**文章编号**：1004-4353(2019)04-0349-07

# 基于改进互信息的微博新情感词提取

柳文婷

（安徽理工大学 计算机科学与工程学院，安徽 淮南 232001）

**摘要**：针对微博新词的情感倾向分析问题，提出了一种改进互信息的微博新情感词提取方法. 首先，对预处理后的微博数据进行 $N$ 元切分，以此得到候选字串；然后，通过计算多字互信息(multiword mutual information，MMI)和左右侧邻接熵对候选字串进行扩展和过滤得到候选新词，再将候选新词与相应词典进行对比得到新词；最后，通过词间情感相似度(sentiment similarity between the words，SW)计算出新词的情感倾向值，从而得到新情感词. 实验结果显示，该方法对新词情感倾向识别的准确率、召回率和 $F_1$ 值比文献[4]方法分别提高了 13.14%、5.81%和 8.59%，因此该方法具有很好的应用价值.

**关键词**：微博；新情感词；$N$ 元切分；多字互信息；词间情感相似度

**中图分类号**：TP391 **文献标志码**：A

# The extraction of micro-blog new sentiment word based on improved mutual information

LIU Wenting

(*School of Computer Science and Engineering*, *Anhui University of Science & Technology*, *Huainan* 232001, *China*)

**Abstract**: Aiming at the problem of sentiment tendency analysis of new words in micro-blog, a method of extracting new sentiment words based on improved mutual information was proposed. Firstly, $N$-gram segmentation method are performed in the preprocessed micro-blog data to obtain candidate string. Then, the candidate word string was expanded and filtered by calculating the multiword mutual information (MMI) and the left and right adjacency entropy to obtain the candidate new words. And the candidate new words were screened to obtain new words by comparing the corresponding dictionaries. Finally, the sentiment tendency value of the new word was calculated by sentiment similarity between the words (SW), and the new sentiment word was obtained. The experimental results show that the precision rate, recall rate and $F_1$ value of the method for micro-blog new sentiment word recognition are 13.14%, 5.81% and 8.59% higher than those in the literature [4]. Therefore, the method has good application value.

**Keywords**: micro-blog; new sentiment words; $N$-gram segmentation; multiword mutual information; sentiment similarity between the words

## 0 引言

微博新情感词的提取一般分为新词提取和新词情感识别[1]. 目前，新词提取方法主要分为基于规则的方法和基于统计的方法. 基于规则的方法提取新词[2]主要是根据构词学原理或词性信息来匹配新词，但由于该方法需要人工标注信息，所以需要耗费大量的人力物力. 基于统计的方法提取新词[3]是使用统

**收稿日期**：2019-08-14 **作者简介**：柳文婷(1996—)，女，硕士研究生，研究方向为自然语言处理及数据挖掘.
**基金项目**：安徽省高校拔尖人才培育项目(gxbjZD15)；安徽省自然科学基金面上项目(1908085MF189)

计学方法来建造模型并判断字串是否为新词,该方法适用于大规模的语料库.也有学者将上述两种方法结合起来提取新词,这种方法虽然效果较为稳定,但在实际应用中很难获得高质量的标记语料[4].

近年来,在微博情感倾向的相关研究中,大多数学者都是通过对微博中的词汇和句子进行情感判断来分析微博的情感倾向[5-7],而对微博新情感词的相关研究较少.对微博的情感分析目前可分为基于词典的方法、基于机器学习的方法和基于词典与机器学习相结合的方法.基于词典的方法[8]主要是通过构造情感词典和制定一系列的规则来计算新词的情感值,该方法虽然判断新词情感的准确率较高,但召回率偏低,且构建不同领域情感词典的成本较高.基于机器学习的方法[9]是将文本的情感分析作为分类问题进行分析,分类算法主要有深度学习的方法和支持向量的方法.前者计算量大,但准确率较高;后者准确率相对较高,但不适用于大规模数据.基于词典与机器学习相结合的方法[10]是将词典融合到机器学习的模型中进行文本情感分析,该方法虽然可提高机器学习性能,但却需要人工收集情感词,因此使得情感词库的覆盖面较低.基于上述研究,本文结合新词构词特点,提出一种基于互信息和构造情感词库的微博新情感词提取方法,并通过实验验证该方法的可行性.

