Journal of Computer Applications
计算机应用,2014,34(10):2869-2873
ISSN 1001-9081
CODEN JYIIDU
2014-10-10
http://www.joca.cn

文章编号:1001-9081(2014)10-2869-05
doi:10.11772/j.issn.1001-9081.2014.10.2869

基于词性标注序列特征提取的微博情感分类

卢伟胜,郭躬德*,陈黎飞
(福建师范大学 数学与计算机科学学院,福州 350007)
(*通信作者电子邮箱 ggd@fjnu.edu.cn)

摘 要:传统的 *n*-gram 文本特征提取方法会产生高维度的特征向量,高维数据不但增大了分类的难度,同时也会增加分类的时间。针对这一问题,提出了一种基于词性(POS)标注序列的特征提取方法,根据词性序列能够代表一类文本的这一个特点,利用词性序列组作为文本的特征以达到降低特征维度的效果。在实验中,词性序列特征提取方法比 *n*-gram 特征提取方法至少提高了9%的分类精度,降低4816个维度。实验结果表明,该方法能够适用于微博情感分类。

关键词:特征提取;词性;标注序列;微博情感分类;极性分类
中图分类号:TP391.1 **文献标志码**:A

Emotion classification with feature extraction based on part of speech tagging sequences in micro blog

LU Weisheng, GUO Gongde*, CHEN Lifei
(*School of Mathematics and Computer Science, Fujian Normal University, Fuzhou Fujian* 350007, *China*)

Abstract: Traditional *n*-gram feature extraction tends to produce a high-dimensional feature vector. High-dimensional data not only increases the difficulty of classification, but also increases the classification time. Aiming at this problem, this paper presented a feature extraction method based on Part-of-Speech (POS) tagging sequences. The principle of this method was to use POS sequences as text features to reduce feature dimension, according to the property that POS sequences can represent a kind of text. In the experiment, compared with the *n*-gram feature extraction, the feature extraction based on POS sequences at least improved the classification accuracy of 9% and reduced the dimension of 4816. The experimental results show that the method is suitable for emotion classification in micro blog.

Key words: feature extraction; Part-Of-Speech (POS); tagging sequence; microblog emotion classification; polarity classification

0 引言

随着新浪微博的出现,以及随后的腾讯微博、搜狐微博、网易微博等各种微博平台的崛起,微博已经融入到了人们的日常生活中,许多人把逛微博当成每天必做的一件事。同时,微博上有许多公众人物、政府机关等具有舆情影响能力的用户,他们发表的每条微博都能通过其他用户的转发快速地传播出去。许多微博用户的粉丝达到了几百万甚至达到了千万的级别。他们发表的微博在社会舆论的形成与引导上具有重大的影响。社会上许多舆论风波都源于微博,如何对微博上这些影响力较大的用户进行舆论关注,及时了解或掌握相关微博可能的舆情导向,成为一个十分重要的问题。

对于那些影响力较大的微博用户,可以对其发表的每条微博进行正向情感和负向情感分类,并且对那些发表具有负向情感的微博用户给予重点的关注。因此,对于微博的情感分类研究有助于对微博平台进行舆情监控,能够及时发现具有煽动性、危害性的负面微博,实现有效的舆情控制。此外,情感分类还可以对经济指标[1]以及股市指标[2]进行预测。

对于文本特征向量,一般都采用 *n*-gram[3](*n* 一般取值为1,2,3,依次对应 unigram,bigram,trigram)进行特征提取,对应的特征值计算方法有词频(Term-Frequency, TF)以及词频-逆向文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)[4]等,通过这些特征提取方法对微博进行特征提取的时候,往往会得到维数非常高的特征向量,甚至达到成千上万维。高维数据不但会影响算法的执行效率,同时也会增加分类的难度。此外还有考虑文本相关特点的特征提取方法,比如以标点符号出现的次数、正负向词出现的次数等作为文本的特征,这类特征往往是对 *n*-gram 的一种补充,需要与 *n*-gram提取出的特征进行组合[5]。

相对于基于词汇的 *n*-gram 特征提取方法,本文提出了一种基于词性标注序列组的特征提取方法,目的在于用词性标注序列组概括文本的内容,相比用词性出现的次数,利用词性对文本进行筛选(比如删除语气词等)的方法与文本内容结合更紧密,能够保存更多的文本信息,其主要思想是利用词性序列能够代表某一类文本的特征这一个特点,以词性序列组作为文本的特征表示,以达到降低特征表示维度的效果。

