中文微博事件情感分布的原因分析*

李泽魁[1]，赵妍妍[2]，秦兵[1]，刘挺[1]

[1]哈尔滨工业大学计算机科学与技术学院信息检索研究中心，哈尔滨，150001

[2]哈尔滨工业大学机电学院媒体系，哈尔滨，150001

E-mail: {zkli, yyzhao, qinb, tliu}@ir.hit.edu.cn

摘　要：微博作为新兴的社交媒体平台，越来越多的网民选择在微博上获取与分享自己感兴趣的信息。在微博日均千万级的大数据面前，分析网民对某一事件的观点与态度是一件非常有意义的工作。调研中本文发现，大众对于单个事件的不同话题存在不同的情感分布。针对这一现象，本文提出了使用无监督学习的层次聚类排序方法和半监督学习的微博话题纠正算法两种方法，进行事件话题及其相关微博的挖掘。最后利用情感分析的相关技术，达到对相关微博进行情感分布统计及分析的目的。通过在人工构建的数据集上测试，结果表明本文的方法能够准确分析事件情感分布的原因。

关键词：情感原因分析；话题聚类；话题纠正；中文微博

Emotion Causation Analysis for the Hot Event on Microblogs

Zekui Li[1], Yanyan Zhao[2], Bing Qin[1], Ting Liu[1]

[1]Research Center for Social Computing and Information Retrieval of Computer Science and Technology School, Harbin Institute of Technology, Harbin 150001

[2] Department of Media Technology and Art, Harbin Institute of Technology, Harbin 150001

E-mail: {zkli, yyzhao, qinb, tliu}@ir.hit.edu.cn

Abstract：As an emerging social media platform, more and more netizens tend to obtain and share information they are interested in by micro-blog. In front of tens of millions level microblogging data per day, analysis of users' attitudes on an event is a meaningful work. This paper found that there are different emotions distributions when the public talk about different topics of an event. In response to that phenomenon, firstly we proposed a combination algorithm of unsupervised learning method based on hierarchical clustering and a kind of supervised learning algorithm used for topic rectification, so that we can mine the topics under events as well as their micro-blogs. Finally we analyze emotional distribution using some related algorithms about sentiment analysis. Fortunately, the experiment results on artificially constructed data set show that our proposed method can accurately analyze the causations for the emotional event distribution.

Key words：Emotion Causation Analysis; Topic Cluster; Topic Rectification; Chinese Micro-blog

1　前言

微博，作为国内最流行的社交媒体平台之一，存在着数以亿计的活跃用户。伴随着网民在微博上不断获取、传播及分享身边的新鲜事，日均都会产生千万级的微博数据。越来越多的学者开始关注微博这样的大数据背后的信息，而本文所提出的中文微博事件的情感分布原因分析就是面向微博语料的一项新兴的研究课题。

在实际分析微博事件时本文发现，一个微博热点事件的情感分布是随着时间而变化的。如 2015 年 5 月的微博热点事件之一“MERS 入侵广东”事件，它的情感分布折线（从 5 月 22 日至 6 月 4 日）如图 1 所示：

* **基金项目**：本课题受到国家自然科学基金重点项目(61133012)、国家自然科学青年基金项目(61300113)及国家自然科学基金面上项目(61273321) 资助。

作者简介：李泽魁(1992—)，男，硕士，主要研究方向为情感分析。

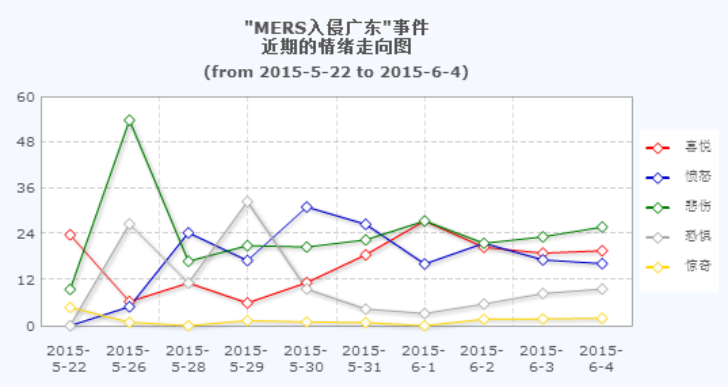

图 1 “MERS 入侵广东”事件的情感分布变化折线图

Fig.1 Emotional Line Chart of Event “MERS Invasion into Guangdong”