## 1 微博新情感词提取流程

微博作为一种服务类的社交网站,它具有公开性、及时性以及多样性.绝大多数用户都能随时随地以文字、图片或视频来表达自己的所思所想,但由于微博用户的教育背景、生活习惯、语言表达等的不同,因此使得微博数据较为混乱,其中最常见的问题有:①重复性.微博上内容重复的网页较多,且其真实性有待确认.②随意性.微博用词(包括文本、图片等)缺少规范,随意性很大.③领域广.微博文本涉及的领域广,仅使用某一领域的提取方法会极大地影响新词的提取准确率.④人造词多.用户使用的一些新词或是来自某地方言或是自创,不存在于字典中,因此难以判断其情感倾向.为解决以上问题,并达到快速、准确地提取新情感词,本文提出一种基于多字互信息和词间情感相似度的微博新情感词提取方法.该方法主要分为两个阶段:新词提取和新词情感倾向分析.

1)新词提取阶段.该阶段的主要工作是对预处理的数据进行 $N$ 元切分,以此得到候选字串;然后再根据多字互信息和左右侧邻接熵计算候选字串的内部统计量和外部统计量的值,以此得到候选新词集;最后将得到的候选新词集与词典进行对比,删除词典中已有的词后即得到新词集.

2)新词情感倾向分析阶段.该阶段主要是根据新词之间的情感倾向影响来改进情感倾向点互信息公式,以此得到词间情感相似度的计算公式;根据该公式计算新词的情感倾向值,并依据该值判断新词的情感倾向,删除中性新词后剩下的即为新情感词.

新情感词的提取过程如图 1 所示.

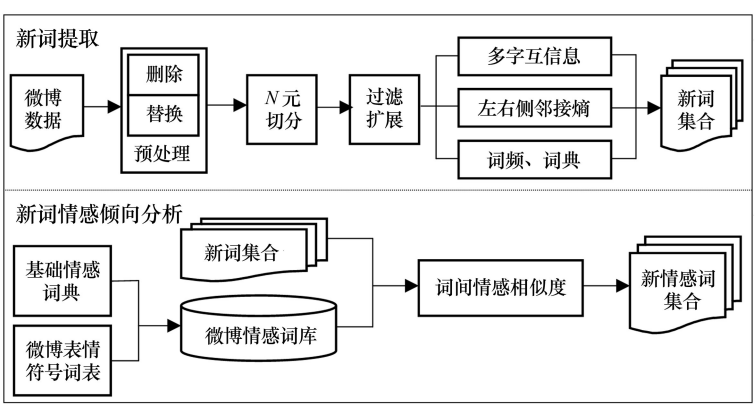


**图 1 微博新情感词的提取过程**

## 2　微博新词的提取

### 2.1　微博数据的预处理

由于微博具有用户背景不同、包含领域广、文本书写不规范、字数有限等特点，因此通过爬取方法所获得的微博数据中存在大量的噪声. 为此，本文对微博数据进行预处理，预处理后文本的处理时间和数据的存储空间可得到有效提高. 预处理的主要方法是删除和替换.

1）删除. 删除微博数据中的链接、重复的标点符号以及“#＊＊＊＊#”、“@＋用户名”等微博程序自带的固定字串.

2）替换. 用 ICTCLAS 词法分析系统切分微博文本中的词句，将切分后得到的标点符号和停用词用空格代替，并将繁体字换成简体字.

### 2.2　微博新词的提取方法

1）$N$ 元切分方法. 一般的分词系统都是根据已有的字典对句子进行切分，这种切分方法可能会造成错误切分或遗漏新词. 例如“王经／理／了／理／袖口”，分词系统有可能会将这个句子划分成“王／经理／了／理／袖口”，遗漏掉“王经”这个新词. 所以，本文采用 $N$ 元切分方法来切分文本. $N$ 元切分方法的基本思想是对文本进行逐一扫描、$N$ 字切分，切分后的每 $N$ 个单字构成一个字串. 因目前二元和三元切分技术较为成熟，且新词一般由 2 ～ 4 个字组成[11]，所以本文中的 $N$ 取 2 和 3.

2）多字互信息. 互信息是表示两个字之间的依赖程度，传统的互信息的表达式为

$$MI(x,y)=\log\frac{P(xy)}{P(x)P(y)}. \tag{1}$$

其中：$P(xy)$ 表示 $x$ 和 $y$ 在语料库中共同出现的频次与语料库中总词数的比；$P(x)$ 和 $P(y)$ 分别表示 $x$ 和 $y$ 单独出现在语料库中的频次与语料库中总词数的比.

