文章编号：1003-0077(2017)03-0048-07

情感词发现与极性权重自动计算算法研究

张华平[1,2]，李恒训[3]，李清敏[4]

(1. 北京理工大学 计算机学院，北京 100081；
2. 北京市海量语言信息处理与云计算应用工程研究中心，北京 100081；
3. 公安部第一研究所 信安部，北京 100048；
4. 工业和信息化部电子科学技术情报研究所，北京 100040)

摘 要：随着互联网电子商务和各种社交网络应用的快速发展，产生了大量的用户评价信息。为满足快速整理这些评价信息的需求，情感倾向性分析应运而生。情感词典是各类情感倾向性识别算法的基础，收集一部全面且权重合理的情感词典，往往可以简单快速而有效地解决情感分析问题。但情感词典规模有限，而网络上新的情感词层出不穷，语言使用不规范，人工整理耗时耗力。已有的情感词收集方法较复杂，且领域性强，收集的情感词可扩展性差。本文提出一种自动挖掘潜在情感词并计算其极性权重的算法，该算法与应用领域无关，具有良好的扩展性。该方法利用共现特性，基于朴素贝叶斯公式能检测出未知的情感词，并根据其情感权重值的大小判断其情感极性，可有效地扩展情感词典，将已有的情感词典进一步量化。在理论研究的基础上，本文分别针对京东、豆瓣及大众点评网三组评论语料做了实验，其结果的准确率都基本在 90%以上，验证了该方法的有效性和实用性，为情感倾向性分析提供了知识库基础。

关键词：情感词；情感权重；情感程度判别；情感词典
中图分类号：TP391 **文献标识码**：A

Research on Automatic Emotional Word Detection and Polarity Weighting Algorithm

ZHANG Huaping[1,2], LI Hengxun[3], LI Qingmin[4]

(1. Department of Computer, Beijing Institute of Technology, Beijing 100081, China;
2. Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application, Beijing 100081, China;
3. First Reaearch Institute of the Ministry of Public Security of PRC, Beijing 100048, China;
4. Electronic Information Institute of Science and Technology, Ministry of Industry and Information Technology, Beijing 100040, China)

Abstract: Rapid development of Internet commerce and various social networking applications leads to a large number of user comment information. To meet the requirement of fast processing these information, sentiment and its polarity analysis arises at the moment. Emotion dictionary is the basis for all kinds of recognition algorithms of emotional polarity. To build a comprehensive emotional dictionary with rational weight, this paper proposes an automatic emotion weight (AEW) algorithm to mine the potential emotional words and estimate the emotion weight, with the advantage of domain independence and good scalability. The method uses special type of co-occurrence, which is based on Bayesian theory, to recognize unknown emotion words, judge the sentiment polarity according to the value of its emotion weight. We verify the theoretical research by three empirical analysis of data form JD.com, douban.com and dianping.com, achieving a precision about 90%.

Key words: sentiment lexicon; polarity weight; emotional orientation degree; emotion dictionary

收稿日期：2014-09-25 **定稿日期**：2015-03-18
基金项目：国家重点基础研究发展计划(973 计划)(2013CB329601)

1　引言

随着 Web 2.0 的迅速发展，互联网上涌现出大量用户参与及评论信息。人们通过微博、论坛等社会媒介发表大量对事件、产品等有价值的评论。评论信息不仅表达人们的情感色彩及情感倾向性，也为潜在用户提供了参考价值，收集评论信息可以了解大众舆论对某一事件或产品的看法，用以支持决策；同时，还可以为生产商提供反馈，通过主观色彩的评论信息了解用户评价，把握用户的需求，改善产品与服务。但由于越来越多的用户在互联网上分享观点与评论，信息迅速膨胀，人工方式难以及时处理这些海量数据，为满足快速获取和整理评价信息的需求，情感倾向性分析技术应运而生[1]。

