Computer Science and Application 计算机科学与应用, 2019, 9(12), 2314-2322
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912257

Hans汉斯

Text Emotional Analysis Based on Hybrid Information Gain Algorithm

Yuqiang Li, Zhiyong Hong, Jinghui Chen

Wuyi University, Jiangmen Guangdong
Email: lyq19950723@gmail.com, hongmr@163.com, 754631011@qq.com

Received: Nov. 22nd, 2019; accepted: Dec. 5th, 2019; published: Dec. 12th, 2019

Abstract

Aiming at the problem of selection bias in the traditional information gain feature selection method and the problem of word frequency between different categories without considering the feature frequency of different elements, a text sentiment analysis algorithm with mixed information gain is proposed. By introducing the inverse document frequency coefficient, the inter-class feature word frequency coefficient and the chi-square statistic coefficient, the text data are feature-selected, so that the word frequency information in the entire document, the word frequency information between each class, and the low-frequency word information of important emotional colors are obtained and used efficiently. The experimental results show that the text sentiment analysis method with mixed information gain can effectively improve the quality of feature selection and improve the accuracy of text sentiment analysis, about 2% to 5%.

Keywords

Information Gain, Feature Selection, Emotional Analysis, Chi-Square Statistic, Text Classification

基于混合信息增益算法的文本情感分析

李育强，洪智勇，陈靖辉

五邑大学，广东 江门
Email: lyq19950723@gmail.com, hongmr@163.com, 754631011@qq.com

收稿日期：2019年11月22日；录用日期：2019年12月5日；发布日期：2019年12月12日

摘　要

针对传统信息增益特征选择方法存在的选择偏向性的现象以及未考虑特征元素在不同类别间词频的问题，提出了一种混合信息增益的文本情感分析算法。通过引入逆文档频率系数、类间特征词频系数和卡方统计量系数，对文本数据进行特征选择，使得整个文档中词频信息、每个类之间的词频信息以及重要情感

文章引用: 李育强, 洪智勇, 陈靖辉. 基于混合信息增益算法的文本情感分析[J]. 计算机科学与应用, 2019, 9(12): 2314-2322. DOI: 10.12677/csa.2019.912257

色彩的低频词信息得到有效利用。实验结果表明，采用混合信息增益的文本情感分析方法可以有效地提高特征选择的质量，进而提高文本情感分析的准确率，大约2%~5%。

关键词

信息增益，特征选择，情感分析，卡方统计量，文本分类

1. 引言

随着社交媒体的不断发展，人们对于数据分析的需求越来越大，文本情感分析正好是数据分析中的一大主流研究方向。例如：在看热点新闻时，发表相应的观点；在购物时，对产品反馈相应的评论等等。文本情感分析[1]是指用自然语言处理、文本挖掘、机器学习和计算机语言学等方法对带有情感色彩的主观性文本进行有效数据的挖掘，最后根据这些有效数据做出其文本的情感倾向的判断，所以情感分析也称为意见挖掘。当前，文本情感分析的两大主流方法[1]分别是基于词典和基于机器学习的文本情感分析，其中基于语义词典的方法则先构造情感词词典，借助词典判断情感倾向[2]；而基于机器学习的方法是利用机器学习[3]的各种分类方法来识别情感。在文本情感分析中，文本数据通常以空间向量模型[4] (VSM)的形式表示，通过该模型，文本数据可以转换为结构化数据，用于后续的任务处理。在一般的文本数据集中，特征元素通常会达到成千上万个，如何选取有效的特征元素来描述文本数据，从而提高文本情感分析的效果，就成为分析文本情感的主要问题[5]，特征的选择就成为了文本情感分析的重要组成部分。

在情感分析算法中，特征选择[6]的主要目的是为了减少特征数量、降维，使模型泛化能力更强，减少过拟合，增强对特征和特征值之间的理解以及提高模型的效率。目前常用的特征选择方法[6]为：信息增益(Information gain, IG)，卡方检验(Chi-squared, CHi)，文档频数(Document Frequency, DF)以及互信息(Mutual Information, MI)。其中在信息增益的方法上，郭亚维，刘晓霞等人[7]通过引入新的平衡因子，再结合传统的信息增益算法得到一种新的算法，虽然提高了情感分析的分类效果，但是该因子取值不能确定；李海瑞[8]考虑了各特征选择算法的优缺点，结合了传统信息增益方法、TF-IDF 特征词权重计算方法与信息熵方法，虽然分类有一定的提高，但是在特征情感词计算上存在不足；蒲国林[9]针对于传统信息增益方法，并结合粗糙集的知识，提出一种新方法，主要是利用粗糙集剔除高冗余性的特征，虽然提高了情感特征提取的准确率，但是忽略了不同类别间情感词的关系，同时花费的时间相对较长；龚安等人[10]针对于文本语法不规则、特征稀疏的问题，设计了一种针对评论文本的多特征融合的情感分类算法，但是特征选择算法只是单纯的融合，并未区分其重要度，存在一定的局限性。