由公式(1)可以看出，传统的互信息只考虑了候选词被划分成两部分的构成模式，即该方法只可对 2 字词进行划分，而无法对多字词进行划分. 因此，本文对传统互信息进行改进，即考虑多字候选词（字数超过 2 的候选词）的构成模式. 3 字词的所有构成模式如图 2 所示. 根据图 2 可类推 4 字词的构成模式（7 种），且所有构成模式包含的元素为 9 种.

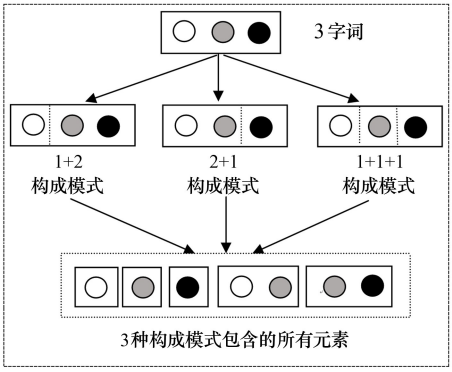


**图 2　3 字词的构成模式**

为了扩大新词识别的范围和提高新词识别的准确率，本文对多字互信息进行如下定义：

**定义 1**　多字互信息（multiword mutual information，MMI）是在多种构成模式下衡量多字词的内部凝聚度，其计算公式为

$$MMI(a_1,a_2,\cdots,a_n)=\log\frac{P(a_1,a_2,\cdots,a_n)}{\sqrt[T]{\prod_{j=1}^{T}P_j(a_1,a_2,\cdots,a_n)}}. \tag{2}$$

其中：$n$ 为字串 $a_1,a_2,\cdots,a_n$ 的总字数；$T$ 为字串 $a_1,a_2,\cdots,a_n$ 在多种构成模式下的元素种类个数，$T=\sum_{i=1}^{n-1}(n-i+1)$；$P(a_1,a_2,\cdots,a_n)=\frac{F(a_1,a_2,\cdots,a_n)}{N}$，$N$ 为语料库中微博的总条数，$F(a_1,a_2,\cdots,a_n)$ 表示字串 $a_1,a_2,\cdots,a_n$ 在语料库中出现的频次；$P_j(a_1,a_2,\cdots,a_n)=\frac{F_j(a_1,a_2,\cdots,a_n)}{N}$，$F_j(a_1,a_2,\cdots,a_n)$ 表示所有构成模式中的第 $j$ 种元素在语料库中出现的频次.

3) 左右侧邻接熵. 左右侧邻接熵能够反映字符串邻接元素的不确定性，熵值越大邻接元素的不确定性越大，即该字符串成为词的可能性也越大. 左侧邻接熵和右侧邻接熵的计算公式[12] 分别为：

$$E_l=-\sum_{i=1}^{|V_l|}\frac{n_i}{n}\log\left(\frac{n_i}{n}\right), \tag{3}$$

$$E_r=-\sum_{j=1}^{|V_r|}\frac{m_j}{m}\log\left(\frac{m_j}{m}\right). \tag{4}$$

其中：$|V_l|$ 和 $|V_r|$ 分别为左右侧邻接字种类的数量；$n$ 和 $m$ 分别为左右侧邻接字的总个数；$n_i$ 和 $m_j$ 分别为左右侧某一种邻接字的个数.

### 2.3 微博新词的提取算法

本文提出的微博新词提取算法(算法 1) 如下：

输入：微博文本集合 $T$，字频阈值 $\theta_1$，多字互信息阈值 $\theta_2$，左右侧邻接熵阈值 $\theta_3$ 和 $\theta_4$，词典 $D$

输出：微博新词集合 $E$

step 1 对微博文本集合 $T$ 进行预处理.

step 2 对 step 1 中得到的内容进行 $N$ 元切分得到二元字串和三元字串，并统计每个字串的频率，删除频率小于 $\theta_1$ 的字串后得到候选字串.

step 3 计算 step 2 中候选字串的多字互信息 $MMI(a_1,a_2,\cdots,a_n)$，删除多字互信息小于 $\theta_2$ 的字串.

step 4 计算 step 3 中剩余候选字串中二元字串的左右侧邻接熵$\{E_l,E_r\}$. 若 $E_l\geqslant\theta_3$，且 $E_r\geqslant\theta_4$，则将该字串添加到新词集合 $E$ 中.

step 5 计算 step 3 中剩余候选字串中三元字串的左侧邻接熵 $E_l$. 若 $E_l\geqslant\theta_3$，字串的左边界确定，执行 step 6；否则，字串向左扩展一个字后，执行 step 6.

step 6 计算 step 5 中得到的字串的左侧邻接熵$\{E_l,E_r\}$. 若 $E_l\geqslant\theta_3$，且 $E_r\geqslant\theta_4$，则将该字串添加到新词集合 $E$ 中.

step 7 对比新词集合 $E$ 和词典 $D$，删除共有的词后即得到最终的新词集合.

