

基于主题模型的微博话题检测算法

黄华军, 谭骏珊, 秦姣华

(中南林业科技大学计算机与信息工程学院, 湖南 长沙 410004)

摘要: 微博数据的实时、大规模、短文本以及富含噪声等特征为话题检测带来新的挑战, 传统向量空模型(VSM)表示文本无法很好地对其进行建模。基于此, 提出一种基于主题模型的微博话题检测算法。首先, 对微博数据构建文档词条矩阵和词语关联矩阵来提取主题词; 然后, 对主题词进行聚类, 得到主题模型; 最后, 利用文本与主题模型相互匹配实现文本聚类, 从而达到话题检测的目的。实验结果表明, 该算法能有效地进行话题聚类并检测出话题, 在最佳参数组合条件下, 其各类别的平均 F 值达到 95% 以上。

关键词: 话题检测; 主题模型; 文档词条矩阵; 词语关联矩阵

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00049

Micro-blog topic detection algorithm based on topic model

HUANG Hua-jun, TAN Jun-shan, QIN Jiao-hua

(College of Computer and Information Engineering, Central South University of Forestry & Technology, Changsha 410004, China)

Abstract: Micro-blog data has the characteristic of real-time, volume, short-text, and noise-rich. So it is a challenge for the traditional topic detection technology. A novel micro-blog topic detection algorithm based on topic model was proposed. Firstly, the micro-blog data was expressed as text word matrix and word relation matrix. The topic word was extracted from the two vectors. Secondly, the topic model was obtained with clustering. Finally, the topic detection of micro-blog was obtained by clustering text and topic model. Experimental results show that the algorithm proposed can effectively detection the text topic, and with the best parameter group of precision, recall rate, F , and the value F is about 95%.

Key words: topic detection, topic model, text word matrix, word relation matrix

1 引言

在 Web 2.0 时代, 以微博为典型代表的社交网络应用取代传统媒介, 占据了信息传播的主导位置。微博门槛低、易使用、方便快捷等特点, 吸引一大批网民用户, 使其产生的网络在线数据

呈爆炸性趋势增长^[1]。一条信息通过微博平台能够在短时间传播并影响到数百万的用户。相比传统媒体, 微博在信息传播过程中的传播时效与传播广度都大大的增强。与此同时, 一些虚假信息通过社交网络平台的传播也能在短时间造成社会恐慌、用户财产损失等问题。社会上许多突发性

收稿日期: 2016-04-13; 修回日期: 2016-05-06。通信作者: 黄华军, hhj0906@163.com

基金项目: 国家自然科学基金资助项目 (No.61304208); 湖南省自然科学基金资助项目 (No.13JJ2031); 中南林业科技大学青年科学研究基金资助项目 (No.QJ2012009A)

Foundation Items: The National Natural Science Foundation of China (No.61304208), The Natural Science Foundation of Hunan Province (No.13JJ2031), Youth Scientific Research Foundation of Central South University of Forestry & Technology (No.QJ2012009A)

话题，往往在微博应用上首发，凭借好友转发机制迅速传播，引起社会广泛共鸣，产生巨大的社会影响。因此，微博话题检测技术，对于社会热点话题发现、网民意见感知、舆情检测、应急处理具有积极意义^[2]。

微博话题以 140 字以内的文本信息、图片、影音等多媒体内容，展示个人最新动态，话题分享。这些数据具有时效性、稀疏性、奇异性和冗余性等特点^[3,4]。同时，这些数据信息的表现形式也发生了变化且凌乱无序，许多有价值的信息都被湮没在大量冗余信息中，从而导致提取和管理有用信息越来越困难。对微博话题检测，不仅可以过滤无效信息、提高内容质量、改善用户体验，更能起到监测、舆情控制、观点挖掘的作用。

话题检测是一种信息挖掘技术，从大量数据中挖掘出相关信息，为后续研究提供基础^[5]。传统方法对文本进行话题检测的一般思路是：主题词提取、文本表示、文本相似度计算、文本聚类、话题检测。1) 提取主题词。话题是由主题词来体现的，从文本中提取能表达主题的关键词集合到一起，能很好地表达出文本所描述的话题。其中，最常用的是 TF-IDF 权重计算算法^[6]；通过建立词库来进行主题词提取也是一种比较常见的方法^[7]。2) 文本表示。目前，文本表示大都采用向量空间模型，通过主题词来组成向量表示文本，进而构成一个向量空间^[8]。3) 文本相似度计算是度量文本间距离的一种有效途径，用向量空间模型表示文本后，一般采用余弦定理的方法来计算其相似度，也有不少学者采用其他方法来进行计算^[9,10]。4) 文本聚类是文本信息处理方面常用的一种方法，已比较成熟，其中，比较常用的有 k -means^[11]、层次聚类^[12]等。5) 话题检测的实质是文本聚类，也有一些话题模型，如最常用的 LDA 主题模型^[13]。