按照文本处理的规模不同，情感倾向性分析的研究工作可分为词语级、句子级[2]、篇章级及海量文本等几个研究层次。本文研究工作主要针对未知情感词计算其情感权重，判断其情感极性，不断扩展情感词典，属于词语级情感倾向性分析研究的范畴，是情感倾向性分析技术的基础性研究工作。典型的有朴素贝叶斯（naïve Bayes，NB）[3-4]、支持向量机（support vector machine，SVM）[5-6]和最大信息熵（maximum entropy，ME）[7]等方法。但由于用户产生内容以短句为主，特征少，机器学习方法的分类效果不理想，而句子情感往往是由句子中的几个情感词决定。此外，网络新词层出不穷，旧词新用等，所以利用情感词典进行情感倾向性分析往往是分析用户产生内容情感分析简单且有效的方法。因此，如何自动构建情感词典并自动计算情感词的情感权重是解决情感倾向性分析问题的首要前提。

同时，在大数据时代下，语料规模大，人工标注情感极性及强度不太可能；随着新型互联网应用的出现，网络用语、新语层出不穷，如何利用大数据自动识别挖掘出这类表达情感的词语更具挑战性。目前，针对此情感精细判别和程度判别还没有合适的算法。面对上述问题与挑战，为解决如何自动计算情感词的情感权重，以及判断情感倾向性来扩展情感词典，本文提出一个情感词的自动发现及情感极性判别算法。该方法基于贝叶斯原理和大数据挖掘，能够挖掘未知情感词，并根据其情感权重值的大小判断其情感极性及情感倾向性程度，可有效扩展情感词典，丰富情感词典的精细化使用。另外，本方法与应用领域无关，具有良好的扩展性。

2　已有研究

在大数据时代下，数据本身有着不可忽视的价值，如何利用已知数据挖掘其潜在价值已引起国内外学者的广泛关注和研究。互联网电子商务及各种社交网络直接产生用户参与信息，迫切需要利用大数据挖掘其价值。随着情感倾向性分析技术的发展，本文利用大数据的原理，基于让数据本身创造价值的特点，解决情感倾向性分析技术中情感词的自动挖掘问题。

2.1　已有工作

目前，对于词的情感倾向性分析有两种研究方法：基于词典的方法及基于语料的方法。

基于词典的方法主要通过词语知识库或种子情感词扩展生成情感词典，中文以 HowNet、朱嫣岚[8]等研究成果为依据，利用 HowNet 的语义相似度计算新词与种子词的相似程度，从而判断新词的情感倾向性。这些方法的缺点是过于依赖种子情感词及其数量。英文情感倾向性判断是在 WordNet 和 General Inquire 的基础上进行的。Hu 和 Liu[9]对评论进行词性标记（POS），给定几个有极性的种子形容词，利用 WordNet 不断查找它们的同义词和反义词以扩大情感词典。该方法比较简单，但是只考虑了形容词（未考虑词典的更新、新型术语、网络新语）。Baccianella 等人[10]基于 WordNet 构建了认可度最高的 SentiWordNet。Hamouda 等人[11]基于机器学习方法，构建了 MLBL（machine learning based senti-word lexicon），取得了较 SentiWordNet 更高的微平均值[12]。

基于语料的方法主要是利用词之间的共现模式来确定其情感倾向性。Turney 等人[13]利用候选情感词与基准情感词的点互信息（PMI）进行词汇的情感倾向判断。这种方法领域针对性较强。Qiu[14]等人针对产品评论的情感词扩展方法，利用词性标注及句法依赖关系发现词与词间的搭配关系与语义关系进行情感词典的扩展。该方法需要大量的人工标注语料。

此外，情感词不同的情感强度对不同的潜在用户产生的影响也会不同。以往的情感强度往往规定为正情感词为＋1，负情感词为－1。目前针对情感强度的研究工作，大连理工大学信息检索研究室在林鸿飞教授的指导下整理和标注了一个中文本体资

源。该资源从不同角度描述一个中文词汇或者短语,包括词语词性种类、情感类别、情感强度及极性等信息,最终将情感共分为七大类21小类。该项工作是在Ekman的六大类情感分类体系的基础上构建的。但该研究是人工整理标注的结果,其情感强度不是自动计算得出的。