从上述的算法中可以看出：算法首先对微博文本进行扫描，以此判断并建立候选字串集，此时的时间复杂度为 $O(n)$，$n$ 为候选字串的个数. 然后再对候选字串集进行多字互信息过滤，此时的时间复杂度为 $O(n)$. 候选字串为二元字串的有 $k$ 个，三元字串的有 $(n-k)$ 个. 通常情况下，三元字串运用左右侧邻接熵进行扩展的次数极少超过 2 次，因此扩展的时间复杂度可记为 $O(n)$.

## 3 新词情感倾向分析

### 3.1 构建微博基础情感词库

在微博中，表情符号和情感词都是用户情感的直接表达，因此本文使用基于词典的方法分析新词情感倾向. 本文将知网情感词典与台湾大学简体中文情感极性词典合并、去重后的情感词集作为基础情感词典，然后选取倾向性明显的(出现频次在前 36 个)的褒贬义表情符号作为微博表情符号词表，如表 1 所示. 最后将基础情感词典与微博表情符号词表去重、合并得到微博基础情感词库.

**表 1　微博褒贬义表情符号**

| 褒义(18 个) | 贬义(18 个) |
|---|---|
| [可爱][威武][鼓掌][嘻嘻][哈哈][爱你] | [哼][怒][酸][泪][挖鼻][悲伤] |
| [亲亲][抱抱][加油][赞啊][给力][笑哈哈] | [晕][衰][吐][白眼][鄙视][怒骂] |
| [好爱哦][太开心][酷][心][耶][good] | [伤心][弱][吃惊][可怜][抓狂][生病] |

### 3.2　新词情感倾向识别

一般情况下，微博中出现的新词是用户情绪的一种宣泄或表达，而情感词是用户情感的直接体现. 因用户发表的微博内容中出现的新词和情感词的情感倾向大多是类似的，所以可以从用户使用的情感鲜明的情感词来分析新词的情感倾向. 由于传统的情感倾向点互信息只考虑了微博情感词库中已经存在的情感词，未考虑新词的情感对与其共同出现的其他新词情感的影响，因此需对传统的情感倾向点互信息进行改进. 本文通过分析新词之间的情感倾向影响，对情感倾向点互信息进行改进，即通过分析词间相似度来分析新词之间的情感倾向影响. 词间情感相似度的相关定义和公式如下：

**定义 2**　词间情感相似度(sentiment similarity between the words，SW)是衡量同一个文本中某个新词和情感词对同一条微博中的其他新词的情感倾向的影响程度，其计算公式为：

$$SW_j=\alpha(PA\_PMI_j-NA\_PMI_j). \tag{5}$$

其中：$PA\_PMI_j=\dfrac{1}{|C^+|}\sum_{i=1}^{|C^+|}\log\dfrac{P(X,C_i^+)}{P(X)P(C_i^+)}$，$P(X,C_i^+)$ 表示新词 $X$ 和第 $i$ 个正向情感词 $C_i^+$ 出现在同一条微博文本中的概率，$P(X)$ 表示新词 $X$ 单独出现的概率，$P(C_i^+)$ 表示第 $i$ 个正向情感词 $C_i^+$ 单独出现的概率，$|C^+|$ 表示正向情感词的个数；$NA\_PMI_j=\dfrac{1}{|C^-|}\sum_{i=1}^{|C^-|}\log\dfrac{P(X,C_i^-)}{P(X)P(C_i^-)}$，$P(X,C_i^-)$ 表示新词 $X$ 和第 $i$ 个负向情感词 $C_i^-$ 出现在同一条微博文本中的概率，$P(C_i^-)$ 表示第 $i$ 个负向情感词 $C_i^-$ 单独出现的概率，$|C^-|$ 表示负向情感词的个数；$\alpha$ 表示语料中正负面情感倾向强度比，$\alpha=\dfrac{\sum_j PA\_PMI_j}{\sum_j NA\_PMI_j}$，$\sum_j PA\_PMI_j$ 表示所有新词与正向情感词的平均点互信息之和，$\sum_j NA\_PMI_j$ 表示所有新词与负向情感词的平均点互信息之和.