收稿日期:2014-04-29;**修回日期**:2014-06-12。 **基金项目**:国家自然科学基金资助项目(61175123)。
作者简介:卢伟胜(1990-),男,福建漳州人,硕士研究生,主要研究方向:数据挖掘、人工智能; 郭躬德(1965-),男,福建龙岩人,教授,博士,主要研究方向:数据挖掘、机器学习; 陈黎飞(1972-),男,福建长乐人,副教授,博士,主要研究方向:数据挖掘、模式识别。

1 相关工作及背景知识

关于情感分类，Pang 等[6]在2002年所作的研究掀起了情感分类的一段热潮。通常有两种方法[7]，如下所示。

一种是构建情感词典，然后根据词典通过特定的算法模型进行情感倾向值的计算，进而根据情感倾向值进行分类。比如朱嫣岚等[8]提出了基于 HowNet 词汇语义相似度和语义相关场的情感词极性计算方法。有关 HowNet 提供的词汇相似性度量得到了较多学者的使用。王振宇等[9]将 HowNet 与 PMI(Point Mutual Information)进行结合得出了改进的词语情感极性计算方法。张成功等[10]则构建了一个全面、高效的包含基础词典、领域词典、网络词词典以及修饰词词典的极性词典，并且在此基础上提出了一种基于极性词典的情感分析方法，得到了较好的分类效果。Turney 等[11]将单词的语义倾向计算方法与统计信息相结合得出一种新的语义倾向计算方法，其可处理的词性有形容词、副词、名词，以及动词。

另一种方法则为传统的文本分类方法，通过构建情感文本语料集，首先将文本转化为特征向量，然后再使用传统的分类算法进行训练分类。Davidov 等[12]以 twitter 里面的标签和表情符作为类标签，不需要人工对微博的类别进行标注，同时尝试用 n-gram，punctuation-based 等多种特征提取方法的组合对微博进行特征提取，然后使用 kNN-like 分类算法对微博进行情感分类，最终得到的多情感(大于两种)分类的效果较差，而二分类效果较好(即判断某条微博是否含有某种情感)。庞磊等[13]利用表情符号、表情图片以及情绪词来对未标注的样本进行自动标注，自动标注的准确率接近90%，然后以 unigram 和 bigram 作为特征构建多种分类器对微博进行情感分类。唐慧丰等[14]对不同特征提取方法以及不同分类器作了相关的实践研究。

词性是指作为划分词类的根据的词的特点，主要包括两大类：一类是实词(名词、动词、形容词等)，另一类是虚词(副词、介词、连词等)。对于一条微博“我是水手”经分词后得到的单词可能为“我”“是”“水手”，这3个单词分别都有其对应的词性序列(Part-Of-Speech tagging，POS)，比如“我”对应的是代词(/r)，“是”是动词(/v)，“水手”是名词(/n)。这3个词按其对应的顺序连接起来可以得到一个词性序列(/r/v/n)，该词性序列往往能够代表一类文本的特征，比如“我爱西瓜”“我恨冬天”等文本其对应的词性序列都与“我是水手”的词性序列一致。

关于词性的研究更多的是解决如何对单词进行更准确的词性标注，比如文献[15]针对 twitter 数据提出的词性标注方法，实验中的准确率接近90%。在分类上，Wilks 等[16]将 POS 与单词结合在一起，用来作低级别的词性歧义消除。Turney[17]使用词性序列作为模板，从原句中抽取出与对应词性序列模板相匹配的短语，然后再利用无监督的分类方法对文本进行情感分类，实验准确率较高，其主要利用词性序列保留一些对情感分类比较有效的短语。利用词性序列可以将文本中认为可能比较有用的信息保留，剔除比较不重要的信息，比如形容词和动词往往比一些名词在情感分类应用中更有辨识度，因此词性序列对情感分类还是有帮助的。

此外，基于序列的特征提取在其他领域也得到了广泛的应用，比如氨基酸序列特征提取，DNA(DeoxyriboNucleic Acid)序列特征提取等，相关的特征提取有熵密度、n-OCC[18]、伪氨基酸组成成分特征提取方法[19]等。