传统向量空间模型忽视了中文的“同义”、“多义”及高维稀疏向量问题，因而，在微博话题检测过程中准确率和速度不尽人意。其次，传统的话题检测技术基本都是针对新闻报道进行研究的，这些新闻报道的文本格式比较规范，篇幅较为统一，与现在的微博数据特征差异很大，使传统的文本处理方法已经不再适合对微博数据进行处理。基于此，本文提出一种基于主题模型的文

本聚类方法来进行话题检测，通过提取主题词来进行文本聚类，可以有效解决微博数据的话题检测。

2 微博文本预处理

微博文本预处理主要包括无效微博过滤、文本提纯、中文分词并标注词性、过滤停用词等。具体步骤流程如图 1 所示。

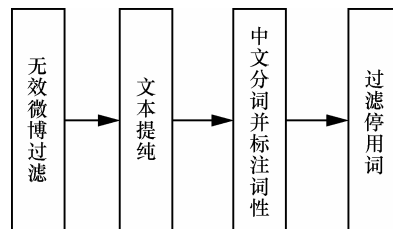


图 1 文本预处理流程

无效微博过滤：在文献[2]中提到，微博中存在一些无效微博，即没有任何话题性、主题性的微博文本，此类微博文本通常由很少的字数组成，基本是由各种表情、符号或链接组成，如“***很百搭 <http://weibo.com/5081446387>”之类的微博。

文本提纯：微博文本中存在大量与主题无关的表情、符号、链接、图片等，如新浪微博中@用户名提到用户，并通知对方，//@用户名则表示转发某用户微博标记，这些符号都与微博内容无关，称这些为噪声数据，在文本处理中，这些噪声数据是与实验处理无关的数据，因此，需要去掉噪声数据，对文本进行提纯。

中文分词：由于中文文本处理的基本单位是词语，且中文文本并不像英文中单词之间都会用空格分开，中文文本中只有句子、单位之间才会用标点符号将其分开，因此，首先需要对中文文本进行分词操作。中科院 ICTCLAS 工具对文本进行分词操作，并标注词性。

过滤停用词：人类语言中通常包含很多功能词，与其他词相比，这些功能词没有特别实际含义，如汉语语言中的“的”、“啊”、“呀”等词。这些词通常用于连接句子成分或表达说话者的感情、强调语调等用途，因此称为停用词。停用词的存在不仅降低实验处理结果的正确率，且由于停用词数量众多，导致文本矩阵存在严重的高维

2) 文档词条矩阵

文档词条矩阵的行与列同样分别表示文本与词语，但是，该矩阵中的词语由经过提取后的关键词构成，且每一个元素的取值为 0 或 1，当某元素所在列对应的词在所在行对应的文本中出现时，则该元素取值为 1；否则为 0，其含义表示的是文档中词语出现的情况。其定义与形式化的表格描述如下。

定义 2 设 $KT = \{kt_1, kt_2, kt_3, \dots, kt_m\}$ (m 表示关键词个数) 表示提取的关键词集合，矩阵 DTM_{mn} 为一个 n 行 m 列的空矩阵，则当 DTM 满足条件(1)时

$$DTM_{ij} = \begin{cases} 1, & kt_i \in d_j \\ 0, & kt_i \notin d_j \end{cases} \quad (1)$$

称 DTM_{mn} 为一个文档词条矩阵。

3) 词语关联矩阵

从上述提到的文档词条矩阵分析可知，从文档词条矩阵的列，即词语的角度考虑，对于每一个出现过该词语的文本，必定会出现一些其他关键词，这些关键词都与该词共同出现过，则表示这些词语具有相关性，称其相互关联。当 2 个词语在文档中共同出现的次数多时，表示其具有强关联性。由此可知，此处的词语关联矩阵的行和列都表示词语，那么，矩阵中每一个元素的值则表示该元素所在行对应的词语与所在列对应的词语同时出现在文档中的次数。其定义与形式化的表格描述如下。

定义 3 设矩阵 TRM_{mm} 为一个 m 行 m 列的零方阵，则当 TRM 满足以下条件时

$$TRM_{ij} = \begin{cases} TRM_{ij} + 1, & kt_i \in d_k, kt_j \in d_k \\ TRM_{ij} + 0, & \text{其他} \end{cases} \quad (2)$$

