文章编号: 1006-2475(2016) 09-0100-04

基于文本过滤的贝叶斯分类算法的改进

路金泉,徐开勇,戴乐育

(解放军信息工程大学,河南 郑州 450000)

摘要: 针对传统贝叶斯分类算法无法满足复杂网络文本过滤需求,提出一种多词-贝叶斯分类算法(Multi Word-Bayes, MWB) 。该算法一方面引入了特征权重(Term Frequency-Inverse Document Frequency,TF-IDF) 的计算思想,优化了传统贝叶斯分类算法只考虑词频不考虑文本间关系的问题; 另一方面将词与词间的关系作为文本分类的重要参考项,克服了传统贝叶斯分类算法在分类器训练上对语义分析的忽视。实验结果表明,MWB 在垃圾文本过滤上具有更好的分类性能。

关键词: 贝叶斯分类算法; TF-IDF; 语义分析; 文本过滤
中图分类号: TP311 文献标识码: A doi: 10.3969/j.issn.1006-2475.2016.09.022

Improvement of Bayes Classification Algorithm Based on Text Filtering

LU Jin-quan, XU Kai-yong , Dai Le-yu

(The PLA Information Engineering University, Zhengzhou 450000, China)

Abstract: As the complexity of the network, traditional Bayes classification algorithm cannot meet the demand of text filtering. Multi Word-Bayes (MWB) classification algorithm is proposed. On the one hand, Term Frequency-Inverse Document Frequency (TF-IDF) feature weight is introduced in MWB algorithm to optimize the traditional Bayes algorithm which only considers the problem of word frequency, but doesn't consider the relationship between the texts. On the other hand, the new algorithm views the relationship between the word and the word as an important reference, which overcomes the traditional Bayes classification algorithm ignoring the semantic analysis on the classifier training. Experiment results show that MWB classification algorithm is of better classification effect on the text filtering.

Key words: Bayes classification algorithm; TF-IDF; semantic analysis; text filtering

0 引 言

根据中国互联网络信息中心(CNNIC) 统计,截至 2015 年 6 月,中国网民数达到了 6.68 亿人,互联网普及率达到 48.8% [1] 。互联网为人们各抒己见提供了平台,但其中不乏有一些国内外敌对势力利用网络的公开性,大肆宣扬国外文化,并对我国历史人物,英雄事迹进行恶意的扭曲,用以瓦解我国民众建立起来的价值观 [2-3] 。攻击形式主要是文本,据统计网络文本总量约占网络总体内容的 80% [4] 。攻击者利用中文语义的多义性,避开了传统的信息监测技术的监测; 而针对大量的网络文本如果依靠人工力量的话无疑会消耗大量的人力、财力。因此基于语义的中文文本过滤技术成为了学术界的一个研究热点。

近年来,国内外学者们做了大量的研究,也获得了大量的研究成果。X. Luo [4] 等人提出针对中文的多特性性质用于文本分类中,指出将 3 个字以上的组合作为一个单位,并引入了权重因子。许珂 [5] 等人提出了使用信息增益值与 TF-IDF 值的乘积作为特征权重的值以将语义相关度作为信息熵,产生了较好的效果。马兆才 [6] 提出两阶段处理文本,增加了文本分类的准确度。邓一贵 [7] 等人提出一种新的针对网页内容的敏感信息过滤算法(SWDT-IFA) ,不再依赖于词典等,而是构建敏感词决策树。黄贤英 [8] 等人针对微博特有的短文本特性提出对动、名词赋予不同权重的方式来达到内容的精确分类。综上虽然考虑到了语义分析在文本分类中的重要性,但针对文本内部词与词之间的相关性没有考虑,而词与词之间的相关性对文本的主题思想具有重要影响。针对历史人物的刻意扭曲的例子如 "和珅一生清正廉明,却被

收稿日期: 2016-03-22
作者简介: 路金泉(1991-) ,男,山西临汾人,解放军信息工程大学硕士研究生,研究方向: 信息安全; 徐开勇(1963-) ,男,研究员,博士,研究方向: 信息安全,可信计算; 戴乐育(1990-) ,男,助教,研究方向: 信息安全,密码协处理器。