2 基于词性序列特征提取的微博情感分类

根据词性序列能够代表一类文本的这一个特点，本文提出了一种基于词性序列组的特征提取方法，用词性序列来筛选文本内容，再对剩余的文本内容进行特征权重的计算，忽略掉了部分与词性序列不匹配的文本内容，从而降低文本特征向量的维度。主要工作包括词性序列组的自动选取算法、词性序列对应的特征值计算方法以及情感语料的自动获取方法的设计。该方法从微博训练集中自动获取词性序列组，以该词性序列组里的每个词性序列作为微博的特征。针对每条微博，利用词性序列特征值计算方法求出词性序列组中每个词性序列特征的特征值，将这些特征值组成一个特征向量。最后以转换后的特征向量作为实例，通过传统的分类方法对微博进行情感分类。

2.1 形式化定义

设微博为一个由多个单词组成的有序单词集合 $W = \{w_1, w_2, \cdots, w_n\}$，其中 w_i 为第 i 个单词，微博对应的词性标注设为有序集合 $P_W = \{t_1, t_2, \cdots, t_n\}$，其中 t_i 表示微博第 i 个单词对应的词性标注。

定义1 词性序列(简称 POSeq)。由至少一个词性标注排列组成的有序序列。

定义2 词性序列长度。词性序列中词性标注的总数。

定义3 k 元词性序列(简称 k-POSeq)。词性序列长度为 k 的词性序列。

定义4 词性序列组(简称 POSeq-Set)。由若干个互不相同的词性标注序列组成的集合。

对于微博“今天很开心”其经过带标注的分词后得到了4个词以及对应的词性标注，分别为“今天/t”“我/rr”“很/d”“开心/a”(/t,/rr,/d,/a 为它们的词性标注)，/t，/t/rr，/t/rr/d/a 都可以称为 POSeq，其中/t 为 1-POSeq，/t/rr 为 2-POSeq，/t/rr/d/a 为 4-POSeq，而集合 {/t,/t/rr,/t/rr/d/a} 则可称为 POSeq-Set。

2.2 词性序列组自动获取算法

作为微博特征的词性序列组自动获取算法的基本思想是从训练集中提取出频率较高的词性序列并以这些词性序列组成一个词性序列组。通过设定参数 k 以及参数 p 来分别限定要抽取的词性序列长度以及要提取的前 p 个频率最高的词性序列。

算法的基本过程：给定一个 k 值集合 $K = \{k_1, k_2, \cdots, k_m\}$ 以及参数 p，对于集合 K 中的每个 k_i 值，分别提取出出现频率最高的前 p 个 k_i-POSeq，并组成一个词性序列组，最终会得到 m 个词性序列组，再将这 m 个词性序列组合并成一个词性序列组。

算法2用来抽取每个微博可能出现的 k-POSeq 组成的词性序列组，算法1用来获取作为微博特征表示的词性序列组。

算法1 词性序列组构造算法。

输入：$\{W_1, W_2, \cdots, W_s\}$ 为 s 条微博；$K = \{k_1, k_2, \cdots, k_m\}$ 为 k 值集合；参数 p。

输出：词性序列组。

第1步 初始化 $set = \{\}$，$i = 1$；

第 2 步　对于 k_i,利用算法 2 求得每条微博的所有可能的 k_i-POSeq;

第 3 步　统计各个 k_i-POSeq 出现的频率,将频率最高的前 p 个 POSeq 加入到集合 set 中;

第 4 步　$i = i + 1$;如果 $i < m$ 跳转到第 2 步;

第 5 步　return set;

算法 2　微博的 k-POSeq 抽取算法。

输入:$W = \{w_1, w_2, \cdots, w_n\}$ 为微博;$P_W = \{t_1, t_2, \cdots, t_n\}$ 为微博的词性标注;参数 k。

输出:词性序列组。

```
第 1 步　set = {};
第 2 步　For i = 1:n - k + 1
           // 创建一个词性序列 seq
           seq = {t_{i+0}, t_{i+1}, t_{i+2}, ..., t_{i+k-1}};
           // 将 seq 加入到 set 中
           set += seq;
         End
第 3 步　// set 包含该微博所有可能的 k-POSeq
  return set;
```

2.3　词性序列特征值计算方法

给定 k-POSeq 以及一条微博 W,首先找出微博中与 k-POSeq 相匹配的单词序列(可能没有也可能有多个),以单词序列和对应的正负向情感语料作为输入,利用式(1)计算出每个单词序列对应的权重,以这些权重的和作为最终的词性序列特征值,若没有相匹配的词序列则对应的特征值为 0。