在图 1 中，红色、蓝色、绿色、灰色和黄色折线分别代表喜悦、愤怒、悲伤、恐惧和惊奇五种情感变化。可以发现从 5 月 22 日 MERS 病毒在韩国出现首例开始，事件微博开始捕捉，随后的几天中，事件在不断地发酵。图中分别在 5 月 26 日、5 月 29 日、5 月 30 日和 6 月 1 日等时间点网民的情绪分布都有所变化。

那么为何网民随着时间推移，对于上述事件会反映出不同的情感分布呢？通过系统地分析，我们发现在 5 月 26 日，新闻报道“韩国 MERS 病毒携带者在广东确诊”，大家的情绪是恐惧加悲伤；5 月 29 日，微博上传闻“接受隔离治疗的病毒患者病情加重”，网民的情绪转为恐惧居多；5 月 30 日新闻揭露了“韩国医疗部门的监管漏洞及技术乏力”等问题，大家的情绪又转为愤怒居多；6 月 1 日各大网站记录了“惠州医院 ICU 的医生护士们为了不使疫情扩散，抽签上岗”的感人事迹，国内一片鼓励、支持和赞美的呼声瞬间高涨了起来，由之而来的是情绪中喜悦的比例上涨。

从“MERS 病毒入侵广东”事件的分析中我们可以看出，其随着时间推移而变化的情感分布，其实代表的是网民对事件的不同侧面的不同看法。例如事件中“恐惧”的情感分布是针对新闻“病毒患者病情加重”、“愤怒”针对的是话题是“韩国医疗部门监管漏洞”等。本文将事件的不同讨论方面，定义为单个事件的话题。

中文微博事件情感分布的原因分析，本文将这项新任务定义为从实时微博数据中自动挖掘微博热点事件及其话题，通过对话题情感分布的统计，最终对事件的不同情感分布的原因进行分析。

2 相关研究

由于微博作为社交媒体的形式走入我们生活的时间并不长，同时本文的任务也是首次定义，所以国内外面向微博的事件情感分布原因分析的相关研究不是很多。

本文首先调研了现阶段的微博事件挖掘方法。2011 年，Weng 等人通过将小波变换的相关原理利用到微博文本中一些词语频率的监听上，通过分析其自相关性过滤筛选出突发词汇，聚类为突发事件[1]，该方法在事件监测方面有一定效果，但是易受噪声干扰；Zhao 等人在为了对微博中的热点词条进行排序，根据含有关键词条的微博的转发率、词频等信息计算出来一个概率值，根据概率得出基于“有趣程度”的排序公式[2]。Spina 等人列举了现有的文本抽取的抽取方式，通过对少量已标注微博语料进行了话题抽取，最后出乎意料的是最简单的基于词频/逆文档频率的抽取方法取得最好的效果，同时证明了名词过滤的预处理在本任务中是有效的[3]。相比前人比较粗糙的工作，Abhimanyu 和 Anitha 在 2014 年的工作[4]就显得充分很多。他们为了挖掘 Twitter 中的热点话题，通过观察微博事件的共性，得出了三项评价指标，分别为“多样性（Diversity）”、“唯一性（Uniqueness）”和“突发性（Burstiness）”，

用弱标注的训练语料通过一个高斯混合模型来拟合数据的分布，从而输出候选角度是否为微博事件，这样的有监督学习的话题抽取方法也可以取得不错的效果，但是很遗憾这个算法没有涉及话题的聚类排序处理。

总结地说，前人的研究基本上以面向英文微博 Twitter 为主，同时偏向于微博的事件监测。而本文的工作，微博事件情感分布的原因分析，本质上是对事件的不同侧面的话题进行情感分布的分析，与上述研究存在一定的交集。