后人视为贪官”。基于上述方法无法对类似句子做一个准确的判断。本文将类似(刻意扭曲历史事实)文本视为垃圾文本,其余文本视为正常文本。

本文针对文本分类这种特有的缺陷,对传统的贝叶斯分类算法提出改进,提出多词-贝叶斯算法(MWB)。本文采用中国科学院计算所开发的汉词词法分析系统对文本进行预处理,考虑中文词与词间的影响对文本语义的重要性,引入词与词的概率公式,进而对传统贝叶斯算法进行改进。

1　相关理论研究

在介绍 MWB 算法之前首先要对文本过滤过程作一个简单的了解,其包括: 语料库选择、文本预处理、数据库建设、过滤器训练、数据库更新和新文本过滤。文本自动过滤流程图如图 1 所示。

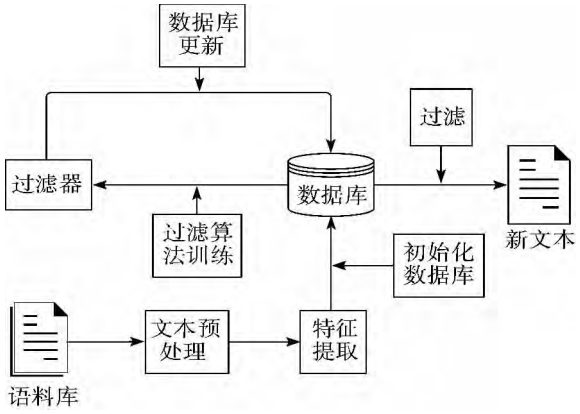

图 1　文本自动过滤流程

1.1　文本预处理

本文采用中国科学院计算所开发的汉词词法分析系统,该系统可以实现中文分词、词性标注等。实践证明该系统分词的正确率可以高达 97.8%,分词与词性标注处理速度为 543.5 kB/s[9]。利用该系统的高准确率以及高速率对语料库文本进行分词与词性标注,将所得结果做去噪声操作,去除如连词(如: 和、或等)、助词(如: 的、啊等)等。

文献[5]对经典特征选择方法 TF-IDF 的改进方式引入了语义分析的思想,能较准确地给出特征词权重,其公式为:

$$\text{weight}(i) = TF \cdot IDF \cdot \overline{IG_c(C,T)} \tag{1}$$

其中,weight(i) 表示词语 i 出现的权重; TF 表示词频; IDF 表示反文本频率; $\overline{IG_c(C,T)}$ 表示词 i 的信息增益值。

通过特征权重的计算去除那些权重极小可以忽略不计的因子,保留较大权重因子,进一步降低数据数量。

1.2　数据库建设

文本预处理结果所得到的是划分好的词语,以及该词项属于哪篇文章,其在文章中出现的次数以及该词与文本之间的从属关系。而依照贝叶斯分类算法需求,只需要知道某词出现的频率,以及该词属于垃圾文本还是正常文本,因此数据库建设思想如下:

设每一类文本对应一个 Hash 表,则定义 Hashtable_good 表内存储正常文本; Hashtable_bad 表内存储垃圾文本; 每个词对应一个 Token 串。存储内容即为 Token 串与词频的映射关系。

通过计算各类 Hash 表中 Token 串出现的频率,结合贝叶斯公式,可得某词出现在垃圾文本中的概率,具体计算过程参见 1.3 节。进而可以得到新的 Hash 表,其存储内容可以表示为: Token→p(B/wi),其中 B 表示垃圾文本,wi 表示某个词; 该公式即为: Token 串对应的词出现垃圾文本中的概率。

1.3　过滤器训练

过滤器训练的核心是对文本分类算法的训练,大家所熟知的文本分类算法有: 朴素贝叶斯算法(Naive Bayes)[10]、k 最近邻分类算法[11]、支持向量机(Support Vector Machine,SVM)[12]、神经网络算法[13]、决策树分类算法[14]等,本文采用贝叶斯分类算法对垃圾文本进行过滤。

文本 d 属于类别 c 的条件概率 p(c|d)[15],其计算公式如下:

$$p(c|d) = \frac{p(c) \times p(d|c)}{p(d)} \tag{2}$$

公式(2)为典型的贝叶斯公式。

若把文本 d 看作一组词组成,即 $d = \{w_1, w_2, \cdots, w_n\}$,其中词与词之间相互独立,任取词 w_i 与 w_j($i \neq j$),其相互独立,根据独立间关系有:

$$p(c) \times p(d|c) = p(c) \times p(w_1, w_2, \cdots, w_n|c) = p(c) \prod_{i=1}^{n} p(w_i|c) \tag{3}$$