3 情感词典构建模型

3.1 算法依据

本文从贝叶斯原理出发,根据已知数据的总体情感信息和样本情感信息,通过自动机器学习,挖掘数据中潜在的情感词,从而实现情感词的自动挖掘及情感倾向性程度判别。下面,对贝叶斯定理进行简单描述。

贝叶斯定理也称贝叶斯推理,根据不确定性信息做出推理和决策需要对各种结论的概率做出估计。简单来说,贝叶斯原理是根据已知的总体信息、样本信息及先验概率,转换为求后验概率的过程。贝叶斯公式表示如式(1)所示。

$$P(A_i \mid X)=\frac{P(A_i,X)}{P(X)}=\frac{P(X \mid A_i)P(A_i)}{\sum_{j=1}^{n}P(X \mid A_j)P(A_j)} \tag{1}$$

其中,A 表示类别,A_i 表示第 i 类别,X 表示样本。$P(X)$ 表示样本信息,$P(X|A_i)$ 表示 A_i 类别中样本 X 的概率,即总体信息。$P(A_i)$ 表示类别 i 的概率,即先验概率。$P(A_i|X)$ 是表示样本 X 属于 A_i 类别的概率,即后验概率。公式(1)表示样本 X 属于 A_i 类别的概率。

3.2 算法原理

情感词典是解决情感分析问题的前提,情感词典的规模及质量影响情感倾向性分析的准确率,如何自动识别挖掘情感词,扩展情感词典尤为重要。本文采用贝叶斯原理,将情感词的识别定义为二元分类问题,即一个候选情感词是否为情感词,从而从语料中自动挖掘潜在情感词。

现给定语料及一部情感词典,有正情感词 $S_{正}$ 和负情感词 $S_{负}$ 之分。假设 C_i 表示一个字,$C_1..C_i..C_n$ 表示一个候选情感词,S^* 表示已知情感词,有正情感词 $S_{正}$ 和负情感词 $S_{负}$ 之分,Freq 表示语料中出现的频率。对于一个未知情感的候选情感词 $C_1..C_i..C_n$,判断其是否为情感词及情感倾向性程度的推导过程如下。

由贝叶斯公式得式(2)。

$$P(C_i \mid S^*)=\frac{P(S^*,C_i)}{P(S^*)}=\frac{\mathrm{Freq}(S^*,C_i)}{\mathrm{Freq}(S^*)} \tag{2}$$

由式(2)可计算情感词 S^* 中每个组成字 C_i 的概率 $P(C_i|S^*)$,由于情感词有正情感词 $S_{正}$ 和负情感词 $S_{负}$ 之分,所以每个组成字都会有正情感和负情感之分。

$$P(S^*)=\frac{\mathrm{Freq}(S^*)}{\mathrm{Freq}(N)}=\frac{\mathrm{Freq}(S^*)}{\sum_{i=0}^{n}\mathrm{Freq}(W_i)} \tag{3}$$

由式(3)可计算语料中情感词的分布。其中,N 表示语料中词的集合,W_i 表示语料中的任意词。$P(S^*)$表示情感词 S^* 的概率。

用 $S^{\sharp}$ 表示候选情感词的情感权重值。$P(S^*|C_1..C_i..C_n)$表示候选情感词 $C_1..C_i..C_n$ 是情感词的概率。由于是对一个未知词 $C_1..C_i..C_n$ 计算其情感权重,所以无论其是正情感词还是负情感词,$P(C_1..C_i..C_n)$是定值,可忽略不计。其次,由于 $P(C_1..C_i..C_n|S^*)$是未知的,不可计算,假设它们每个字是情感字的概率是条件独立的,则可以表示成以下形式,如式(4)所示。

$$\begin{aligned}S^{\sharp}&=\arg\max_s P(S^* \mid C_1..C_i..C_n)\\&=\arg\max_s \frac{P(S^*,C_1..C_i..C_n)}{P(C_1..C_i..C_n)}\\&\approx\arg\max_s P(S^*,C_1..C_i..C_n)\\&=\arg\max_s P(C_1..C_i..C_n \mid S^*)P(S^*)\\&=\arg\max_s \prod P(C_i \mid S^*)P(S^*)\end{aligned}$$