称 TRM_{mm} 为词语关联矩阵。

4) 主题词聚类

通过对话题微博内容分析可知，话题类微博包含的字数可多可少，这些微博中可能只包含一个或包含多个与话题相关的关键词；而对于类别相近的一些话题，可能存在一个关键词同时与 2 个话题的相关。包含多个关键词的话题微博可能具有两面性：一方面这些关键词更具有类别代表性；另一方面，这些关键词中存在与多个话题相关的词，导致

话题类别不够明确。由此可知，对于同一类的微博话题，其文本中出现的词基本都是关联性强的词语；而对于不同类别的微博话题，文本中出现的词关联程度则会相当小。于是，可以通过提取这些关联性强的关键词，然后，对这些词进行聚类分析，就可以得到每一个类别对应的关键词，此时，称这些关键词为主题词，定义如下。

定义 4 设 $TP = \{tp_1, tp_2, \dots, tp_n\}$ 表示话题集合。 n 表示话题数， tp_i 表示属于第 i 个话题的主题词的集合，则当 kt_i 仅属于 tp_i 时，称该 kt_i 为主题词。

通过对话题主题词进行分析，可得以下 3 点性质。

① 当一个话题出现多个主题词时，则这些主题词共同出现的次数必定比跟其他词出现的次数要高。

② 当一个话题中只有一个主题词时，则其单独出现的次数比与其他任何词共同出现的次数要高出很多。

③ 当一个主题词同时出现在多个内容相近的话题中时，该词与这些话题中其他有区别性的主题词共现次数均多，则将此主题词定为无效主题词，不具备区分效果，将其除去，不作考虑。

对于无效主题词，设定一个阈值参数 α ，用来判断一个主题词是否为无效主题词。

定义 5 当一个主题词和关联性最强的主题词所关联的次数除以与该主题词有关联的其他所有主题词数量之和的值不超过 α 时，则定义该主题词为无效主题词。

对于每一个主题词，与其关联性强的主题词基本是属于同一话题，再设定阈值参数 β ，对于每一个主题词，分别选取与其关联关系最强的前 β 个主题词进行分析。对每一个主题词之间的强关联词集合进行交集运算，若交集不为空，则判断这 2 个主题词表示为同一话题；否则，表示不同话题。由此，实现对主题词的聚类操作，整个模型构建的具体流程如图 4 所示。

其中， $TRM_{i\cdot}$ 表示词语关联矩阵的第 i 行， $\max(TRM_{i\cdot})$ 则表示 TRM_{mm} 第 i 行中的最大值， $\text{num}(TRM_{i\cdot} \neq 0)$ 表示 TRM_{mm} 第 i 行中不为 0 的值的个数， $TRM_{i\beta}$ 和 $TRM_{j\beta}$ 分别表示排序后的 TRM_{mm} 中的第 i 行和第 j 行的前 β 个元素构成集合，