比如给定一个 3 元词性序列{/r/v/n},以及微博"我爱西瓜 我爱冬瓜"(微博对应的词性标注为"/r/v/n/r/v/n"),微博中与该 3 元词性序列相匹配的单词序列有两个("我爱西瓜"和"我爱冬瓜"),得到对应的特征值为 V("我 爱 西瓜",C_1, C_2) + V("我 爱 冬瓜",C_1, C_2);

$$V(W, C_1, C_2) = \frac{P(W, C_1) - P(W, C_2)}{P(W, C_1) + P(W, C_2)} \tag{1}$$

其中:$W = \{w_1, w_2, \cdots, w_n\}$ 为有序单词集合,C_1 表示在训练集中正向情感语料对应的单词集合,C_2 表示训练集中负向情感语料对应的单词集合。

$$P(W, C) = \sum_{i=1}^{n} p(w_i, C) \tag{2}$$

$$p(w, C) = \begin{cases} -\mathrm{lb}(N_w / N_c), & w \in C \\ -\mathrm{lb}(\beta), & w \notin C \end{cases} \tag{3}$$

其中:N_w 表示单词 w 在集合 C 中出现的次数;N_c 表示集合 C 中单词的总数;β 为自定义参数,实验中取值为 0.1。

式(2)意在求出文本属于类别 C 的概率,前提假设是每个词出现的概率是相互独立的,同时考虑到多个概率相乘可能导致乘积太小,因此引进用对数的累加来替换概率的连乘。式(3)为每个单词在对应类别中的"概率"计算方法,其中对于未出现过的单词给定一个默认出现的概率 β。算法 3 为具体的特征值计算算法。

算法 3　词性序列特征值计算算法。

输入:$s = \{t_1, t_2, \cdots, t_k\}$ 为 k 元词性序列;$W = \{w_1, w_2, \cdots, w_n\}$ 为微博;$P_W = \{t_1, t_2, \cdots, t_n\}$ 为微博的词性标注;C_1 为正向语料单词集合;C_2 为负向语料单词集合。

输出:特征值。

```
第 1 步　value = 0
第 2 步　for i = 1:n - k + 1
           seq = {t_{i+0}, t_{i+1}, t_{i+2}, ..., t_{i+k-1}};
           if seq = s then
             W_sub = {w_{i+0}, w_{i+1}, w_{i+2}, ..., w_{i+k-1}};
             // 利用式(1) 对 W_sub 求值
             tmp = V(W_sub, C_1, C_2);
             value = value + tmp;
           end
         end
第 3 步　return value
```

2.4　微博情感分类

2.4.1　分类流程

基于词性序列特征提取的微博情感分类流程主要分为分类器构建和待分类样本分类两部分:

1)分类器构建。首先,从带有类别标签的训练集中选出作为微博特征的词性序列组;然后,将训练集里的每一条微博根据词性序列组转化成对应的特征向量;最后,利用这些特征向量构建分类器。

2)待分类样本分类。将待分类样本根据在分类器构建阶段得到的词性序列组转化为对应的特征向量,然后利用已经构建好的分类器对其进行分类。

2.4.2　时间复杂度分析

本文提出的分类算法的时间复杂度主要包含三个部分,分别是分词以及词性标注的时间复杂度、词性序列特征提取的时间复杂度以及分类器的时间复杂度。主要对词性序列特征提取的时间复杂度进行分析,词性序列特征提取包含多个步骤,为了便于分析,假定训练极性中的微博单词总数为 l,k 值只取一个,那么 k 元词性序列特征提取的时间复杂度为 $O((l-k)*k)$;频繁词性序列选取主要操作在于排序上,所以其时间复杂度为 $O(h \log h)$,其中 h 为排名不超过 p 的 h 个词性序列,即 $h \leqslant p$。特征值计算的时间复杂度为 $O((l-k)*k*k)$,训练集单词频度的统计可以使用哈希表进行存储,因此时间复杂度为 $O(l)$。综上,词性序列的特征提取时间复杂度为 $O((l-k)*k) + O(h \log h) + O((l-k)*k*k) + O(l)$,在实际运行中,因为 k 值为用户设定的参数,h 值不大于参数 p,而且它们一般都比较小,可以把它们看成常数,因此词性序列特征提取的时间复杂度可以简化为 $O(dl)$,其中 d 为常数,且与 k 和 p 的取值有关,所以总体效率还是很高的。