其中,p(c) 为属于某类文本总数占训练集的总文本数的比值,$p(w_i/c)$ 表示词 w_i 在类 c 下出现的次数与类 c 中包含词语总数的比值,n 表示文本中所含的词语总数,w_i 表示文本中任一词语。

根据独立性的假设,文本 d 属于类别 c_1 的概率表示如下,其中 n 为词 i 的数量:

$$p(c_1|d) = \frac{p(c_1) * p(d|c_1)}{p(d)} = \frac{p(c_1) \prod_{i=1}^{n} p(w_i|c_1)}{\sum_{j=1}^{2} p(c_j) \prod_{i=1}^{n} p(w_i|c_j)} \tag{4}$$

其中,j 表示类别数,分别为正常文本、垃圾文本; 分母即表示常数 p(d) 的全概率公式。公式(4)中假设的是词与词之间是独立的,而现实当中词与词之间是不

独立的,因此,这种方式只分析了词间独立的情况,而没有考虑文本自身的含义。

2 MWB 算法

考虑到贝叶斯算法只能适用于词与词之间相对独立的情况,且概率计算忽略了文本间的相关性问题; 本文通过基于语义分析的思想,结合贝叶斯算法与传统词频计算方法,提出一种改进算法——MWB 算法。

2.1 文本间关系引入

从 1.3 节获知,计算 $p(c|d)$ 的核心是计算 $p(w_i/c)$。而对于中文文本,其处理对象是离散的,利用传统的概率计算方式表示为 $p(w_i/c)=\frac{n_{w_i}}{N_c}$,$n_{w_i}$ 表示词 w_i 出现在类 c 中的次数,N_c 表示类 c 中所含的所有词语总数。显然这种方式能够计算某个词出现在某类时的概率,但它的类比情况是与出现在该类的其他词作比较,而非是其自身的概率分布情况。因此,这种方式不适合用作于垃圾文本的过滤。

针对垃圾过滤思想,结合传统算法没考虑词在文本内部的分布情况以及类别间的概率分布,现给出如下定义:

定义 1 任取词语 w_i 属于垃圾文本的概率是指词 w_i 出现在垃圾文本中的概率之和与词 w_i 出现在语料库各类文本中的概率之和的比值 $p(w_i)$,其表达式如下:

$$p(w_i)=\frac{\sum_{i=1}^{n}\frac{B_i}{N_{B_i}}}{\sum_{i=1}^{n}\frac{B_i}{N_{B_i}}+\sum_{j=1}^{m}\frac{G_j}{N_{G_j}}} \tag{5}$$

其中,B_i 表示词语 w_i 出现在垃圾文本 i 中的次数,N_{B_i} 表示垃圾文本 i 的词总数; G_j 表示 w_i 出现在正常文本 j 中的次数,N_{G_j} 表示正常文本 j 中的总词数。

通过这种方式得到的结果将词语出现在垃圾文本与正常文本的情况考虑进去,将侧重点转移到垃圾文本过滤中,更有针对性,解决了传统贝叶斯公式不考虑文本间关系的问题。

2.2 词与词间关系引入

分析完单个词语出现时的概率表达公式之后,引入语义相关性概念; 如上所述 “和珅一生清正廉明,却被后人视为贪官”。对句子划分并去多余项得到: 和珅/一生/清正廉明/后人/视为/贪官。经验判断,对于“和珅”与“清正廉明”相关性不高,其出现在垃圾文本中的概率应该较大。考虑到词与词的这种相关性,可得出如下定义。

定义 2 任取词 w_i 与词 w_j(其中 $i\neq j$),其同时出现在垃圾文本中的概率是指其同时出现在垃圾文本中的概率之和与其同时出现在语料库中不同类别文本中的概率之和的比值 $p(wi\cdot wj)$,其表达式如下:

$$p(w_i\cdot w_j)=\frac{\sum_{x=1}^{n}\frac{N_{B_x}(w_i\cdot w_j)}{N_{B_x}}}{\sum_{x=1}^{n}\frac{N_{B_x}(w_i\cdot w_j)}{N_{B_x}}+\sum_{y=1}^{m}\frac{N_{G_y}(w_i\cdot w_j)}{N_{G_y}}} \tag{6}$$