微博的情感分类多种多样，例如可以按照传统分类方法粗粒度地分为“褒贬”两类，也可以细粒度地分为“喜怒哀恐惊”五类[5]。本文采用后者的分类体系作为情感分类标准。Rosenthal 等人在 SemEval2015（Semantic Evaluation, 语义评测）中使用的一套基于特征抽取的微博情感分类的算法达到了世界最优[6]，无疑在中文微博的情感分类任务中是有借鉴价值的。本文借鉴了 Rosenthal 等人的相关情感分类算法，但是情感分类部分不是本文重点，所以不在此赘述。

3 微博事件情感分布的原因分析算法

由于中文微博的语言特殊性，本文结合本任务针对性地提出了一套微博事件情感分布的原因分析算法，其算法总体流程如图 2：

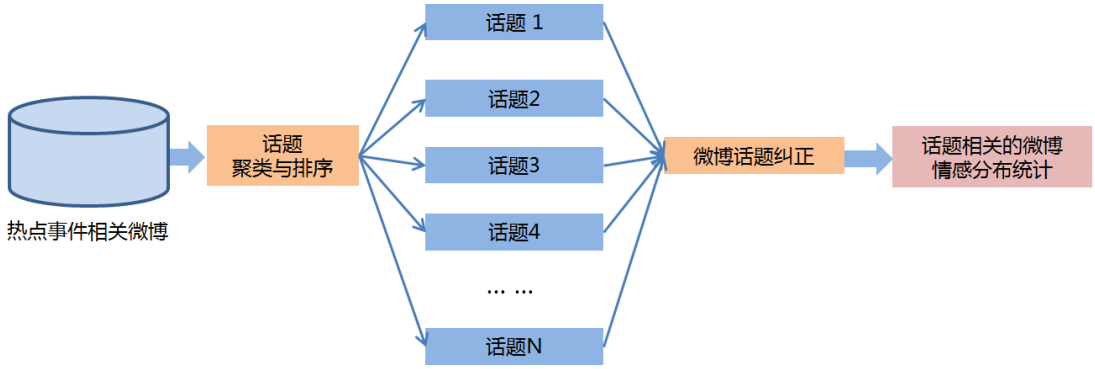

图 2 微博事件情感分布的原因分析算法

Fig.2 Flow Chart of Analysis in Micro-blog Emotional Distribution

图 2 中，第一步输入为单个热点事件的相关微博，为了快速收集微博资源，本文使用了哈工大语言技术平台[7]对文本预处理并通过关键词匹配方法判断微博是否相关；输入相关的微博，我们需要找出具有代表性的话题信息，这里本文使用了 3.1 节中的微博话题层次聚类与排序算法，为下一步提供 N 个话题候选及其对应微博；接着，为了对聚类排序结果进行自动修正，本文在 3.2 节中介绍了基于半监督学习的微博话题纠正算法，对排序靠前的话题的相关微博进行自扩充，使得话题对应的微博分布更趋于真实值；最终利用相关的细粒度情感分类算法得到话题的情感分布。

3.1 微博的关键话题发现

给定热点事件的微博，我们首先需要自动挖掘热点事件微博中的话题信息。实际实验中，通过统计微博中的 Hashtag（即话题信息，微博中两个“#”符号之间的文字）信息，发现热点事件的话题非常多，代表性的比如 2015 年 5 月发生的“庆安枪击事件”，爬取的相关微博中不同的话题就有 343 个，从“事件主人公徐纯合”到“开枪民警李乐斌”，从“支持警察开枪”到“同情受害人”等等方面，角度众多。面对这么多的话题信息，本文要做的是对话题进行聚类排序和针对关键话题及其微博的挖掘。3.1.1 节和 3.1.2 节就针对这两部分问题展开介绍。

3.1.1 话题的聚类排序算法

在得到热点事件的相关微博后，首先需要对话题进行抽取与聚类排序。话题的抽取工作，是指将微博所描述的话题信息进行抽取总结；话题的聚类排序，是指先将部分相似的话题进行聚类处理，例如“携程官网被黑或为人祸”和“携程服务器被黑或为人祸”这两个话题就十分类似，同时将合并后的话题根据其热议程度排序。这部分算法的流程图如图 3：