针对上述有关信息增益方法存在的局限性，本文通过引入了卡方统计量系数、逆文档频率系数以及类间频率系数，提出一种混合信息增益特征选择方法。通过相关理论分析以及实验证明，该算法能有效地利用卡方统计量信息、逆文档信息以及词频信息提高特征选择的质量，从而提高情感分类的精确度。

2. 背景知识

2.1. 信息熵

熵在信息论中是一个非常重要的概念，说明了空间中任何能量分布的均匀性，能量分布越均匀，越

不确定，熵就越大。shannon [11]在信息处理中引入了熵，并总结出了“信息熵”的概述。信息熵是指信息的量化度量，是对随机变量不确定性的度量。

设 X 为一个取值有限个的离散随机变量，其概率分布为：

$$P\left(X=x_i\right)=P_i,\ i=1,2,\cdots,n\ 。\tag{1}$$

则随机变量 X 的信息熵定义为[12]：

$$H\left(X\right)=-\sum_{i=1}^{n}p\left(x_i\right)\log p\left(x_i\right)。\tag{2}$$

设有随机变量(X, Y)，其联合分布为：

$$P\left(X=x_i,Y=y_j\right)=P_{ij},\ i,j=1,2,\cdots,n\ 。\tag{3}$$

通过观察在已知随机变量 X 的条件下随机变量 Y 的不确定性，在 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望为：

$$H\left(Y\mid X\right)=-\sum_{i=1}^{n}p\left(x_i\right)\sum_{i=1}^{n}p\left(y_j\mid x_i\right)\log p\left(y_j\mid x_i\right)。\tag{4}$$

2.2. 信息增益

根据上节信息熵与条件熵的介绍，信息增益恰好就是信息熵与条件熵的差，并表示在消除不确定性后获得的信息量，一般定义[12]为：

$$\mathrm{IG}\left(Y,X\right)=H\left(Y\right)-H\left(Y\mid X\right)。\tag{5}$$

在机器学习中的信息增益方法(IG)一般定义为：待分类的集合的熵 $H(Y)$和选定某个特征的条件熵 $H\left(Y\mid X\right)$之差。如果计算出来的信息增益值越大，则这个特征的选择性越好，也说明该特征项对于该类就越具有代表性。在文本分类系统中，它通过统计某个特征元素 w 在类别 Class 中出现与否的文档总数来计算特征元素 w 对类别 Class 的信息增益值，也就是系统原本的熵与固定特征 w 后的条件熵之差，其定义式为：

$$\begin{aligned}\mathrm{IG}\left(w\right)&=H\left(C\right)-H\left(C\mid w\right)\\&=-\sum_{i=1}^{m}p\left(c_i\right)\log p\left(c_i\right)+p\left(w\right)\sum_{i=1}^{m}p\left(c_i\mid w\right)\log p\left(c_i\mid w\right)\\&\quad+p\left(\overline{w}\right)\sum_{i=1}^{m}p\left(c_i\mid\overline{w}\right)\log p\left(c_i\mid\overline{w}\right)\end{aligned}\tag{6}$$

上式中：$p(c_i)$表示 c_i 类文本在语料中出现的概率，即 c_i 类文本数除以总的文本数；$p(w)$表示语料库中包含特征元素 w 的文本概率，即包含特征元素 w 的文本数除以总的文本数；$p\left(c_i\mid w\right)$表示文本包含特征元素 w 时属于 c_i 类文本的条件概率，即包含特征元素 w 且属于 c_i 类文本数除以包含特征元素 w 的文本数；$p\left(\overline{w}\right)$ 表示语料中不包含特征元素 w 的文本概率，即不包含特征元素 w 的文本数除以总的文本数；$p\left(c_i\mid\overline{w}\right)$表示文本不包含特征元素 w 时属于 c_i 类的条件概率，即不包含特征元素 w 且属于 c_i 类文本数除以不包含特征 w 的文本数；m 为类别数。

2.3. 卡方统计量

在文本情感分析中，卡方统计量也经常被用来特征选择，基本思想[13]是通过观察实际值和理论值的偏差程度来确定理论的有效性，并度量特征元素 w 和类 c 间的独立性。设 w 为特征元素，包含 w 且属于

类 c 的文档数，记为 A；包含 w 但是不属于类 c 的文档数，记为 B；不包含 w 且属于类 c 的文档数，记为 C；不包含 w 且不属于类 c 的文档数，记为 D。基本元素如表 1 所示：

Table 1. Basic element settings
表 1. 基础元素设置

类别	属于 C 类	不属于 C 类	总数
包含特征 t	A	B	$A+B$
不包含特征 t	C	D	$C+D$
总数	$A+C$	$B+D$	$A+B+C+D=N$