(条件独立性假设 T) (4)

利用式(2)和式(3)计算得出 $P(C_i/S^*)$及 $P(S^*)$,并代入式(4),就可计算出候选情感词的情感权重。为了更好地表示结果,对其取 log,表示如式(5)。

$$S^{\sharp}=\arg\max_s \sum_{i=0}^{n}\log P(C_i \mid S^*)+\log P(S^*) \tag{5}$$

由于每一个候选情感词都有正情感权重 $S^{\sharp}_{正}$ 和负情感权重 $S^{\sharp}_{负}$,则其最终的情感倾向性可以表示为两者的情感之差,即

$$S=S^{\sharp}_{正}-S^{\sharp}_{负} \tag{6}$$

其中,S 表示候选情感词的情感倾向值,S 大于 0 表示其是正情感词,S 小于 0 表示其是负情感词。

存在一种特殊情况:若词 $C_1\cdots C_i\cdots C_n$ 中的某个字 C_i 在语料中的情感词中未出现,则 $P(C_i \mid S^*)$

为 0，所以此时要进行数据平滑，表示如式(7)。

$$P(C_i \mid S^*) = \frac{P(C_i, S^*)}{P(S^*)} = \frac{n(c_i, s^*)}{n(s^*)} \approx \frac{n(c_i, s^*) + \delta}{n(s^*) + \delta \cdot \text{字总数}} \quad (7)$$

δ 取较小的数值，本文取为中文汉字总数的倒数。式(7)表示给任意一个字 C_i 的词频加一个很小的值，避免词频为 0 整体为 0 的现象，从而影响实验分析的准确性，以及某些候选情感词的选择。

通过式(7)推导，可计算出每个候选情感词的情感权重，且为准确数值，这与以往假定的正情感词为 +1，负情感词为 −1 不同。此外，不同语料，情感词不同，情感权重也不同，该算法实现了跨领域挖掘相关情感词，也符合大数据时代下数据本身价值的再利用。

根据计算得出的情感倾向值，按其值大小进行排序可得到情感倾向性程度排序表。通过排序表可明确表示情感倾向性程度，值越大，情感倾向性程度也会越大，即情感强度越强。

3.3　算法过程

根据上述模型，通过计算语料中情感字的分布，挖掘潜在情感词并对其极性做判别。情感词典构建系统框图如图 1 所示。

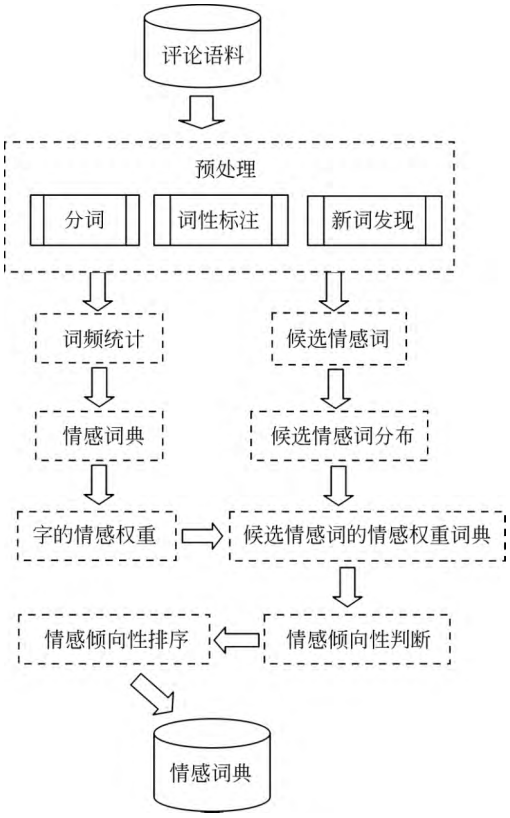


图 1　情感词典构建系统框图

4　实验

本文通过对语料进行机器学习训练，分别计算语料正、负情感字的概率；依据计算出的字的情感概率计算候选情感词的情感权重，判别其情感倾向性以得到情感词，加入情感词典，实现情感词的自动发现及情感极性判别。