$TRM_{i\beta} \cap TRM_{j\beta}$ 表示对这 2 个集合取交集运算。通过词语关联矩阵进行主题词聚类的算法如下。

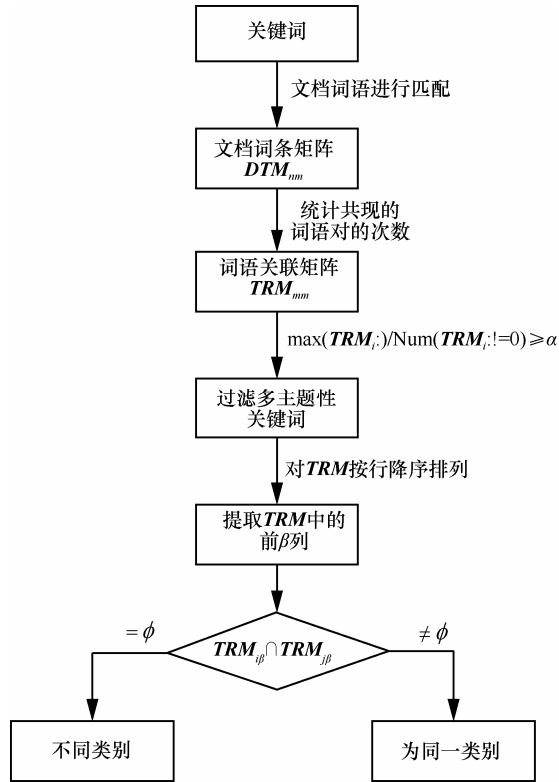


图4 主题模型算法流程

算法1 基于词语关联矩阵的主题词聚类

输入：词语关联矩阵 TRM_{nm}

输出：聚类后的各类主题词集合 TP

1) 对 TRM_{nm} 的每一行进行降序排列；

2) 提取 TRM_{nm} 的前 β 列，得到 $TRM_{m\beta}$ ；

3) While($i \leq m \ \&\& \ j \leq m$) {

 If($TRM_{i\beta} \cap TRM_{j\beta} = \phi$) {

$TRM_{i\beta}$ 和 $TRM_{j\beta}$ 属于不同的类别，

不对其进行操作；

 }else {

$TRM_{i\beta}$ 和 $TRM_{j\beta}$ 属于同一类别，此

时，取 $TRM_{i\beta} = TRM_{i\beta} \cup TRM_{j\beta}$ ， $TRM_{j\beta} = \phi$ ，

即只取其并集再进行同样的操作

 }

}

3.2 话题检测

文本聚类是话题检测的前提，话题检测技术的本质就是无监督的文本聚类，即在无人工干预的情况下，通过判断一个文本是否属于已有话题

来进行归类。通过文本聚类后，找到聚类类别，从而提取话题，达到话题检测结果。由上述可知，主题模型创建后，可以得到各类别话题的主题词，因此，可以根据文本与各类话题的主题词匹配情况来确定各文本所属的话题类别，从而实现话题聚类。假设设定阈值参数 γ ，当文档中出现该类主题词的数量不小于 γ 时，该文档属于该类话题，从而实现话题检测。即通过利用 D 中每一个文档的词语与这些归好类的话题词进行对比匹配，当 $\text{num}(d_i \cap tp_i) \geq \gamma$ 时，定义该文本 d_i 属于 tp_i 这个话题，其中， $\text{num}(d_i \cap tp_i) \geq \gamma$ 表示的是取交集后的个数。

4 实验结果与分析

4.1 实验准备

实验数据取自中国计算机学会信息网科研数据平台，包含 14 个热点话题，共 3 455 条微博文本，各话题内容的数量如表 1 所示。

表1 微博话题及其数量

序号	话题	数量
1	#北京楼顶别墅主人#	386
2	#斯诺登棱镜事件#	131
3	#李天一案件#	356
4	#郑钧刘芸马尔代夫婚礼#	138
5	#姚贝娜 vs 林育群#	143
6	#七夕，一起看英仙座流星雨#	233
7	#光大证券乌龙事件#	480
8	#埃及清场#	153
9	#林丹 vs 李宗伟#	350
10	#台风尤特#	609
11	#泰国白龙王病逝#	160
12	#中国好外婆#	109
13	#昌都地震#	120
14	#新疆吐鲁番地表温度 70℃ 火焰山烤鸡蛋生意红火#	87

实验评估采用信息检测领域最常用的 3 个评价指标：准确率 P (precision)、召回率 R (recall) 和综合评价指标 F 值 (F -value)。准确率是指结果中有多少是准确的；召回率就是指所有准确的结果中有多少被检测出来。 F 是准确率与召回率的综合评价参数，设 M_1 为检索到的正确文档数， M 为所有的正确文档数， N 为所有检索到的文档数。具体计算公式分别如下

$$P = \frac{M_1}{M} \quad (3)$$

$$R = \frac{M_1}{N} \quad (4)$$

$$F = \frac{2PR}{P+R} \quad (5)$$

4.2 结果分析

首先, 读取微博文本数据, 对其进行预处理, 得到有效词 4 987 个, 再提取文本的关键词, 得到 112 个关键词, 从而大大缩减了文本处理的维度问题。再根据关键词构造文档词条矩阵, 此时, 可知文档词条矩阵为一个 3 455 行、112 列的二维矩阵, 并由此可以得到一个 112×112 的词语关联矩阵。从矩阵大小来看, 该算法能有效解决传统方法中存在的高维稀疏等问题。

实验需要确定 3 个参数值: 阈值参数 α , 用来判断一个主题词是否为无效主题词; 阈值参数 β , 对于每一个主题词, 分别选取与其关联关系最强的前 β 个主题词进行聚类; 阈值参数 γ , 当文档中出现该类主题词的数量不小于 γ 时, 对该文档进行归类。通过对实验数据的分析, 可以大致确定这 3 个参数的取值范围分别规定为:

$1 \leq \alpha \leq 10$, $1 \leq \beta \leq 4$, $1 \leq \gamma \leq 10$, 由此可知, 对于每一个类别, 根据每一种参数组合值就可能存在 $10 \times 4 \times 10 = 400$ 种可能, 14 个类别就有 $400 \times 14 = 5\,600$ 种情况, 这些图全部表示在一个图中过于密集, 因此, 本课题选取每一种参数组合条件下所有类别的 F 值的平均值作为参考标准。表 2 给出了部分参数组合条件各类话题的准确率、召回率和 F 值, 其中, 行表示各类话题, 根据表 2 的序号排列, 列分别表示各参数组合条件下各类的准确率、召回率和 F 值。从表中可以看出, 不同参数组合条件下各取值的不同, 当 $\alpha = 5$, $\beta = 2$, $\gamma = 2$ 时结果最好。

图 5 给出了各参数组合条件下所有类别 F 值的平均值曲线。由图 5 可知, 该算法在不同的参数取值条件下, F 平均值的取值变动范围很大, 通过实验数据得出, 当参数取值组合为 $\alpha = 5$, $\beta = 2$, $\gamma = 2$ 时, 各话题类别的 F 值取平均值最好, 其 F 平均值达到 95.8。此时, 各个类别的准确率、召回率与 F 值取值情况如图 6 所示。

表 2 不同参数组合条件下各类话题的准确率、召回率和 F 值

参数话题	$\alpha=1, \beta=1, \gamma=1$			$\alpha=3, \beta=2, \gamma=5$			$\alpha=5, \beta=2, \gamma=2$		
	P	R	F	P	R	F	P	R	F
1	24.55	56.48	34.22	100.00	56.21	71.97	98.97	100.00	99.48
2	11.27	100.00	20.26	95.42	95.41	95.42	99.20	94.66	96.87
3	14.64	36.51	20.90	97.79	87.07	92.12	98.72	87.08	92.54
4	8.78	56.52	15.20	100.00	92.75	96.24	100.00	92.75	96.24
5	35.46	69.93	47.06	97.10	93.71	95.37	99.24	91.61	95.27
6	88.78	37.33	52.57	75.57	99.57	85.92	82.01	97.85	89.24
7	41.22	99.79	58.34	26.42	56.04	35.91	99.35	95.83	97.56
8	6.08	35.29	10.37	84.70	94.12	89.16	96.57	92.16	94.31
9	29.05	73.71	41.68	20.82	60.57	30.99	99.69	94.57	97.07
10	60.23	75.86	67.15	29.37	49.09	36.75	98.22	99.67	98.94
11	9.47	68.75	16.64	95.18	98.75	96.93	95.18	98.75	96.93
12	5.51	44.95	9.82	100.00	52.29	68.67	99.03	94.50	96.71
13	6.12	39.17	10.69	96.46	90.83	93.56	100.00	94.17	97.00
14	6.64	58.62	11.94	100.00	88.51	93.90	100.00	87.36	93.25

