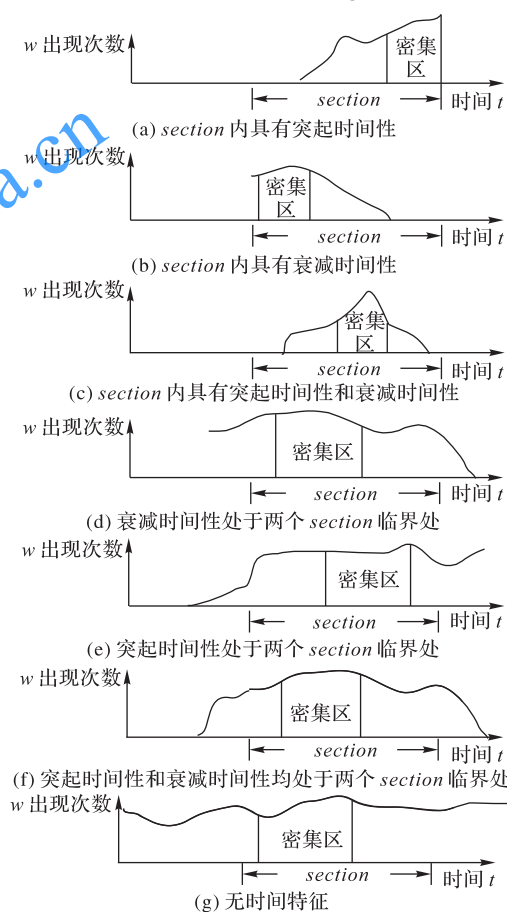
2.1 热点词时间特征值计算

本文的新词识别是热点新词的识别, 和普通新词识别不同。热点新词其实就是一种由不出现或极少出现的非频繁词串变得频繁出现的词串, 并且这个词串在新闻热度退去后又变为非频繁词串。因此热点新词具有开始短期内变得频繁出现的突起和之后衰减的时间特征。例如, 2014 年 3 月 8 日一架载有 239 人的马来西亚航空公司 MH370 客机在从吉隆坡飞往北京的途中失踪之后, 网络新闻里就爆发性出现“马航”这个新词, 具有短时间里突然增多的特性。而一些频繁的非热点词串, 如“外交部回应”“外媒关注”, 具有持续性, 不具有时间突起和衰减的特征。

本文将时间特征分为两种: 一是突起时间性, 由很少出现或不出现变为频繁出现; 二是时间衰减时间性, 由频繁词变为很少出现或不出现。在一个时间段 *section* 内, 热点新词可能在此前已经经过了突起时间性, 也可能在这个 *section* 内经历突起时间性, 或者突起时间性处于两个 *section* 临界处: 前一个

section 为非频繁, 而进入后一个 *section* 立刻变为频繁的。同样, 热点新词在这个 *section* 内可能经历衰减时间性, 可能在两个 *section* 临界处衰减, 也可能在下个 *section* 或之后才衰减。因此, 可以将 *section* 内热点词的时间特征分为以下七种情况 (图 2 中密集区表示 *section* 时间段内达到频繁计数一半的最密集处):

- 1) 在 *section* 时间段内具有突起时间性, 如图 2(a);
- 2) 在 *section* 时间段内具有衰减时间性, 如图 2(b);
- 3) 在 *section* 时间段内既具有突起时间性又具有衰减时间性, 如图 2(c);
- 4) 在 *section* 时间段内既不具有突起时间性又不具有衰减时间性, 但是和下一个 *section* 临界处具有衰减时间性, 可在处理下一个 *section* 获得衰减时间性, 但数据不继续采集则无法判断, 如图 2(d) 和图 2(f);
- 5) 在 *section* 时间段内既不具有突起时间性又不具有衰减时间性, 在和上一个 *section* 临界处具有突起时间性, 前移半个 *section* 获得突起时间性, 如图 2(e) 和图 2(f);
- 6) 无法获得突起时间性和衰减时间性, 热度维持时间长的事件的热点词具有此种情况, 如图 2(g)。