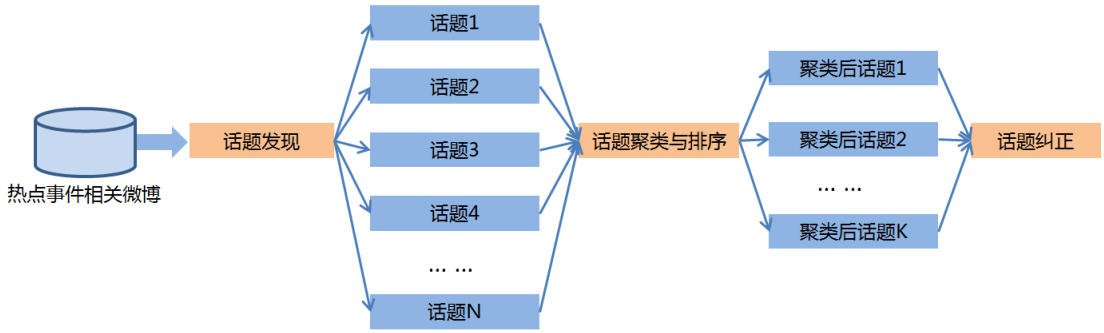

图 3 话题聚类排序算法

Fig.3 Flow Chart of Topic Clustering and Ranking Algorithm

图 3 中，热点事件相关微博作为输入进入话题发现模块，该模块首先通过 Hashtag 匹配等方式将对应的话题标题抽取出来，经过基于微博数目的排序和过滤等策略挖掘 N 个话题；话题聚类模块通过层次聚类（Hierarchical Cluster）算法[8]以及一些过滤规则，得出候选的合并后的 K 个话题，经过排序算法，作为关键话题抽取的输入。

- 层次聚类算法

上述算法中，层次聚类的相似度计算方法是一个关键点，很大程度上决定了聚类结果的好坏。本文通过对比基于微博词频/逆文档频率（TF/IDF，Term Frequency/Inverse Document Frequency）的相似度算法与 Hashtag 字符串相似度算法的实验结果，最终选定了后者，即字符串相似度作为聚类中距离计算的依据。两种相似度计算公式如下：

$$Similarity_{TFIDF}(\mathrm{S_A},\mathrm{S_B}) = cosine(TFIDF(\mathrm{S_A}),\ TFIDF(\mathrm{S_B})) \quad (1)$$

$$Similarity_{Hashtag}(\mathrm{H_A},\mathrm{H_B}) = \frac{Length(LCS(\mathrm{H_A},\mathrm{H_B}))}{min(Length(\mathrm{H_A}),Length(\mathrm{H_B}))} + (1 - \frac{Edit\ Distance(\mathrm{H_A},\mathrm{H_B})}{max(Length(\mathrm{H_A}),Length(\mathrm{H_B}))}) \quad (2)$$

公式 1 中，给定两句微博文本$\mathrm{S_A}$和$\mathrm{S_B}$，基于 TF/IDF 的相似度距离是文本的 TFIDF 向量之间的余弦夹角。公式 2 中，假设两个$\mathrm{S_A}$和$\mathrm{S_B}$中的 Hashtag 字符串分别为$\mathrm{H_A}$和$\mathrm{H_B}$，约定两个字符串的最长公共子序列（LCS， Longest Common String）越长、编辑距离（Edit Distance）越短，他们的相似度越高。为了使公式具有普适性，将两个字符串相似度的数值作了归一化处理，即分别除以了两个字符串的长度的最小最大值。

- 话题聚类结果排序算法

除了聚类算法，从中抽取中关键话题同样尤为重要。本文对比了两种排序算法：简单的根据微博数目排序（公式 3）和根据微博数目与聚类结果话题数的加权关系排序（公式 4），本文选取了考虑更多因素的公式 4 作为排序公式。

$$Ranking\ Score(\mathrm{topic}) = \mathrm{topic_{weibonumber}} \quad (3)$$

$$Ranking\ Score(\mathrm{topic}) = log(\mathrm{topic_{weibonumber}}) \cdot \mathrm{topic_{num}} \quad (4)$$