3　实验结果与分析

实验中需要用到 NLPIR 汉语分词系统 2013 版(http://ictclas.nlpir.org/)对微博进行分词和词性标注,并且使用 LIBSVM[20] 作为支撑向量机(Support Vector Machine, SVM),同时通过腾讯开放平台(http://dev.t.qq.com/)的开放接口来获取相关的微博数据。实验中所采用的机器配置环境为:CPU 为 Intel Core i3-2350M 2.30 GHz,内存 4 GB,Windows7 操作系统,Eclipse 集成开发环境,JDK1.7(Java development kit v1.7)。

3.1　情感语料自动获取方法

微博平台可供用户在发表微博的时候设定相应的心情,以腾讯微博平台为例,主要有 5 种心情状态可供用户选择,分别为"狂喜""偷乐""无感""伤心""咆哮",如图 1 所示(每种心情状态可以有多种配图),同一条微博只能有一种心情状态。心情设定是非必须的,多数用户在发表微博的时候并没有去设定,所以有设置心情标签的用户一般发表的内容都会

与对应的心情标签相关。因此,可以利用该心情状态来获取相关的微博情感语料。

通过腾讯微博开放平台提供的时间线接口(statuses/public_timeline)可以收集发表在微博广播大厅的微博数据。返回的数据中包含 emotiontype 字段,其取值为{0,1,2,3,4,5},依次对应"未设定心情""狂喜""偷乐""无感""伤心""咆哮"这 6 种状态。将字段值为 1 的微博作为正面情感数据,字段值为 5 的微博作为负面情感数据。实验中一共用到 2052 条不重复的微博,其中表示"狂喜"的微博数量为 1069,表示"咆哮"的微博数量为 1083(数据共享链接:http://pan.baidu.com/s/1o6qbXzk)。

从数据集中随机抽取 100 条微博进行人工分类,其中属于"狂喜"状态的微博有 87 条是属于正面情感,属于"咆哮"状态的微博有 85 条属于负面情感,虽然存在点噪声,但是其可靠性还是较高的,因此可以此作为实验数据使用。相比用表情符号来自动获取情感语料,通过心情状态来获取情感语料也是一种新的微博情感语料获取途径。

图 1 腾讯微博的心情设定图例

3.2 实验结果分析

实验共分为 3 部分:第一和第二部分用来研究参数 k 和参数 p 对分类效的果影响,第三部分是将词性序列特征提取方法与 n-gram 的特征提取方法在分类精度上进行了比较。从 3.1 节获取的 2153 条微博语料中抽取 236 条微博作为待分类样本,其余的作为训练样本。

实验 1 采用 LIBSVM 作为分类器,固定 k 值集合 $K=\{1,2,3,4\}$,p 分别取 1 到 25 中的奇数,最终得到的分类结果如图 2 所示。从图 2 中可以看出,当固定 k 值集合后,随着参数 p 取值的变化,分类的精度有所波动,但是总体上是比较稳定的。当 p 大于某个值后,等价于将所有的词性序列作为特征,因为词性序列的个数是有限的。当 p 大于某个值后,也就意味着之前不管 p 再怎么增大,分类精度也不会再改变。