图 2 词 *w* 在时间段 (*section*) 内计数情况Fig. 2 Counts of word *w* in one period (*section*)

由图 2 可看出, 具有时间性的热点词分布是不均匀的, 具有集中性: 图 2(a)、(b)、(c) 三种情况, 密集区在 *section* 内所占时间比例要比 1/2 小很多; 而图 2(e)、(f) 两种情况, 若把时间段时间向前推移 1/2 的 *section* 时间段, 可以看出密集区所占时间比例也要比 1/2 小很多; 同样图 2(d)、(f) 两种情况, 在推后 1/2 的 *section* 时间段也可以统计到其密集区时间比例远低于



1/2;而不具有时间性的情况图2(g)就无法判断。考虑密集区数据比例太低无法判断 w 整体是否具有集中性,而比例太高则造成稀疏部分影响过大,因此选择了密集区包含词 w 的50%计数。因此本文根据包含一半计数的最频繁时间长短来判断词 w 的集中性,据此判定其时间性。本文设计了时间特征值计算式(1),判定候选新词 w 的时间特征值:

$$T(w) = a \times \frac{\text{section}}{\min_t (\text{time}(t) - \text{halftime}(t))} \quad (1)$$

其中: section 为选定的统计新闻的时间段的天数; $\text{time}(t)$ 为词 w 在这个时间段内出现的某个时间点 t ; $\text{halftime}(t)$ 为从时间点 t 开始词 w 出现次数达到该时间段内50%的时间点; $\min()$ 求最小值,即词 w 在 section 时间段内出现次数达到该段内总数50%的最短连续天数; a 是调节系数($a \geq 1/2$, $a=1/2$ 时均匀出现词的时间特征值在1左右)。新闻有时间性,大多数热点新闻很难持续高热度,少数新闻持续受关注,但热度也会降低;并且选择时段过长会加大数据计算量,因此 section 时间段不宜过长。考虑开始追踪新闻热点时,刚刚已经爆发的热点需要处理,且图2(e)、(f)情况也需要判断出时间性,因此对于 section 时间段内均匀出现的高频词串可做二次处理,计算方法如式(2):

$$T(w) = \max(T_{01}(w), T_1(w)) \quad (2)$$

其中: $T_1(w)$ 是词串 w 在 section 的时间特征值; $T_{01}(w)$ 是词串 w 在上一个 section 后半段时间和当前 section 的前半段时间内的时间特征值。

对于图2(d)的情况,可将频繁出现且未判定为热点词的新词在下一个 section 处理。

2.2 多元TPMI

文献[14]给出的 PMI^k 是二元的互信息计算公式,如式(3)。

$$\text{PMI}^k(x, y) = \log \frac{p^k(x, y)}{p^k(x)p^k(y)} \quad (3)$$

其中: $p^k(x)$ 和 $p^k(y)$ 分别表示词串 x 和 y 的概率的 k 次幂; $p^k(x, y)$ 表示字串 x 和 y 的联合概率的 k 次幂。当 $k=1$ 时, PMI^k 即 PMI 。本文采用的是 PMI 。

本文候选新词至少由2个词组成,需要用多元 PMI 计算相关度。因此,需要扩展二元 PMI 为多元 PMI 。从式(3)可以看出, PMI 计算两个词的结合度,其实是计算两个词属于某个词组成部分的程度,并不能确定一个完整的词。如“尸位素餐”这个词,在现代文中计算“位”“素餐”的 PMI 可以发现其结合度很高,这说明“位素餐”很可能是一个词的一部分。因此本文设计了一种扩展 PMI 方式。对于词串 $w_1w_2 \cdots w_{n-1}w_n$ (记为 w),首先寻找其中 PMI 最高的相邻两个词,并认为最大可能成为某个词一部分,所以将两个词合为一个词,然后继续如此扩展。选择最后一次 PMI 值并乘时间特征系值形成该词的 TPMI 值。