其中,$N_{B_x}(w_i,w_j)$ 表示 w_i 与 w_j 出现在垃圾文本 x 中的数目,$N_{G_y}(w_i,w_j)$ 表示 w_i 与 w_j 出现在正常文本 y 中的数目,N_{B_x} 表示 x 文本中所含的词语总数,N_{G_y} 表示 y 文本中所含的词语总数。

同理,当 3 个词同时出现时可得定义如下:

定义 3 任取词 w_i、词 w_j 与词 w_k(其中 $i\neq j\neq k$)同时出现在垃圾文本中的概率是指其同时出现在垃圾文本中的概率之和与其同时出现在语料库中不同类别文本中的概率之和的比值 $p(w_i\cdot w_j\cdot w_k)$,其表达式如下:

$$p(w_i,w_j,w_k)=\frac{\sum_{x=1}^{n}\frac{N_{B_x}(w_i,w_j,w_k)}{N_{B_x}}}{\sum_{x=1}^{n}\frac{N_{B_x}(w_i,w_j,w_k)}{N_{B_x}}+\sum_{y=1}^{m}\frac{N_{G_y}(w_i,w_j,w_k)}{N_{G_y}}} \tag{7}$$

2.3 贝叶斯算法改进

基于上述思想通过对训练集的训练,得出单个词语或者是多个词语的条件概率,即其在垃圾文本中出现的概率。当新文本出现时,判断该文本为垃圾文本的概率有如下定义。

定义 4 词项是指从文本中提取出来的,之间相互独立的个体,可以是单个词语也可以是 2 个或者 2 个以上词语; 记为: $u_i(i=1,2,\cdots,n)$,其中 n 表示文本所含词项个数。u_i 属于垃圾文本的条件概率记为 $p(u_i)$。

定义 5 文本 d 属于垃圾文本的概率是指文本 d 所含词项属于垃圾文本的概率之积与出现 u_i 的概率积的总和的比值。记为: $p(C_{bad}/d)$,其表达式如下:

$$p(C_{bad}|d)=\frac{\prod_{i=1}^{n}p(u_i)}{\prod_{i=1}^{n}p(u_i)+\prod_{i=1}^{n}1-p(u_i)} \tag{8}$$

在文本训练阶段要做好概率阈值的选取,使得当所求概率值大于阈值时为垃圾文本,小于阈值时为正常文本。综上,通过对贝叶斯分类算法的改进,将其运用到文本过滤系统中去,通过对词语相关性的分析,提高了对文本自身语义分析的准确性,进而可以提高文本过滤的准确性。

2.4 算法实现过程

根据斯坦福的《Introduction to Information Re-

trieval》[17] 所说,多项式模型计算准确率更高。即以单词为粒度计算其条件概率所得分类结果准确率更高。文本过滤器的核心即是对文本分类算法的训练,而 MWB 算法核心即是对词语属于垃圾文本的条件概率进行计算。因此算法的实现也是围绕这个核心进行的。

算法实现过程如下:

1) 语料库选择。选取具有代表性的文本库,对文本进行预处理,得出"类别—词语—在该文本中出现频率"对应关系 f(x) 。

2) 初始化数据库。建立 Hash 表,即 Hashtable_good 表示正常文本类别,Hashtable_bad 表示垃圾文本类别,将 f(x) 转化为数学表达。

3) 词项划分。根据实际需求将词语划分为单个、2 个或者 3 个一组,划分原则要求所选词语同时出现在一个文本中。

4) 数据库更新。通过计算所选词项在训练集语料库中的概率,即利用公式(5) ~公式(7) 计算所选词项属于垃圾文本的概率,调整所属类别词项频率。

5) 整合 Hashtable_good 与 Hashtable_bad 中数据,计算文本属于垃圾文本的概率阈值。

6) 根据实际需求引入新的语料库,转步骤 2。

7) 新文本测试。对新文本进行预处理,选取文本 10 个以上词项,利用公式(8) ,计算其属于垃圾文本概率,并与概率阈值作比较。

8) 算法结束。

整个算法流程可以归纳为 3 个大步骤: ①准备阶段(步骤 1 ~3) ; ②算法训练阶段(步骤 4 ~6) ; ③应用阶段(步骤 7) 。通过对训练语料库的不断调整,进而调整所得条件概率,使得算法更适用于复杂多变的网络环境。