上述公式中Ranking Score(topic)是话题topic对应的排序得分，$\mathrm{topic_{weibonumber}}$为话题下含有的微博数目。公式 4 在公式 3 的基础上考虑了聚类后单个簇下的话题数目，其中的

$topic_{num}$为结果中合并的话题数目。同时为了使指标具有可比性，对微博数目进行了对数处理。

3.1.2 微博话题纠正算法

按照 3.1.1 节的聚类算法可以初步返回一系列排序完的聚类结果，由于聚类结果主要集中在微博的 Hashtag 字符串表面层次上，所以聚类容易使 Hashtag 相似的微博聚类在一起。但是在实际语料中，存在很多 Hashtag 和微博内容不匹配的现象，例如表 1 中的示例：

表 1 微博中 Hashtag 与内容不匹配示例

Tab.1 Examples of Mismatch between Hashtags and Contents in Micro-blog

ID	微博内容	Hashtag	内容指向对像
1	#柴静雾霾视频#污染企业快倒闭吧！	柴静雾霾视频	污染企业（贬义）
2	#贵州男童自尽#当地政府的职责哪去了？	贵州男童自尽	当地政府（贬义）

从表 1 的两个微博示例中我们可以看出，如果简单地将 Hashtag 对应的微博进行情感分布统计，会造成话题对应的情感分布不准确，甚至出现某些明显大众有支持情绪的话题中有负面情绪，例如表 1 中的“柴静雾霾视频”中，网民其实是对“污染企业”持有负面情绪而非对雾霾视频。

本节所做的工作是针对上述现象所展开的，目标是将事件相关的微博划分到正确的主题下。本文将这项任务定义为微博话题纠正任务。针对此任务，设计的算法流程如图 4：

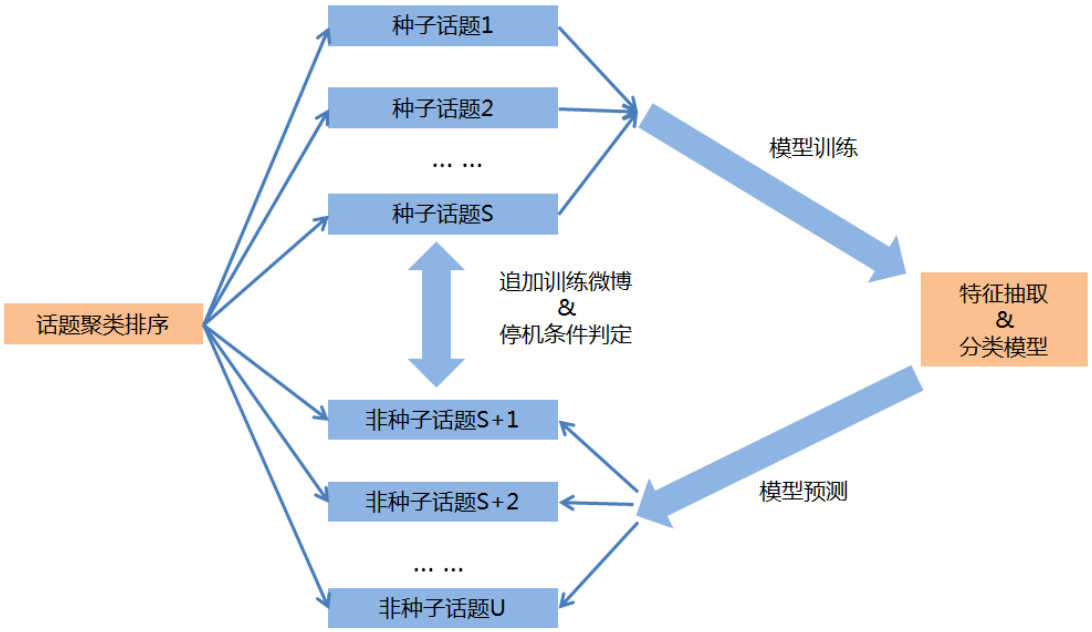

图 4 基于半监督学习的微博话题纠正模型

Fig.4 Flow Chart of Micro-blog Topic Rectification Based on Semi-supervised Method