对于词串 w , TPMI 计算方法如下:

- 1)选择出现 PMI 最高的两个连续字串 w_kw_{k+1} 为初始种子, $\text{seed} = w_kw_{k+1}$, $k = k - 1$, $j = k + 2$, w_kw_{k+1} 的 PMI 值为 tpmi ;
- 2)while($k > 1$ OR $j < n$)
 - if($k \geq 1$)计算 w_k 和 seed 的 PMI 值 pmil ; else $\text{pmil} = 0$;
 - if($j \leq n$)计算 w_j 和 seed 的 PMI 值 pmir ; else $\text{pmir} = 0$;
 - if($\text{pmil} > \text{pmir}$) { $\text{seed} = w_k\text{seed}$, $k = k - 1$, $\text{tpmi} = \text{pmil}$ };

else { $\text{seed} = \text{seed}w_j$, $j = j + 1$; $\text{tpmi} = \text{pmir}$ }; }

3)加入时间特征系数, w 的 $\text{TPMI} = T(w) \times \text{tpmi}$ 。

2.3 基于TPMI和邻接熵的新词判定

本文判定候选词是否为新词过程:首先计算改进 TPMI ,若 TPMI 大于一定阈值,再计算该词的左右边界的邻接熵^[15];若左右边界熵在一定阈值,则判定该词为一个完整词,为热点新闻。左邻接熵的计算如式(4),右邻接熵的计算如式(5):

$$H_L(w) = - \sum_{x_i \in CL} p(x_i|w) \log p(x_i|w) \quad (4)$$

$$H_R(w) = - \sum_{x_j \in CR} p(x_j|w) \log p(x_j|w) \quad (5)$$

其中: CL 、 CR 分别是候选词 w 的左、右邻接词的集合; $p(x_i|w)$ 是候选词 w 的左邻接词概率, $p(x_j|w)$ 是候选词 w 的右邻接词概率。候选新词邻接熵越大,邻接字词不确定性越大,成为新词边界可能性越大。

3 实验与结果分析

3.1 测试数据集

为了验证本文算法的正确性和有效性,本文采集网络新闻作为测试数据集进行验证。新闻热点词往往是短期内出现比较集中的词,而热点新闻短期内爆发,因此无需采集长时间的数据集进行测试。2018年12月新浪国内新闻中各种热度的新闻事件较多,本文采用2018年12月的新浪国内新闻作为测试集,共采集新闻7 222条。通过人工处理,发现数据集中包含热度非常高的新闻热点1个,一般热度新闻事件16个,以及热度低的新闻事件16个,具体新闻事件如表1所示。

新闻热度是该事件新闻被关注的情况,网络新闻上可以根据该事件新闻出现量判定其热度,不同需求设定不同。为了研究本文方法,设置比较低热度值,将在半月内出现30次以上新闻事件设为高热度新闻(平均每天出现2次及以上),半月内出现15~29次的为一般热度(平均每天1~2次)。低热度的新闻是否算作新闻热点需要根据实际需要决定,可能算热点也可能不算热点,本文低热点新闻为半月内相关新闻8~14条的新闻事件。此外,是否为新词和分词软件有关,本文采用gooseeker的分词工具对数据集进行分词。

本文实验使用软硬件环境为:处理器为Intel Core i7-8750H CPU @ 2.20 GHz 2.21 GHz,内存大小为16 GB,所用软件为Microsoft Visual C++ 2015。

3.2 采用改进FP-tree获得热点新闻的候选集

采用不同的参数可采集不同程度热点新闻的热点新闻。在本文实验中,采用最小频繁计数为8时,可基本采集所有程度热点的新词。若采用频繁计数16,则低热度新闻的新词大多无法采集。采集所有程度热点的新词,识别更困难,本文处理包括低热度(最小频繁计数为8)的热点新闻。