由以上可知，由算法 1 得到的微博新词只有经过词间情感相似度的判断才能确定某个新词是否是具有情感的新词. 微博新情感词提取算法(算法 2) 如下：

输入：微博新词集合 $E$，微博基础情感词库 $C$，词间情感相似度阈值 $\theta_5$ 和 $\theta_6$

输出：微博新情感词集合 $S$

step 1　结合微博基础情感词库，计算新词集合中每个词的词间情感相似度 $SW$；

step 2　如果 $SW<\theta_5$ 或 $SW>\theta_6$，则判定该词为微博新情感词.

从算法 2 中可看出，只需对微博新词集合进行 1 次扫描即可完成所有新词的情感倾向分析，其时间复杂度为 $O(m)$，新词个数为 $m$.

## 4　实验

### 4.1　实验数据

1)新情感词提取. 爬取 2018 年 11 月—2019 年 3 月 3 个不同热门话题(“军训式应援”“杨超越登上《人物》杂志”“翟天临学术造假”)的 40 000 条微博，用于微博新情感词提取.

2)停用词库. 对哈工大停用词词库、四川大学机器学习智能实验室停用词库、百度停用词表进行整

理、去重后得到本文的停用词库共计1 598个停用词,(不包括英文词和中文标点符号),用于微博数据预处理阶段.

3)词典.将知网文本词库与同义词词林相结合后得到的词语集合即为本文中使用的词典,用于候选新词筛选.

4)微博基础情感词库.将知网情感词典、台湾大学简体中文情感极性词典以及本文列出的微博褒贬义表情符号(表1)进行合并、去重,所得的情感词集即为本文的基础情感词典,用于新词情感倾向识别阶段.

### 4.2 实验评估指标

采用准确率 $P$(precision)、召回率 $R$(recall)、综合指标 $F$($F$-score)来评价算法的准确性,各指标的计算公式为:

$$P = TP/(TP + FP) \times 100\%, \tag{6}$$

$$R = TP/(TP + FN) \times 100\%, \tag{7}$$

$$F = 2PR/(P + R) \times 100\%. \tag{8}$$

其中:$TP$ 表示将待测的词语预测为新(情感)词,实际也为新(情感)词的数量;$FP$ 表示将待测的词语预测为新(情感)词,实际为非新(情感)词的数量;$FN$ 表示将待测的词语预测为非新(情感)词,实际为新(情感)词的数量.

### 4.3 实验分析

为了验证本文算法的有效性,进行两方面实验:一是微博新词提取算法(算法1)的有效性验证,二是微博新情感词提取算法(算法2)的有效性验证.由于各个话题讨论的内容不同,所以本文将他们分开讨论.话题1("军训式应援")是针对国内某个明星而提出的,由于该明星形象良好,因此微博中出现的都是情感偏正向的词.话题2("杨超越登上《人物》杂志")是刚出道的明星(杨超越)登上《人物》杂志而出现的各种不同评价.话题3("翟天临学术造假")由于是因翟天临学术造假而引起的话题,所以该话题中的词大部分都是情感负向的词.算法1在3个话题中得到的部分高频新情感词如表2所示.

**表2 高频新词识别结果示例**

| 话题 | 二元词 | 三元词 | 四元词 |
|---|---|---|---|
| 话题1 | 开挂、C位、牛逼 | 吊炸天、眼镜杀 | C位出道、前方高能 |
| 话题2 | 冲鸭、杠精、控评 | 键盘侠、走花路 | 炒鸡厉害、喜大普奔 |
| 话题3 | 学霸、脱粉、实锤 | 零容忍、演技派 | 顶级流量、压力山大 |
| 总数 | 932 | 353 | 132 |

表3为本文算法1、传统的 $N$ 元方法和文献[3]算法的新词提取结果.由表3可知,本文算法1比传统的 $N$ 元方法的准确率、召回率和 $F_1$ 值分别提高了25.86%、31.52%和30.60%,比文献[3]算法的准确率、召回率和 $F_1$ 值分别提高了10.16%、11.47%和11.03%.算法1的提取效果优于传统的 $N$ 元方法和文献[3]算法的主要原因是:传统的 $N$ 元方法对多字词的识别率较低,而且也未考虑新词内部统计量和外部统计量对新词识别的影响;文献[3]的算法过度依赖于分词系统,使一些词被错分.