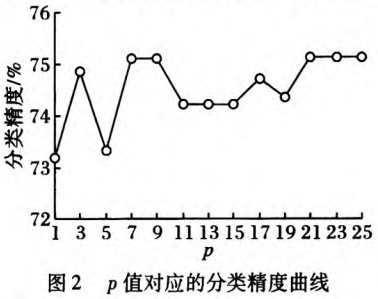

图 2 p 值对应的分类精度曲线

实验 2 以 LIBSVM 作为分类器,固定 p 值为 7,分别取 13 个集合 k 值集集合,并编号成 1 到 13 组,分别对应集合:{1},{1,2},{1,2,3},…,{1,2,3,4,5,6,7,8,9,10,11,12,13}。最终得到的分类精度如图 3 所示,图 3 中的横坐标分类代表 k 值集合对应的组序号。从第 6 个集合开始,后面几个集合的 k 值对应的分类精度是一样的,通过对第 13 个 k 值集合提取得到的特征向量的观察,发现里面存在了很多特征值为 0 的特征。分析得当 k 值越大,以其对应的 k-POSeq 作为特征,微博中与该 k-POSeq 相匹配的词性序列就会越少,计算得到的特征值为 0 的可能性就越大。比如在实验中存在一个出现频率较高的 9-POSeq 为"/w/v/n/w/v/n/w/v/n",其对应的一个微博语句为"/呲牙/呲牙/呲牙",其中"/呲牙"被分词系统分为成了 3 个词,分别是"/""呲""牙",对应词性标注为"/w/v/n",即为 3 个连续的表情符号(腾讯微博以/+标签的形式来表示一个表情符号)。一般很少有微博中存在与这个词性序列相匹配的文本内容,所以计算的特征值为 0。因此,当 k 值过大的时候,近似于给每一个实例增加一维特征值为 0 的特征,所以对分类效果影响不大。

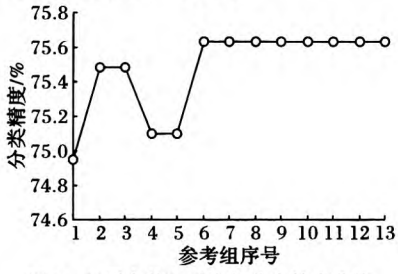

图 3 不同参数组对应的分类精度曲线

实验 3 综合实验 1 和实验 2 的分析,为了方便表示,本文提出的特征提取算法简记为 POSE,其中参数设定为 $K=\{1,2,3,4\}$,$p=7$。实验加入了 unigram,bigram,unigram + bigram 这 3 种特征提取方法,并且以 tf 作为它们的权重。采用 LIBSVM 以及 kNN 分类器来对提取后的特征向量进行分类,其中 kNN 中的近邻个数设定为 5。由于实验的预处理部分做的比较少,所以需要对 n-gram 系列所提取的特征进行过滤,只保留频率排名在区间为[5%,95%]的特征(因为频率高的词很可能是停止词,同时频率少的词也可能是噪声,实验过程中得出限定频率区间比不设定频率区间的效果要来得好)。实验结果如表 1 所示。从表 1 可以得出采用基于词性序列特征提取方法有较好的分类效果,并且大大减少了特征维数。单独使用 bigram 的效果并不理想,但是将 unigram 和 bigram 结合成一个特征的时候反而能得到更好的分类效果。同时当维数过多的时候必须采用稀疏矩阵进行存储,否则就会造成堆的溢出,这也说明了维数过多会增加算法设计的复杂度。

表 1 不同特征提取在两种分类器上的效果以及得到的维数

特征提取方法	分类精度/%		特征维数
	LIBSVM	kNN	
POSE	75.10	69.84	28
unigram	64.66	61.34	4844
bigram	50.00	50.00	26853
unigram + bigram	66.10	63.59	33557

综合上述实验结果,基于词性特征序列组的特征提取是有效的,使用词性序列组作为微博的特征,不但能够减少特征向量的维度,降低分类复杂性,同时也具有较好的分类精度,较高的分类速率。

4 结语

与 n-gram 相比,本文提出的基于词性序列的特征提取方

法具有较少的特征维度也有较好的分类效果,可用于微博的情感分类。此外,根据微博心情设定来收集微博情感语料也是可行的。特征提取只是传统情感分类过程中的一个子模块,在分类精度上可以改进的地方还很多,比如提升微博语料的质量、与其他特征相结合、改进分类算法等。对于词性序列特征提取方法的进一步的工作可以考虑将该特征提取方法拓展到多类别的文本分类中。

参考文献:

[1] LEVENBERG A, PULMAN S, MOILANEN K, *et al*. Predicting economic indicators from Web text using sentiment composition [J]. International Journal of Computer and Communication Engineering, 2014, 3(2): 109 - 115.

[2] ZHANG X, FUEHRES H, GLOOR P A. Predicting stock market indicators through twitter "I hope it is not as bad as I fear" [J]. Procedia-Social and Behavioral Sciences, 2011, 26: 55 - 62.

[3] CAVNAR W B, TRENKLE J M. *N*-gram-based text categorization [C]// Proceedings of the 1994 3rd Annual Symposium on Document Analysis and Information. Las Vegas: [s. n.], 1994: 161 - 175.

[4] JONES K S. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28(1): 11 - 21.

[5] KOULOUMPIS E, WILSON T, MOORE J. Twitter sentiment analysis: the good the bad and the OMG! [C]// Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Palo Alto: AAAI Press, 2011: 538 - 541.