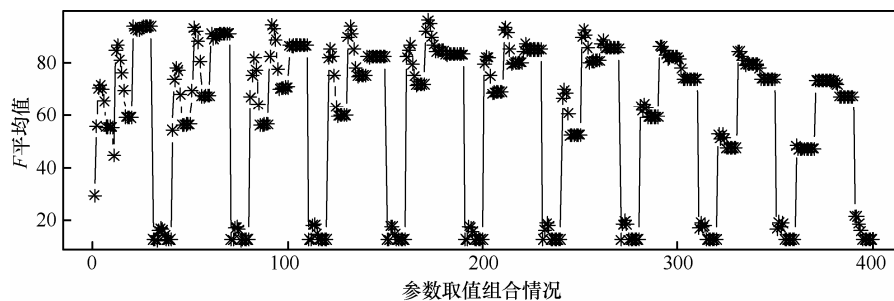


图 5 各参数取值组合条件下 F 平均值取值情况

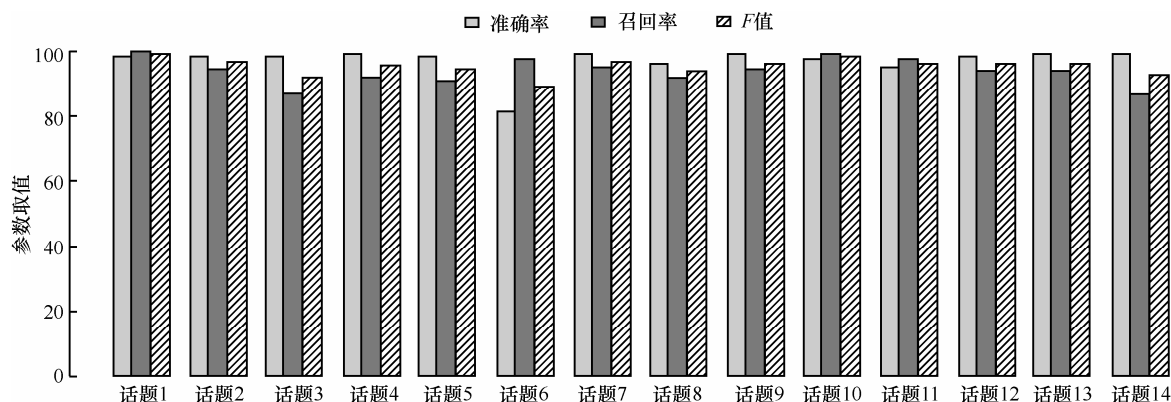


图6 最佳参数取值下各个类别的准确率、召回率与F值取值情况

由上述知，通过关键词提取后，关键词的数量是112个，其按照出现的频次从大到小排列分别为：光大、证券、台风、北京、别墅、尤特、中国、李宗伟、楼顶、林丹、流星雨、强台风、龙王、外婆、事件、李天一、乌龙、婚礼、埃及、登陆、广东、影响、新闻、视频、网友、暴雨、出现、交易、地震、拆除、分享、姚贝娜、中心、记者、美国、书法、声音、表示、昌都、酒吧、期货、斯诺登、假山、曝光、冠军、支持、时间、张必清、马尔代夫、部门、高温、羽毛球、鸡蛋、问题、外孙、泰国、位于、进行、决赛、梦鸽、顶盖、消息、流星、风力、看到、预计、发生、世界、市场、地区、小时、安全、发布、微博、造成、报道、死亡、火焰山、关注、棱镜、郑钧、

对手、系统、教授、西藏、刘芸、广州、暴涨、天气、股市、预警、工作、阳江、林育群、成为、律师、行动、进入、山庄、比赛、广西、海面、地表、人员、媒体、南海、小区、政府、投资、现场、同志、希望。得到最佳参数取值后，主题模型中得到的关于各个话题类别的关键主题词如表3所示。

由表3可知，每一个话题所对应的主题词均能很好地代表此类话题，表明该主题词的主题性非常明确。从上面关键词中看到，“中国”跟“事件”等出现频次很高的一些词没有出现在对应话题的主题词中，通过对这些主题词进行分析得到，这些均属于多话题性主题词，如“中国”一词，在话题#中国好外婆#、#姚贝娜 vs 林育群#、#光

表3

各个话题类别的主题词

话题	主题词
#北京楼顶别墅主人#	北京、别墅、楼顶、拆除、假山、张必清、顶盖、教授、山庄、小区
#斯诺登棱镜事件#	美国、斯诺登、棱镜
#李天一案件#	李天一、书法、酒吧、梦鸽、律师
#郑钧刘芸马尔代夫婚礼#	婚礼、马尔代夫、郑钧、刘芸
#姚贝娜 vs 林育群#	姚贝娜、声音、林育群
#七夕，一起看英仙座流星雨#	流星雨、流星、看到
#光大证券乌龙事件#	光大、证券、乌龙、交易、期货、市场、暴涨、股市
#埃及清场#	埃及、行动
#林丹 vs 李宗伟#	李宗伟、林丹、冠军、羽毛球、决赛、世界、对手、比赛
#台风尤特#	台风、尤特、强台风、登陆、中心
#泰国白龙王病逝#	龙王、泰国
#中国好外婆#	外婆、外孙、同志
#昌都地震#	地震、昌都、西藏
#新疆吐鲁番地表温度70℃火焰山烤鸡蛋生意红火#	鸡蛋、火焰山、地表

大证券乌龙事件#中均有出现,且出现次数不集中;“事件”一词分别在话题#光大证券乌龙事件#、#斯诺登棱镜事件#中出现,且出现次数差不多,此类词明显为多话题性主题词,不具备区分话题类别的价值,将其舍去。

在确定参数及性能后,采用通过欧氏距离计算得到的层次聚类及 k -means 聚类方法对相同实验数据进行对比实验操作,各自的聚类结果分别如图7和图8所示,综合对比结果如图9所示。