对应的算法流程如下：

算法 1 基于半监督学习的微博话题纠正算法

Algorithm 1: Topic Rectification Algorithm Based on Semi-supervised Method

输入：话题聚类排序后的结果，共 K 个

输出：自扩充的 S 个话题及其相关微博

1. 前 S 个话题划分为“种子话题”，后 U 个话题划分为“非种子话题”。U 集合按照与 S 集合的相似度排序划分为待预测集 U1 和训练反例集 U2，这里当然 K=S+U，U=U1+U2；
2. 通过表 2 的特征模板对 S 类话题的语料进行特征抽取以及模型训练；
3. 将训练得到的模型预测非种子待预测微博集合 U1；
4. 将 U1 中微博分类结果概率大于阈值的直接加入到对应 S 集合中，同时将微博从 U1 集合中删除；
5. 从第 2 步开始循环，直至达到停机条件（S 个话题对应微博的添加率小于阈值）。

表 2 关键话题自动抽取特征模板

Tab.2 Feature Template in Seed Topic Extraction

ID	特征名称	特征抽取样例	特征抽取结果
1	Unigram	#乐嘉耍酒疯#真是醉了。	# 乐嘉 耍 酒疯 # 真是 醉了 。
2	Topic Unigram	#乐嘉耍酒疯#真是醉了。	乐嘉 耍 酒疯

4 实验结果

这部分将对微博事件情感分布的原因分析算法进行真实语料的实验。主要分为话题聚类距离算法对比实验、话题聚类结果排序算法对比实验和微博事件话题纠正算法实验。

4.1 实验数据

本文收集了 2015 年 5 月到 7 月间发生的微博热点事件 15 个及其相关的微博（如表 3 所示），并对每个事件的最热门的 5 个代表不同侧面的话题进行人工标注，作为标准答案。部分示例如表 4 中所示。

表 3 实验对应的 15 个微博热点事件（部分）

Tab.3 Part of Fifteen Hot Event in Micro-blog Example

ID	Topic	Micro-blogs' number
1	毕福剑不雅视频事件	121204
2	云南女导游辱骂游客	18338
3	四川女司机被当街暴打	324775
	……	
15	贵州自尽男童	30848

表 4 实验数据标注样例

Tab.4 Annotated Example of Dataset

ID	Event	Topic ID	Annotated Topic
5	庆安枪击事件	5.1	请央视公布完整监控视频
		5.2	开枪民警属正当履职
		5.3	庆安官场全面崩塌
		5.4	真相别总靠“倒逼”
		5.5	开枪民警李乐斌偿命

表 3 中第二列为微博热点事件的名称，第三列为对应的微博数目，微博总数共计 121 万句。表 4 中列举了“庆安枪击事件”及标注的五个不同侧面的话题。

4.2 评价指标

本任务最终采用了准确率@5 指标来反映算法的排序结果的优劣性，使用微博数目平均添加率和微博平均添加准确率作为微博自扩充算法的评价指标。

准确率@5（P@5，Precision at Five）指标在本文任务中的定义为排序结果最前的 5 条预测正确的话题数目与前 5 条标准答案中话题数目的比值，即公式 5：

$$P@5 = \frac{\text{前 5 条结果中命中的话题数目}}{\text{标准答案话题数目}} \tag{5}$$

微博数目平均添加率，是每个话题相关的微博自扩充后的添加率平均值，即公式 6：

$$Average\ Ratio\ of\ Addiction = \frac{1}{S}\sum_{i=1}^{S}\frac{\text{话题 i 微博自扩充数目}}{\text{话题 i 微博总数目}} \tag{6}$$

追加微博的平均命中率，即算法中追加到现有话题的微博正确命中的数目与当前话题的微博数目比值，即公式 7：

$$Average\ Hit\ Ratio\ of\ Addiction = \frac{1}{S}\sum_{i=1}^{S}\frac{\text{话题 i 微博自扩充命中数目}}{\text{话题 i 微博总数目}} \tag{7}$$