利用本文改进的FP-tree算法,获得频繁词串作为热点词的候选集,结果如表1所示。

3.3 热点新闻识别结果

为了验证本文时间特征值的作用,实验中先采用不加时间特征的多元 PMI 和边界熵进行热点新闻识别(简称多元 PMI 实验),再采用融入时间特征的 TPMI 和左右信息熵获得新闻热点新闻(简称 TPMI 实验),并对两个实验数据进行分



析。本文实验时间特征计算中,section 选择为一个月;从出现次数最多的前 200 个多字词中随机抽样 50,计算平均多元 PMI 值作为多元 PMI 的阈值和 TPMI 的阈值,计算平均边界熵作为边界熵的阈值。通过观察大部分新闻热点爆发、持续情况和考虑处理数据量,建议 section 小于等于一个月且大于等于 2 个星期。

表 1 热点词的候选集
Tab. 1 Candidates for hot words

热度	热点事件	找到的候选词
高热	孟晚舟事件	孟晚舟 孟晚舟事件 获保释 孟晚舟保释后 孟晚舟保释 孟晚舟获保释
	华为新闻	华为 5G 5G 华为董事长
	中美贸易问题	中美贸易 贸易战
	非洲猪瘟	非洲猪瘟疫区 非洲猪瘟疫区解除封锁 非洲猪瘟 非洲猪瘟疫情
	嫦娥四号奔月	嫦娥四号
	国家公祭日	国家公祭日 公祭日 南京大屠杀 大屠杀 公祭
	ofo 破产退押金	ofo 退押金
	四川宜宾兴文 5.7 级地震	5.7 级地震 四川兴文 四川宜宾 震源深度 四川兴文地震 四川宜宾 四川兴文 级地震 兴文 兴文地震
	高通诉苹果禁售令	高通 禁售 禁售令 iPhone
	刘强东事件	刘强东
	权健事件	权健 权健事件
	个人所得税变革	个税 个人所得税 专项附加
	马英九出书	马英九
	公交劫持案	
	二月河去世	二月河
一般	江丙坤生病逝世	江丙坤
	张首晟去世	张首晟
	加拿大鹅	加拿大鹅
	刘国梁任乒协主席	刘国梁 乒协主席
	越南旅行团台湾脱团	脱团
	千亿矿权案	千亿矿权案
	章莹颖案	章莹颖 章莹颖案
	杀妻骗保案	杀妻 骗保 杀妻骗保
	张家口爆燃事故	爆燃事故 张家口爆燃事故
	四川叙永山体滑坡	山体滑坡 四川叙永山体滑坡 叙永山体滑坡
	山东菏泽取消楼市限售政策	限售 山东菏泽
	芬太尼问题	芬太尼
	印尼海啸	印尼海啸
	浙江取消高考英语加权赋分	加权赋分
	烧散煤“被拘”	烧散煤 被拘
低热	中央经济工作会议召开	中央经济工作会议
	考研热	考研
	加公民在华被拘	加公民 康明凯
	其他统计出的候选词	外媒 外媒关注 美媒 最高法 发改委 环球时报社评 十大 十大新闻 被拘 市人 原副主任 市委常委 委常委 预 警 一带一路 改革开放 委副书记 国台办回应 退役军人…(184 个)

分析实验数据发现:非洲猪瘟相关报道是在 8 月份开始频繁出现,之后一直不断,所以 TPMI 实验未能找回“非洲猪瘟”“非洲猪瘟疫情”两个词,但比较集中的有些地区宣布解除疫情,因此获得了“非洲猪瘟疫区解除封锁”这个词。如果用于持续检测新闻热点,连续一个时段、一个时段采集,那么上述热点新词均可在这些新闻爆发时识别出。此外,通过分析多元 PMI 实验识别的新词发现有 2 个词为从 2018 年 11 月下旬开始变为频繁,但在 12 月属于低热度且无法判别时间特征的词。本文算法进行时间特征二次计算时只考虑了高频词,无法识别这两个低热度词的时间特征,造成 a 三种取值的 TPMI 实验中均未曾识别。