[6] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? sentiment classification using machine learning techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002, 10: 79 - 86.

[7] ZHAO Y, QIN B, LIU T. Sentiment analysis [J]. Journal of Software, 2010, 21(8): 1834 - 1848. (赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834 - 1848.)

[8] ZHU Y, MIN J, ZHOU Y, *et al*. Semantic orientation computing based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1): 14 - 20. (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] WANG Z, WU Z, HU F. Words sentiment polarity calculation based on HowNet and PMI [J]. Computer Engineering, 2012, 38(15): 187 - 189. (王振宇, 吴泽衡, 胡方涛. 基于 HowNet 和 PMI 的词语情感极性计算[J]. 计算机工程, 2012, 38(15): 187 - 189.)

[10] ZHANG C, LIU P, ZHU Z, *et al*. A sentiment analysis method based on a polarity lexicon [J]. Journal of Shandong University: Natural Science, 2012, 47(3): 47 - 50. (张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法[J]. 山东大学学报: 理学版, 2012, 47(3): 47 - 50.)

[11] TURNEY P, LITTMAN M. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315 - 346.

[12] DAVIDOV D, TSUR O, RAPPOPORT A. Enhanced sentiment learning using twitter hashtags and smileys [C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg: Association for Computational Linguistics, 2010: 241 - 249.

[13] PANG L, LI S, ZHOU G. Sentiment classification method of Chinese micro-blog based on emotional knowledge [J]. Computer Engineering, 2012, 38(13): 156 - 158. (庞磊, 李寿山, 周国栋. 基于情绪知识的中文微博情感分类方法[J]. 计算机工程, 2012, 38(13): 156 - 158.)

[14] TANG H, TAN S, CHENG X. Research on sentiment classification of Chinese reviews based on supervised machine learning techniques [J]. Journal of Chinese Information Processing, 2007, 21(6): 88 - 94. (唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88 - 94.)

[15] GIMPEL K, SCHNEIDER N, O'CONNOR B, *et al*. Part-of-speech tagging for twitter: annotation, features, and experiments [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011, 2: 42 - 47.

[16] WILKS Y, STEVENSON M. The grammar of sense: using part-of-speech tags as a first step in semantic disambiguation [J]. Natural Language Engineering, 1998, 4(2): 135 - 143.

[17] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 417 - 424.

[18] LUO R, FENG Z, LIU J. Prediction of protein structural class by amino acid and polypeptide composition [J]. European Journal of Biochemistry, 2002, 269(17): 4219 - 4225.

[19] CHOU K, CAI Y. Predicting protein structural class by functional domain composition [J]. Biochemical and Biophysical Research Communications, 2004, 321(4): 1007 - 1009.

[20] CHANG C, LIN C. LIBSVM: a library for support vector machines [EB/OL]. [2014-01-26]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(上接第 2868 页)

[8] CHENG L, HE P, SUN Y. Study on Chinese keyword extraction algorithm based on naive Bayes model [J]. Journal of Computer Applications, 2005, 25(12): 2780 - 2782. (程岚岚, 何丕廉, 孙越恒. 基于朴素贝叶斯模型的中文关键词提取算法研究[J]. 计算机应用, 2005, 25(12): 2780 - 2782.)

[9] ZHANG Y, LIU T, WEN X. Modified Bayesian model based question classification [J]. Journal of Chinese Information Processing, 2005, 19(2): 100 - 105. (张宇, 刘挺, 文勖. 基于改进贝叶斯模型的问题分类[J]. 中文信息学报, 2005, 19(2): 100 - 105.)

[10] TONG B. Introduction to the theory of journalism [M]. Beijing: China Renmin University Press, 2002: 118 - 223. (童兵. 理论新闻传播学导论[M]. 北京: 中国人民大学出版社, 2002: 118 - 223.)

[11] YAN T W, GARCIA-MOLINA H. Index structures for information filtering under the vector space model [C]// Proceedings of the 10th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 1994: 37 - 47.

[12] LI G, CHEN C, LI Z, *et al*. Automatic Web structured data extraction based on tag path [J]. Computers Science, 2013, 40(6A): 141 - 145. (李贵, 陈成, 李征宇, 等. 基于标签路径的 Web 结构化数据自动抽取[J]. 计算机科学, 2013, 40(6A): 141 - 145.)