4.3 话题聚类距离算法对比实验

在 3.1.1 中，本文介绍了两种层次聚类时的距离计算方法，分别是基于微博词频/逆文档频率相似度方法和基于 Hashtag 字符串相似度方法。其 P@5 结果如表 5 所示：

表 5 话题聚类距离计算方法对比实验

Tab.5 Contrast Experiment on Distance Formula in Clustering Algorithm

指标\距离公式	Similarity_TFIDF	Similarity_Hashtag
准确率@5（%）	53.3	**78.7**

可见，使用表 5 中的 TF/IDF 方法（即基于微博词频/逆文档频率相似度方法）的准确率要远低于基于字符串相似度计算方法。由于单个事件相关的微博所用词语的重合度比较大（例如“携程官网被黑”事件相关微博中频繁出现“携程”、“服务器”等词汇），干扰了聚类过程，导致层次聚类的最优相似度阈值难以确定。反而使用 Hashtag 的字符串相似度方法，经过层次聚类的迭代过程，在本文任务中取得了不错的效果。

4.4 话题聚类结果排序算法对比实验

本小节主要对比了不同排序算法的实验效果，即简单的根据微博数目排序（公式 3）和根据微博数目与聚类结果话题数的加权关系排序（公式 4），此实验中固定了层次聚类的距离计算公式为字符串相似度方法。实验结果如表 6 所示：

表 6 话题聚类结果排序算法对比实验

Tab.6 Contrast Experiment on Ranking Method in Clustering Algorithm

指标\排序公式	排序公式 3	排序公式 4
准确率@5（%）	66.7	**78.7**

从表 6 中的实验结果可以看出，如果单纯地按照聚类后簇对应的微博数目作为排序指标，效果不如加入簇下的话题数目信息的排序方法。其原因是由于单个簇下聚类的话题数目也在很大程度上代表了该话题的热议程度，即话题聚类结果所含话题越多，其热议程度越高。

4.5 微博事件话题纠正实验

微博事件话题纠正算法的目标是将Hashtag信息与微博文本谈论的主题矛盾的那部分微博进行话题纠正，达到对话题相关的微博进一步自扩充的效果。那么算法平均可以给数据集增加多少微博，同时追加的微博的命中率又如何，是评价这个算法优劣的指标。

欲对这两项进行评估，本文设计了微博数目平均添加率和追加微博的平均命中率两个指标。由于算法是个迭代的半监督学习算法，所以采用折线图来直观表示实验结果，横轴为迭代次数，纵轴分别为两项指标。如图 5、图 6 所示。

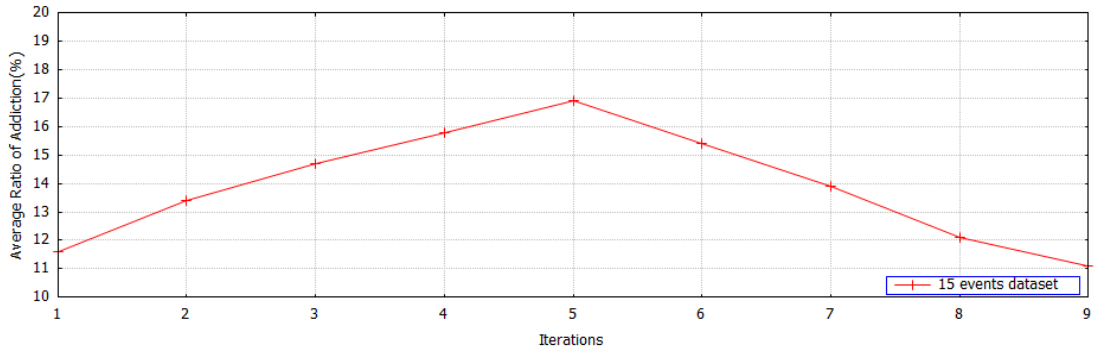

图 5 微博话题纠正算法——微博数目平均添加率

Fig.5 Line Chart Results of Micro-blog Average Addiction Ratio in Topic Rectification Method