实验中, a 取值不宜选择过大,否则会造成持续热度的词更难识别; a 取值也不宜过小,否则造成时间特征弱化。本文测试了 $a=1/2, 2, 8$ 的情况,除了均无法识别上述两个低热点词外:

1) 当 $a=1/2$ 时,正确识别热点新词 51 个,丢失识别“非洲

猪瘟”“非洲猪瘟疫情”2 个热点词(与 $a=2$ 相同),错误识别 5 个,将非热点词“环球时报社评”“涉黑”“九二共识”“红通人员”“加媒”5 个词错误识别成热点新词。错误识别的这 5 个词出现比较零散,每个词对应多个事件,不是热点词。 $a=1/2$ 时,出现比较均匀的词的时间特征为 1,时间特征影响小。

2) 当 $a=2$ 时,正确识别热点新词 51 个,丢失识别“非洲猪瘟”“非洲猪瘟疫情”2 个热点词。

3) 当 $a=8$ 时,正确识别热点新词 50 个词,丢失识别“非洲猪瘟”“非洲猪瘟疫情”“经贸磋商”3 个热点词。“经贸磋商”属于这段时间内时间特征相对弱的词,强化时间特征造成了未识别出该词。

$a=2$ 时的 TPMI 实验结果和多元 PMI 实验结果如表 2 所示。

多元 PMI 实验明显比 TPMI 实验多找到“震源深度、级地



震”和其他统计词 97 个,这些词都不是热点新闻,震源深度存在于所有地震新闻中,不属于哪个地震相关报道,其他词不是错误就明显是一些常用词组。而 TPMI 实验识别出“经贸磋商、中美元首”是因为孟晚舟事件后,中美关系和中美贸易的相关新闻不断出现造成的,而“二手房”是因为有一小段时间

各地楼市信息提到二手房问题,没有将二手房新闻列入低热度新闻,这其实也可以算一个低热度新闻点。人工标注时并未发现,实验后发现关于退役军人的新闻也在短期少量出现,也可算关于退役军人的低热度新闻,因此这四个词的识别不能认为是错误识别。

表 2 多元 PMI 和 TPMI 的实验结果
Tab. 2 Experimental results of multivariant PMI and TPMI

热点事件	多元 PMI 实验识别的新词	TPMI 实验识别的新词
孟晚舟事件	孟晚舟	孟晚舟
华为新闻	5G 华为董事长	5G 华为董事长
中美贸易问题	中美贸易	中美贸易
非洲猪瘟	非洲猪瘟疫区解除封锁 非洲猪瘟 非洲猪瘟疫情	非洲猪瘟疫区解除封锁
嫦娥四号奔月	嫦娥四号	嫦娥四号
国家公祭日	国家公祭日 南京大屠杀	国家公祭日 南京大屠杀
ofo 破产退押金	ofo 退押金	ofo 退押金
四川宜宾兴文 5.7 级地震	5.7 级地震 震源深度 四川兴文地震 四川宜宾 四川兴文	5.7 级地震 四川兴文地震 四川宜宾 四川兴文
高通诉苹果禁售令	高通 禁售 禁售令 iPhone	高通 禁售 禁售令 iPhone
刘强东事件	刘强东	刘强东
权健事件	权健	权健
个人所得税变革	个税 专项附加	个税 专项附加
马英九出书	马英九	马英九
公交劫持案		
二月河去世	二月河	二月河
张首晟去世	张首晟	张首晟
江丙坤生病逝世	江丙坤	江丙坤
加拿大鹅	加拿大鹅	加拿大鹅
刘国梁任乒协主席	刘国梁 乒协主席	刘国梁 乒协主席
越南旅行团台湾脱团	脱团	脱团
千亿矿权案	千亿矿权案	千亿矿权案
章莹颖案	章莹颖案	章莹颖案
杀妻骗保案	杀妻骗保	杀妻骗保
张家口爆燃事故	爆燃事故 张家口爆燃事故	爆燃事故 张家口爆燃事故
四川叙永山体滑坡	山体滑坡 四川叙永山体滑坡	山体滑坡 四川叙永山体滑坡
山东菏泽取消楼市限售	限售 山东菏泽 菏泽	限售 山东菏泽 菏泽
芬太尼问题	芬太尼	芬太尼
印尼海啸	印尼海啸	印尼海啸
浙江取消高考英语加权赋分	加权赋分	加权赋分
烧散煤“被拘”	烧散煤	烧散煤
考研新闻	考研	考研
中央经济工作会议召开	中央经济工作会议	中央经济工作会议
加拿大公民在华被拘	加公民 康明凯	加公民 康明凯
其他统计出的候选词	外媒 外媒关注 美媒 最高法 发改委 环球时报社评市委常委 一带一路 委副书记 经贸磋商 中美元首 退役军人…(101 个)	经贸磋商 中美元首 二手房 退役军人