由图9可知,主题模型算法针对该实验数据进行操作得到的结果明显优于层次聚类及 k -means 聚类方法。通过对实验数据进行分析可知,在经过预处理后,根据这些短文本创建向量空间模型,再分别利用欧氏距离进行距离计算从而进行层次聚类和 k -means 聚类,由于短文本数据具有很大的稀疏性,使其在利用向量空间模型进行计算过程中存在较大的误差,导致普遍准确率等都偏低。

5 结束语

针对传统话题检测技术处理微博数据的不足,提出了一种基本主题模型的文本聚类方法进行话题检测。以微博数据作为研究对象,详细分析了微博文本的特征,得到主题词在话题性微博文本中的重要性,从而推断可以从主题词的角度进行研究。首先,对预处理后的文本提取出关键词,降低文本表示的维度,减小计算量;然后,由提取的关键词构建文档词条矩阵,分析每一个关键词在文本中出现的情况;其次,由文档词条矩阵分析总结出共同出现的词语之间的关系,统计每一对共同出现的词语及其共现的次数,并以此构造一个词语关联矩阵;再次,根据同一话题中出现的词语关联性最强的特点,对词语关联矩阵中的强关联性词语集合取交集运算,将同一话题的主题词聚为一类,不同话题的主题词则分为不同类,此时主题模型构建完成;最后,利用文

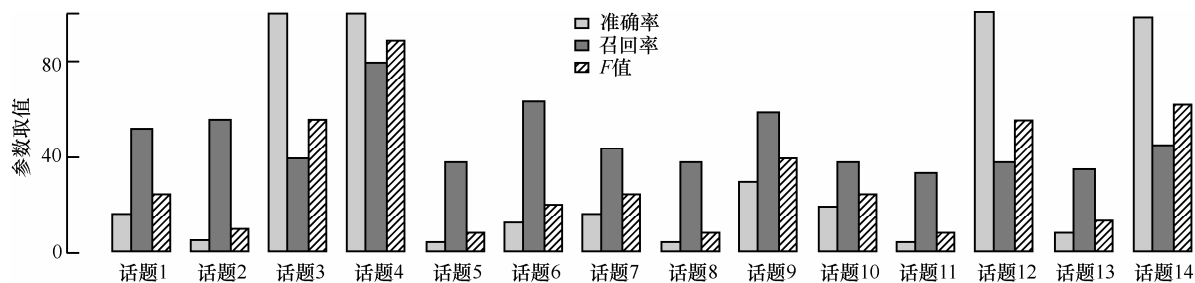


图7 层次聚类结果

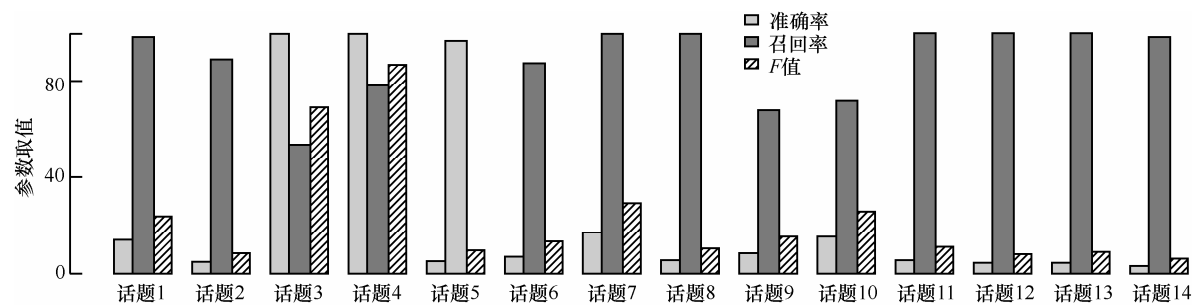


图8 k -means 聚类结果

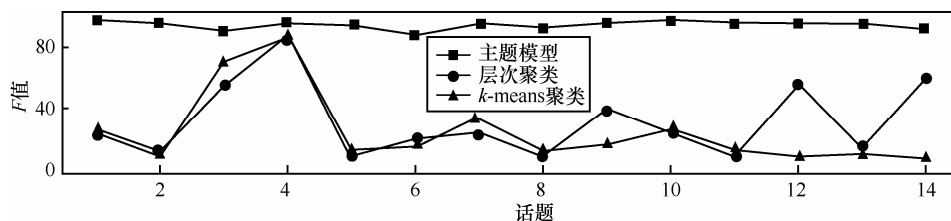


图9 3个实验对比结果

本与主题模型中已经归好类别的主题词进行匹配,即可得到文本的类别,从而实现话题检测。实验结果表明,该方法能取得较好的效果,在最佳参数组合的条件下,其各类别的平均 F 值达到95%以上。由于现在网络技术的发展,有海量的短文本数据存在在线网络上,因此,下一步的研究工作将考虑在大数据环境下,对其进行话题检测技术的研究,并尝试将该方法运用到大数据平台下。