结合上述模型，本文进行三组实验验证上述方法的有效性。

4.1　实验设计

本文分别针对三组不同领域的数据做了情感词的自动发现及情感极性判别的实验，实验数据信息如下。

(1) 京东 THINKPAD 的评论数据，大小为 16 MB，共包含 4 000 条正面评论信息和 4 000 条负面评论信息。本文只根据评论数据挖掘潜在未知情感词，所以语料的正、负面信息对本实验没有影响。

(2) 豆瓣 700 部电视剧的评论数据，大小为 65 MB，其中每部电视剧都包含一定的评论信息，并以文本的形式进行存储。

—大众点评网的餐饮业评论数据，大小为 407M，内容包括用户 ID、店家 ID、评论内容及时间等信息。在本文的实验过程中仅对评论内容作分析，其余部分不做处理。

上述三组数据没有对其进行任何人工的标注，通过实验从上述三组生语料中挖掘潜在情感词，并计算该方法的准确性。

此外，本文实验过程中采用的情感词典以台湾大学的情感词典 NTUSD 为依据，共包含 2 810 个正面词语和 8 276 个负面词语。

以上评论数据和情感词典都可通过数据堂获得。

由于本实验是针对生语料进行情感词的挖掘，为验证该方法的准确性，我们采取如下措施：根据上述模型与方法，计算出现在语料中已知情感词的情感权重，这些情感词的情感倾向性在情感词典中是给定的。然后，根据计算出的情感倾向性，结合已知的情感倾向性来判断正确性。

4.2　实验结果与分析

实验一：封闭实验

本文选取词性为形容词、名词、动词的未知情感

词为候选情感词，三组实验的实验结果正确率如表1所示。

表1 实验结果

正确率/%＼候选词	京东	豆瓣	大众点评
形容词	91.67	90.11	90.35
名词	93.62	92.78	92.81
动词	92.41	89.23	90.57

从表1可以看出，实验结果的正确率比较高，但未达到作者的期望值。原因在于，实验过程中存在一种情况：若词 $C_1 \cdots C_i \cdots C_n$ 中的某个字 C_i 在语料中的情感词中未出现，则 $P(C_i/S^*)$ 为0，所以要进行数据平滑。

实验二：平滑后的封闭实验

在进行数据平滑之后，本文分别对上述三组实验数据重新做了一组候选情感词极性是形容词的实验，实验结果如表2所示。

表2 平滑后的实验结果

正确率/%＼候选词	京东	豆瓣	大众点评
形容词	95.76	91.04	92.56

从表2可以看出，在进行数据平滑之后，实验结果的正确率有了明显的提高。

在这里，我们选择形容词为实验的候选情感词，是因为在表达情感或对某一事件的看法时，形容词是主要的表达情感的词汇。所以，本文只选择了形容词进行上述实验，同时其结果也是有代表性的。

4.3 实验三：开放实验

在取得良好的实验结果后，我们发现上述实验是针对整批语料进行的，训练集和测试集并未分开，所以实验结果可能拟合很好，从而带来较高的正确率，因此该实验结果不具有很强的说服力。为此，本文接下来的工作是将豆瓣语料分成了训练集和测试集两部分重新进行上述实验，同样选择最具有代表性的形容词为候选情感词进行实验，验证实验效果。实验结果如表3所示。

表3 豆瓣实验结果

正确率/%＼候选词	豆瓣
形容词	90.73

对比前两组实验结果，由表3可以看出，正确率稍有回落，但是幅度不大，正确率仍在90%以上。以上结果都足以说明该方法的有效性。

为验证实验结果的正确率，本文还采取了另一种实验思路：由于计算结果中含有情感权重值，能够判断候选词的正、负极性，因此可对实验结果人工标注正、负极性，判断实验计算结果是否与人工标注的一致，从而计算该实验的正确率。表4给出了本文计算出的部分候选情感词。