从以上分析可以看出:引入时间特征值后,可以将一些常用词组合去掉,TPMI 明显大大提高了热点新闻的正确识别率。

可以看出,本文的新词识别是应用于热点新闻识别当中,所以有些词不是真正意义上的词。例如,对应新闻“四川宜宾兴文 5.7 级地震”识别出的“5.7 级地震”“四川兴文地震”“四川宜宾”“四川兴文 5.7 级地震”,每个词都代表了一个地点或一个事件,当然它们也是相关的,不是传统意义的词,但在新闻识别中若拆开会影响识别效果。

本文采集的数据中,没有“”和《》分割的热点词。

本文实验中的 TPMI 和边界熵的参数是根据最频繁出现的词进行计算平均值得到的。这造成了强化时间特征,时间特征强的热点词容易识别,而长时间热度词由于时间特征被

弱化反而易丢失。因此无论 α 还是 TPMI、边界熵的阈值都应该研究更合理的选择方案。

4 结语

本文通过分析新闻热点词特征提出了一种用于网络热点识别的热点新闻发现方法。本文利用非频繁 1、2-词串删除和切分新闻标题来删除大量无用信息;设计融入时间特征的 TPMI 使得热点新闻识别率大幅度提升。

本文方法适用于网络热点新闻的新词发现。而网络热点不仅仅涉及新闻,还包括微博类开放社交媒体。这些平台所发布内容的标题不够正规或无标题,甚至有时候为了吸引网民注意力故意歪曲标题,本文方法还需考虑提炼发布内容和标题的基础上进行改进才可应用。此外,单独处理某段新闻



则会出现之前已经成为热点词且热度始终持续的新词无法识别的问题,可考虑用聚类的方法解决此问题。

参考文献 (References)