参考文献:

- [1] 王仲远,程健鹏,王海勋,等. 短文本理解研究[J]. 计算机研究与发展, 2016, 53(2):262-269.
WANG Z Y, CHENG J P, WANG H X, et al. Short text understanding: a survey[J]. Journal of Computer Research and Development, 2016, 53(2): 262-269.
- [2] 贺敏,杜攀,张谨,等. 基于动量模型的微博突发话题检测方法[J]. 计算机研究与发展, 2015,52(5):1022-1028.
HE M, DU P, ZHANG J. et al. Microblog bursty topic detection method based on momentum model[J].Journal of Computer Research and Development,2015, 52(5): 1022-1028.
- [3] 刘全超,黄河燕,冯冲. 基于多特征微博话题情感倾向性判定算法研究[J]. 中文信息学报, 2014, 28(4): 123-131.
LIU Q C, HUANG H Y, FENG C.Multi-feature based sentiment orientation identification algorithm for micro-blog topics[J]. Journal of Chinese Information Processing, 2014, 28(4): 123-131.
- [4] 程俊霞,李芝棠,邹明光,等. 基于 SVM 过滤的微博新闻话题检测方法[J]. 通信学报, 2013, 34(Z2): 74-78.
CHEN J X, LI Z T, ZOU M G, et al. Novel topic detection method for microblog based on SVM filtration[J]. Journal of Communications, 2013, 34(Z2): 74-78.
- [5] NIST. The 2003 topic detection and tracking task definition and evaluation plan[EB/OL]. <http://www.nist.gov/speech/tests.tdt/tdt2003/evalplan.html>.
- [6] 周学广,高飞,孙燕. 基于依存连接权 VSM 的子话题检测与跟踪方法[J]. 通信学报, 2013, 34(8): 1-9.
ZHOU X G, GAO F, SUN Y. Sub-topic detection and tracking based on dependency connection weights for vector space model[J]. Journal of Communications, 2013, 34(8):1-9.
- [7] 付艳,周明全,王学松,等. 面向互联网新闻的在线事件检测[J]. 软件学报, 2010, 21(Z): 363-372.
FU Y, ZHOU M Q, WANG X S, et al. On-line event detection from web news stream[J].Journal of Software,2010, 21(Z): 363-372.
- [8] 郝文宁,冯波,陈刚,等. 基于领域本体的文档向量空间模型构建[J]. 计算机应用研究, 2013, 30(3): 764-767.
HAO W N, FENG B, CHEN G, et al. Document vector space model construction based on domain ontology[J].Application Research of Computers, 2013, 30(3): 764-767.
- [9] PHUVIPADAWAT S, MURATA T. breaking news detection and tracking in twitter[C]//2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto. c2010: 120-123.
- [10] YANXIANG H, YE T, QIANG C, et al. Summarizing microblogs on network hot topics[C]//2011 International Conference on Internet Technology and Applications. c2011: 1-4.
- [11] 谢娟英,高红超. 基于统计相关性 k -means的区分因子集选择算法[J]. 软件学报, 2014, 25(9): 2050-2075.
XIE J Y, GAO H C. Statistical correlation and k -means based distinguishable gene subset selection algorithms[J]. Journal of Software, 2014, 25(9): 2050-2075.
- [12] DAI X Y, CHEN Q C, WANG X L, et al. Online topic detection and tracking of financial news based on hierarchical clustering[C]//2010 International Conference on Machine Learning and Cybernetics, Qingdao. c2010: 3341-3346.
- [13] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[M]//Advances in Information Retrieval. Berlin: Heidelberg, 2011: 338-349.

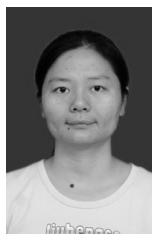
作者简介:



黄华军(1978-),男,湖南宜章人,博士,中南林业科技大学教授、硕士生导师,主要研究方向为网络与信息安全、网络钓鱼防御。



谭骏珊(1963-),男,湖南益阳人,博士,中南林业科技大学教授、博士生导师,主要研究方向为数据库信息与管理、数据挖掘。



秦姣华(1973-),女,湖南益阳人,博士,中南林业科技大学教授、硕士生导师,主要研究方向为网络与信息安全、加密图像检索。