表4 部分候选情感词

候选情感词及情感权重	
正面候选情感词	负面候选情感词
纯熟 62.0463	身败名裂 −31.776453
松懈 34.670124	没羞 −34.058186
骁勇 31.303799	粉身碎骨 −35.578373
美目盼兮 30.47678	繁复 −35.631252
原汁原味 28.78999	头破血流 −35.768387
朗朗上口 26.98067	支离破碎 −36.279156
光彩照人 26.518188	五雷轰顶 −39.63511
素雅 26.5076	狗血喷头 −40.09977
丰沛 26.117752	漏洞百出 −43.665276
求贤若渴 22.281967	两败俱伤 −46.865143
美若天仙 22.09887	怪里怪气 −47.15445

由表4可看出，每个计算出的候选情感词，都有一个情感权重，情感权重值的大小代表其本身在这批语料中的情感强度。情感权重值有正有负，值大于0表示其是正面候选情感词，小于0表示其是负面候选情感词。那么，根据计算出的情感权重值判断该词的极性，再与人工标注的情感极性做对比，从而判断出该实验结果的正确率。

本文对计算出的候选情感词按其情感权重值的大小进行排序，从中计算出正、负情感词top10、top100、top300、top500的正确率，具体结果表示如表5所示。

表5 正确率统计情况

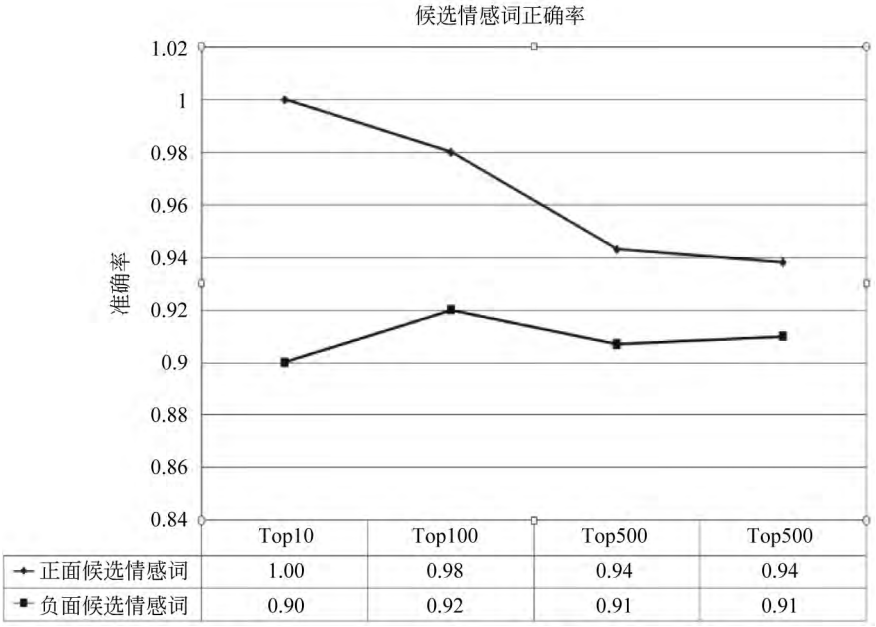
数据统计	正情感词正确数目	正情感词正确率/%	负情感词正确数目	负情感词正确率/%
Top10	10	100	9	90.00
Top100	98	98	92	92.00
Top300	283	94.33	272	90.67
Top500	469	93.8	455	91.00