- [1] 张华平,商建云. 面向社会媒体的开放领域新词发现[J]. 中文信息学报, 2017, 31(3): 55-61. (ZHANG H P, SHANG J Y. Social media-oriented open domain new word detection [J]. Journal of Chinese Information Processing, 2017, 31(3): 55-61.)
- [2] 杜丽萍,李晓戈,于根,等. 基于互信息改进算法的新词发现对中文分词系统改进[J]. 北京大学学报(自然科学版), 2016, 52(1): 35-40. (DU L P, LI X G, YU G, et al. New word detection based on an improved PMI algorithm for enhancing segmentation system [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.)
- [3] 禾荣朋,许国艳,宋健. 基于改进互信息和邻接熵的微博新词发现方法[J]. 计算机应用, 2016, 36(10): 2772-2776. (YAO R P, XU G Y, SONG J. Micro-blog new word discovery method based on improved mutual information and branch entropy [J]. Journal of Computer Applications, 2016, 36(10): 2772-2776.)
- [4] LI W, GUO K, SHI Y, et al. DWWP: domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain [J]. Knowledge-Based Systems, 2018, 146: 203-214.
- [5] 刘伟童,刘培玉,刘文锋,等. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究, 2019, 36(5): 1293-1296. (LIU W T, LIU P Y, LIU W F, et al. New word discovery algorithm based on mutual information and branch entropy [J]. Application Research of Computers, 2019, 36(5): 1293-1296.)
- [6] 张婧,黄锬宇,梁晨,等. 面向中文社交媒体语料的无监督新词识别研究[J]. 中文信息学报, 2018, 32(3): 17-25, 33. (ZHANG J, HUANG K Y, LIANG C, et al. Unsupervised new word extraction from Chinese social media data [J]. Journal of Chinese Information Processing, 2018, 32(3): 17-25, 33.)
- [7] ZHANG S, ZHU H, XU Z. The extraction method of new logging word/term for social media based on statistics and N-increment [EB/OL]. [2020-03-20]. https://www.onacademic.com/detail/journal_1000040155947110_203b.html.
- [8] 韩彦昭,乔亚男,范亚平,等. 基于条件随机场模型和文本纠错的微博新词词性识别研究[J]. 南京大学学报(自然科学), 2016, 52(2): 353-360. (HAN Y Z, QIAO Y N, FAN Y P, et al. Part-of-speech tagging of microblog unknown words based on conditional random fields and error correction [J]. Journal of Nanjing University (Natural Sciences), 2016, 52(2): 353-360.)
- [9] 李少峰. 面向食品安全的新词发现和热词排行方法的研究与应用[D]. 广州: 中山大学, 2015: 15-26. (LI S F. Research and application on new word discovery and hot word ranking for food security [D]. Guangzhou: Sun Yat-sen University, 2015: 15-26.)
- [10] 张长. 金融知识自动问答中的新词发现及答案排序方法[D]. 哈尔滨: 哈尔滨工业大学, 2017: 16-26. (ZHANG C. The method of new words discovery and answers ranking in finance question answering [D]. Harbin: Harbin Institute of Technology, 2017: 16-26.)
- [11] 刘昱彤,吴斌,谢韬,等. 基于古汉语语料的新词发现方法[J]. 中文信息学报, 2019, 33(1): 46-55. (LIU Y T, WU B, XIE T, et al. New word detection in ancient Chinese corpus [J]. Journal of Chinese Information Processing, 2019, 33(1): 46-55.)
- [12] 王馨,王煜,王亮. 基于新词发现的网络新闻热点排名[J]. 图书情报工作, 2015, 59(6): 68-74. (WANG X, WANG Y, WANG L. Hot news ranking of network news based on new words detection [J]. Library and Information Service, 2015, 59(6): 68-74.)
- [13] PECINA P, SCHLESINGER P. Combining association measures for collocation extraction [C]// Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Stroudsburg: ACL, 2006: 651-658.
- [14] BOUMA G. Normalized (pointwise) mutual information in collocation extraction [C]// Proceedings of the 2009 International Conference of the German Society for Computational Linguistics and Language Technology. Berlin: Springer, 2009: 31-40.
- [15] HUANG J H, POWERS D. Chinese word segmentation based on contextual entropy [C]// Proceedings of the 2003 17th Pacific Asia Conference on Language, Information and Computation. Piscataway: IEEE, 2003: 152-158.

This work is partially supported by the National Social Science Foundation of China (17FTQ002), the Social Science Foundation of Hebei Province (HB15SH064).

WANG Yu, born in 1971, Ph. D., professor. Her research interests include text mining, information retrieval.

XU Jianmin, born in 1966, Ph. D., professor. His research interests include personalized information retrieval, Web community discovery, topic identification and tracking, social network modeling.