实验语料使用了 4.1 节介绍的微博事件数据。在算法迭代过程中，微博数目添加率从开始的 11.6%上涨到峰值 16.9%随后降低。通过对比实验数据，其原因是迭代初期，随着分类模型追加的训练微博数量增多，引起使模型的特征增多，从而可以预测更多的样本；后期由于所剩预测样本数目不足，导致添加率降低。

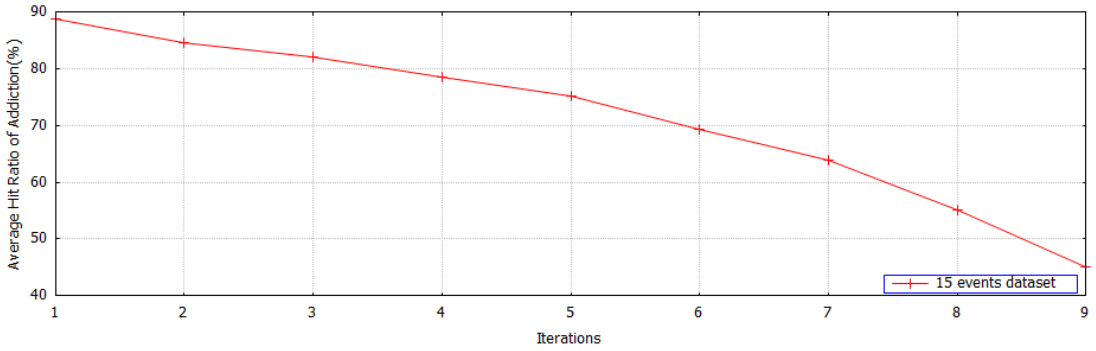

图 6 微博话题纠正算法——追加微博的平均命中率

Fig.6 Line Chart Results of Micro-blog Average Hit Ratio of Addiction in Topic Rectification Method

通过对追加的微博进行抽样标注，本文得出了图 6 的实验结果。结果显示，追加的微博平均命中率从第一轮迭代时的 88.5%到第 9 轮的 44.5%，是一直降低的。造成这种结果的原因是随着模型迭代更新，一方面训练语料不断在增加，另一方面训练噪声也在不断累加。最终造成分类器分类能力不断减弱。

最终本文折中了微博数目平均添加率和追加微博的平均命中率，将迭代次数设置为 2，保证了微博的自扩充数目和追加微博的命中率综合效果最优。

5 结论及展望

本文创新性地提出了一项新任务，即微博事件情感分布的原因分析。通过问题的引入及对比当前国内外的研究现状，本文提出了一套针对性的解决方案。具体工作由定义相关语言现象、任务而展开，例如“话题”、“微博话题纠正任务”等。同时构建了百万级的中文微博事件相关的语料数据集。随后通过一套无监督学习的层次聚类算法和半监督学习的微博话题纠正算法的结合算法，对微博话题进行了相似聚类、基于热点的排序以及微博内容话题不匹配的错误纠正。给定单个事件，算法输出为挖掘出的话题及相关的微博，最后通过情感分类的相关技术实现了对话题对应的微博的情感分布统计，达到对事件情感分布的原因进行分析的目的。在下一步工作中，我们将进一步探索聚类的改进方法，同时加大对情感分类算法的研究，进一步提高微博事件情感分布的原因分析准确率。

参 考 文 献

[1] Weng J, Lee B S. Event Detection in Twitter[J]. ICWSM, 2011, 11: 401-408.

[2] Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388.

[3] Spina D, Meij E, de Rijke M, et al. Identifying entity aspects in microblog posts[C]//Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 1089-1090.

[4] Das A, Kannan A. Discovering topical aspects in microblogs[C]//Proceedings of COLING. The 25th International Conference on Computational Linguistics: Technical Papers , 2014: 860-871.

[5] Zhao Y, Qin B, Liu T, et al. Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on micro-blog[J]. Multimedia Tools and Applications , 2014, 22(1): 1-18.

[6] Rosenthal S, Nakov P, Kiritchenko S, et al. Semeval-2015 task 10: Sentiment analysis in twitter[C]//Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval. 2015.

[7] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 13-16.

[8] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis[M]. John Wiley & Sons, 2009.