上述结果具体如图2所示。

由图2可以看出，无论正面候选情感词还是负面候选情感词，其正确率随着候选词数目的增加都

趋近于 92%～93%，这表明其整体的正确率都基本在 90%以上，充分证明了该方法的有效性。

上述实验可以得出，本文提出的情感词自动挖掘算法是有效的，并且能够取得良好效果。但在本实验过程中还发现一种情况：若是某个字 C_i 在语料中的正、负情感词中都没有出现，用上述平滑方法进行平滑时 $n(C_i, S^*)$ 为 0，$n(S^*)$ 小的一方占优势，即语料中正面情感词典、负面情感词典哪个词典规模越小，其 $P(C_i \mid S^*)$ 越大。这对某些词的情感倾向性的判断是不利的，所以这种情况下可回退到根据已知情感词典进行其情感倾向性的判断。这将是我们接下来要进行的情感词挖掘方面的工作。



	Top10	Top100	Top500	Top500
正面候选情感词	1.00	0.98	0.94	0.94
负面候选情感词	0.90	0.92	0.91	0.91

图 2　候选情感词正确率

5　总结与展望

本文基于贝叶斯原理及大数据挖掘，通过计算语料中正、负情感字的概率，得出候选情感词的情感权重，并判别其情感倾向性以得到情感词。通过本方法能够挖掘未知的情感词，准确率能够达到 90%以上，实现情感词的自动发现与极性判别，避免人工整理情感词典的工作，节省人力、物力，大大扩展了情感词典；且本方法与应用领域无关，具有很好的扩展性，奠定了情感分析工作的基础。但由于本方法没有考虑上下文语境的影响，所以接下来的工作重点将对情感词进行外部条件的限制，利用其位置信息、上下文信息进行情感判断，以弥补上述工作的不足，取得更好的效果。

参考文献

[1] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848.

[2] Zhang Jianfeng, Xia Yunqing, Yao Jianmin. A review towards microtext processing [J]. Journal of Chinese Information Processing, 2012, 26(4): 21-27.

[3] Yang Aimin, Zhou Yongmei, Lin Jianghao. A method of Chinese texts sentiment classification based on Bayesian algorithm [J]. Applied Mechanics and Materials, 2013, (263/266): 2185-2190.

[4] Lin Jianghao, Yang Aimin, Zhou Yongmei, et al. Classification of microblog sentiment based on naïve Bayesian [J]. Computer Engineering and Science, 2012, 34(9): 86-90.

[5] Ren Yong, Kaji N, Yoshinaga N, et al. Sentiment classification in resource-scarce languages by using label propagation[C]//Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25), Singapore, 2011: 420-429.

[6] Escalante H J, Montes-Y-Gómez M, Solorio T. A weighted profile intersection measure for profile-based authorship attribution[C]// Proceedings of the 10th Mexican International Conference on Artificial Intelligence (MICAI '11). Berlin, Heidelberg: Springer-Verlag, 2011: 232-243.

[7] Jung J J. Maximum entropy-based named entity recognition method for multiple social networking services [J]. Journal of Internet Technology, 2012, 13(6):

931-937.

[8] **朱嫣岚,闵锦,周雅倩等.基于** HowNet **的词汇语义倾向计算**[J].**中文信息学报**,2006,20(1):14-20.

[9] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Seattle, WA, USA. 2004:168-177.

[10] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10), Valletta, Malta, 2010: 2200-2204.

[11] Hamouda A, Marei M, Rohaim M. Building machine learning based senti-word lexicon for sentiment analysis [J]. Journal of Advances in Information Technology, 2011, 2(4): 199-203.

[12] **阳爱民,林江豪,周咏梅.中文文本情感词典构建方法**[J].**计算机科学与探索**,2013,7(11):1033-1039.

[13] J Turney Peter. Thumbs up or thumbs down?semantic orientation applied to unsupervised classification of review[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 417-424.

[14] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation[J]. Computational Linguistics, 2011(37): 9-27.

张华平(1978—),**副研究员,博士,硕士生导师,主要研究领域为大数据搜索与挖掘、自然语言处理、社交网络等。**

E-mail: kevinzhang@bit.edu.cn

李清敏(1988—),**硕士,主要研究领域为大数据搜索与挖掘、自然语言处理等。**

E-mail: lqm2011@126.cm

李恒训(1985—),**硕士,主要研究领域为大数据搜索与挖掘、自然语言处理。**

E-mail: dereklee1985@126.com