3　实验结果与分析

3.1　语料选择

本文采用数据堂提供的新浪微博积极、消极微博数据[13] 进行分析,选取其中的 2000 条长微博,其中包含 950 条正常微博、950 条违背事实的垃圾微博作为训练集,另 100 条作为测试集。

3.2　评价标准

文本分类系统采用准确率和召回率来衡量分类的准确性和完整性。本文虽是以文本过滤作为出发点,但过滤的前提是要把文本分为垃圾文本和正常文本。因此,本文仍是采用准确率和召回率来作为衡量的标准。

3.3　实　验

3.3.1　文本预处理

将文本的预处理分为以下 3 个步骤:

1) 采用中国科学院计算所开发的汉词词法分析系统(ICTCLAS) [9] 对长微博进行分词划分,并对其划分结果进行词性标注;

2) 将标注结果进行噪声去除,除去那些冗余的无意义词语,降低数据处理范围;

3) 运用上述改进的 TF-IDF 方法,求特征词权重,将权重极小值去除,进一步降低噪声干扰。

3.3.2　过滤器训练

本文对计算精度有较高的要求,而贝叶斯算法本身所计算出来的结果会产生较长的尾数,进而超过双精度浮点数的范围,即出现溢出。本文采用 Java 中的 BigDecimal 类去定义浮点数,增加了处理精度。将改进后的贝叶斯算法进行训练,分别对一个词出现的情况,2 个词共同出现时的情况,以及 3 个词共同出现时的情况分别训练。最后将训练结果用于文本测试。

3.3.3　结果分析

通过上述实验产生的结果如表 1 所示。

表 1　学习 1900 个文本后的实验结果

测试方式	准确率/%	召回率/%
单个词语	80.3	81.2
2 个词语	80.1	80.8
3 个词语	82.3	83.4
传统贝叶斯	79.5	76.4

从实验产生的结果可以看出,当训练文本数较小时,单个词语分类的结果要比 2 个词语分类结果准确性更高,而 3 个词语所产生的分类效果最好。因此将词语间的相关性加入到无论是文本分类还是文本过滤都有很高的必要性,不再依赖于关键词过滤,避免很多有意义的文本过滤掉,同时防止一些以次充好的文本流进来。

4　结束语

本文通过对传统的贝叶斯算法提出改进,利用 TF-IDF 权重计算思想对垃圾文本的概率求解提供新思路; 同时将词与词之间的相关性加入到文本过滤系统中去,增加了文本过滤的精度。实验结果表明,改进后的贝叶斯算法在文本过滤的准确性上得到了提高。但在实验过程中发现前期的文本预处理阶段消耗的时间较长,其处理时间与后期的文本分类时间相当,这大大增加了文本过滤的时间量,不能很好适应网络文本处理要求。因此,下一阶段的主要研究方向是降低文本预处理所消耗的时间量。

(下转第 108 页)