**表3 不同方法的新词提取结果** %

| 方法 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 传统的 $N$ 元方法 | 24.65 | 12.03 | 16.17 |
| 文献[3]算法 | 40.35 | 32.08 | 35.74 |
| 本文算法1 | 50.51 | 43.55 | 46.77 |

表 4 是通过本文算法 2 得到的微博新情感词示例，表 5 是算法 2 和文献[4]算法的新情感词提取结果. 从表 5 可以看出，算法 2 比文献[4]算法的准确率、召回率和 $F_1$ 值分别提高了 13.14%、5.81%和 8.59%. 其主要原因是文献[4]的算法只考虑了内部统计量和外部统计量对新情感词提取的影响，并没有考虑新词的语义信息，进而导致提取结果中有很多新词不是情感词；而算法 2 在统计量的基础上加入了词在情感词典中的语义信息，进而使得新情感词的提取准确率有所提高.

**表 4　微博新情感词示例**

| 正向 | 负向 |
|---|---|
| 给力、牛逼、开挂、比心、冲鸭、吊炸天、<br>演技派、走花路、C 位出道、家里有矿、宝藏少女 | 杠精、尼玛、学渣、辣鸡、凉凉、伤不起、<br>键盘侠、辣眼睛、天凉王破、一脸懵逼 |

**表 5　2 种算法的新情感词提取结果**　%

| 方法 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 文献[4]算法 | 47.10 | 36.54 | 41.15 |
| 本文算法 2 | 60.24 | 42.35 | 49.74 |

## 5　结论

本文基于新词构词模式多样的特点，提出了一种基于多字互信息和词间情感相似度的微博新情感词提取方法. 实验结果表明，本文方法的微博新词提取的准确率(50.51%)、召回率(43.55%)和 $F_1$ 值(46.77%)均优于传统的 $N$ 元方法和文献[3]的方法，新词情感倾向识别的准确率、召回率和 $F_1$ 值比文献[4]方法分别提高了 13.14%、5.81%和 8.59%，因此本文算法具有很好的应用价值. 在今后的研究中，我们将通过完善微博情感词库以及使用融合机器学习的方法来进一步提高新词情感倾向识别的准确率和召回率.

## 参考文献：


[1] HUANG M L, YE B R, WANG Y C, et al. New word detection for sentiment analysis[C]//52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: ACL, 2014:531-541.

[2] TANG Z, FU Z M, GONG Z R, et al. A parallel conditional random fields model based on spark computing environment[J]. Journal of Grid Computing, 2017,15(3):1-20.

[3] 雷一鸣，刘勇，霍华. 面向网络语言基于微博语料的新词发现方法[J]. 计算机工程与设计，2017(3):789-794.

[4] 王非. 基于微博的情感新词发现研究[J]. 软件学报，2015,36(11):6-8.

[5] 李勇敢，周学广，孙艳，等. 中文微博情感分析研究与实现[J]. 软件学报，2017,28(12):3183-3205.

[6] 唐晓波，刘广超. 细粒度情感分析研究综述[J]. 图书情报工作，2017,61(5):132-140.

[7] LI W W, LI Y Q, WANG Y. Chinese microblog sentiment analysis based on sentiment features[C]//Asia-Pacific Web Conference. Suzhou, China: Springer, 2016,9932:385-388.

[8] ZHAO C J, WANG S G, LI D Y. Exploiting social and local contexts propagation for inducing Chinese microblog-specific sentiment lexicons[J]. Computer Speech and Language, 2019,55:57-81.

[9] HAO Z F, CAI R C, YANG Y Y, et al. A dynamic conditional random field based framework for sentence-level sentiment analysis of Chinese microblog[C]//IEEE International Conference on Computational Science & Engineering. Guangzhou, China: IEEE, 2017:135-142.

[10] 张仰森，郑佳，黄改娟，等. 基于双重注意力模型的微博情感分析方法[J]. 清华大学学报(自然科学版)，2018,58(2):122-130.

[11] 张婧，黄锴宇，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.

[12] 张华平，高凯，黄河燕，等. 大数据搜索与挖掘[M]. 北京：科学出版社，2014:107.