参考文献:
[1] 胡水星. 大数据及其关键技术的教育应用实证分析[J]. 远程教育杂志, 2015(5):46-53.
[2] 徐鹏,王以宁,刘艳华,等. 大数据视角分析学习变革——美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J]. 远程教育杂志, 2013(6):11-17.
[3] 马红亮,袁莉,郭唯一,等. 反省分析技术在教育领域中的应用[J]. 现代远程教育研究, 2014(4):39-46.
[4] 顾小清,郑隆威,简菁. 获取教育大数据:基于 xAPI 规范对学习经历数据的获取与共享[J]. 现代远程教育研究, 2014(5):13-23.
[5] Siemens G. Learning and Knowledge Analytics—Knewton - the future of education? [EB/OL]. http://www.learninganalytics.net/?p=126, 2011-04-14.
[6] NMC Horizon Repor 2011 Higher Ed Edition [EB/OL]. http://www.nmc.org/publications/horizon-report-2011-higher-ed-edition, 2011-01-10.
[7] 李艳燕,马韶茜,黄荣怀. 学习分析技术:服务学习过程设计和优化[J]. 开放教育研究, 2012,18(5):18-24.
[8] 吴青,罗儒国. 学习分析:从源起到实践与研究[J]. 开放教育研究, 2015,21(1):71-79.
[9] 胡艺龄,顾小清,罗九同,等. 教育效益的追问:从学习分析技术的视角[J]. 现代远程教育研究, 2014(6):41-48.
[10] 孟玲玲,顾小清,李泽. 学习分析工具比较研究[J]. 开放教育研究, 2014,20(4):66-75.
[11] 胡艺龄,顾小清,赵春. 在线学习行为分析建模及挖掘[J]. 开放教育研究, 2014,20(2):102-110.
[12] 葛道凯,张少刚,魏顺平. 教育数据挖掘:方法与应用[M]. 北京:教育科学出版社, 2012:69.
[13] 魏顺平. 学习分析技术:挖掘大数据时代下教育数据的价值[J]. 现代教育技术, 2013,23(2):5-11.
[14] 魏顺平. Moodle 平台数据挖掘研究——以一门在线培训课程学习过程分析为例[J]. 中国远程教育, 2011(1):24-30.
[15] 吴青,罗儒国,王权于. 基于关联规则的网络学习行为实证研究[J]. 现代教育技术, 2015,25(7):88-94.
[16] 百度百科. "Weka"词条[EB/OL]. http://baike.baidu.com/view/1380214.htm, 2016-03-22.
[17] 郁晓华,顾小清. 开放教育下的学习分析——2015 AECT 夏季研讨会评述与延伸[J]. 远程教育杂志, 2015(5):14-23.
[18] 李青,王涛. 学习分析技术研究与应用现状述评[J]. 中国电化教育, 2012(8):129-133.
[19] 魏顺平,韩艳辉,王丽娜. 基于学习过程数据挖掘与分析的在线教学反思研究[J]. 现代教育技术, 2015,25(6):89-95.

(上接第 103 页)

参考文献:
[1] 中国互联网信息中心. 中国互联网信息中心第 36 次中国互联网络发展状况统计报告[R/OL]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf, 2015-07.
[2] 徐健锋,许圆,许元辰,等. 基于语义理解和机器学习的混合的中文文本情感分类算法框架[J]. 计算机科学, 2015,42(6):61-66.
[3] 石海明,曾华峰. 科技与战争视角下的国家认知空间安全战略[J]. 国防科技, 2014,35(3):83-87.
[4] Luo Xi, Ohyama Wa, Wakabayashi T, et al. Improvement of automatic Chinese text classification by combining multiple features[J]. Transactions on Electrical and Electronic Engineering, 2015,10(2):166-174.
[5] 许珂,蒙祖强,林启峰. 基于语义关联和信息增益的 TFIDF 改进算法研究[J]. 计算机应用研究, 2012,29(2):557-560.
[6] 马兆才. 文本分类中的两阶段特征降维[J]. 甘肃科技, 2014,30(20):27-29.
[7] 邓一贵,伍玉英. 基于文本内容的敏感词决策树信息过滤算法[J]. 计算机工程, 2014,40(9):300-304.
[8] 黄贤英,陈红阳,刘英涛,等. 一种新的微博短文本特征词选择算法[J]. 计算机工程与科学, 2015,37(9):1761-1767.
[9] 中国科学院计算技术研究所. 汉语词法分析系统 ICTCLAS [EB/OL]. http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/200706/t20070628_2121143.html, 2016-03-21.
[10] 郑炜,沈文,张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]. 西北工业大学学报, 2010,28(4):622-627.
[11] 张宁,贾自艳,史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程, 2005,31(8):171-172.
[12] 赵辉. 支持向量机分类方法及其在文本分类中的应用研究[D]. 大连:大连理工大学, 2005.
[13] 刘钢. 基于神经网络的文本分类系统 NNTCS 的设计和实现[D]. 北京:中国科学院(软件研究所), 2003.
[14] 张青. 决策树分类算法的研究与改进[D]. 郑州:郑州大学, 2002.
[15] Otsuka T, Deng Deyue, Ito T. Text filtering for harmful document classification using three-word co-occurrence and large-scale data processing[J]. Electronics and Communications in Japan, 2015,98(10):168-175.
[16] Guo Xiaoli, Sun Huiyu, Zhou Tiehua. SAW classification algorithm for Chinese text classification[J]. Sustainability, 2015,7(3):2338-2352
[17] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval[M]. 王斌,译. 北京:人民邮电出版社, 2010:175-199.
[18] 数据堂. 新浪微博积极、消极、矛盾微博数据[EB/OL]. http://www.datatang.com/data/47209, 2015-05-07.