特征元素 w 与类别 c 的卡方统计量一般定义形式如下式(7)所示[14]：

$$\chi^2(w,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{7}$$

其中：$\chi^2(w,c)$ 值越大，表示特征元素 w 与类别 c 相关程度越强。当 $\chi^2(w,c)=0$ 时，特征元素 w 中不包含任何与类别 c 相关的信息。在式(7)中，$A+C$ 项表示属于 c 类的文档总数，$B+D$ 项表示不属于 c 类的文档总数。给定一个训练集和类别，在特征选择中只需考虑特征项在某个特定类别中的卡方值大小顺序，无需考虑具体的值，那么对于同一类的特征项来说，相关性没有影响，因此式(7)可以继续简化，简化后的表示形式[13]如下式(8)所示：

$$\chi^2(w,c)=\frac{(AD-BC)^2}{(A+B)(C+D)}。 \tag{8}$$

3. 基于混合信息增益的文本情感分析方法

一般采用传统的信息增益特征选择算法的情感分析过程为：首先对文本数据进行传统的信息增益特征选择时，然后按照计算得到的信息增益值从大到小排序，根据维度的选取需求，选取其中较大的前 n 项值所对应的特征元素进行文本的向量的表示，最后选择合适的分类方法进行分类器的训练，最终利用得到的分类器进行预测。

根据上述的 1 到 6 式，并结合传统信息增益特征选择算的情感分析过程，可以看出基于传统信息增益算法的整个计算过程只是考虑了特征对整体分类器的影响，并未考虑特征所属的某个类别对整体的影响，是一种“全局”的特征选择算法；如果有的特征对于某一类别的识别度特别高，但是对于其他类别没有什么识别度给分类带来的有效信息则偏少，故此引入类间词频信息系数 α 来进行优化。类间词频信息表示词条 t 在属于某一类 class 中文档 d 中出现的频率。假定特征词集合为 Word，且为 $\{w_1,w_2,w_3,\cdots,w_n\}$，语料集合中类别为 $Class_j\ (1<j<m)$ 的文本数有 $doc_{j1},doc_{j2},\cdots,doc_{jN_j}$，$N_j$ 为 c_j 类的文档总数，特征 $w_i\ (1<i<n)$ 在文本 $doc_{jk}\ (1<k<N_j)$ 中出现的频度为 $tf_{jk}(w_i)$。由于同一个词语在长文件里可能会比短文件有更高的词数[15]，这里采用对每个特征词的词频进行归一化处理，则特征词 w_j 在类别 $class_i$ 中出现的频度 α 表示为：

$$\alpha=\frac{\sum_{k=1}^{N_j}tf_{jk}(w_i)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{N_j}tf_{jk}(w_i)}。 \tag{9}$$

引入类间词频系数避免了传统信息增益没有重视到自身出现某一类文档中的次数的问题，然而，仅使用特征词的频率就会导致一些少的特征元素具有很高的 α 值，如“你”和“是”等词，但这些词并不实际意义。通常情况下，如果特征元素出现在越多的文档中，特征元素对文档的影响较小，那么特征元素区分文档的能力相对较弱。因此，为了增加特征元素的影响力，选出更具代表性的某一类特征元素，故此引入逆文档频率系数 β [16]进行调节，从而解决该问题。逆向文件频率(inverse document frequency, IDF)是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。逆文档频率系数 β 的定义形式为：

$$\beta = \log \frac{N}{f(w_i)+1} \tag{10}$$

其中：N 为训练集中文档的总数，$f(w_i)$为包含特征元素 w_i 的文本数，分母后加上 1，是为了确保分母不为 0，以保证系数的有效性。通过引入逆文档频率系数，可以降低一些常见词作为特征元素对于最终分类的影响，提高特征选择的质量。

在分析信息增益算式时，不难发现此特征选择方法没有负相关性但是它存在一定的偏向性问题。简单来说：在数据样本分布与特征词分布不均匀时，大多数特征词在特定的类别中是不会出现的，此时信息增益值主要由式(6)中特征词不出现情况下的那部分值所决定。这样会使在一个类别中出现次数少然而在其它类中经常出现的特征词被筛选出来，而不倾向于选取在一个类别中出现次数较多而在其它类中出现次数较少的更具代表性的特征词，最终导致信息增益特征选择方法的效果大大降低[17]；如果该特征的每个取值下的样本数非常少且特征词为主要情感词，则会对情感分析结果产生巨大的影响，从而大大减少了情感分类的准确率；故此可以引入 2.3 节中的卡方统计量系数来进行优化，其式如下：

$$\chi^2(w,c) = \frac{(AD-BC)^2}{(A+B)(C+D)}。$$

由于在某个特征的每个取值下的样本数非常少的情况下，传统的信息增益一般偏向于取值较多的特征，而忽略一些取值较少的特征，这样就会导致不良的特征选择的效果；但是由于卡方校验存在“低频词偏袒”的特征，恰巧对信息增益偏向性的问题有一定的帮助；同时为了区分情感特征词与普通的特征词，在选择特征词时，对于特征情感词赋予一定的权重参数 b 来进行调节。

通过对于传统的信息增益特征选择算法的不足，综合上述三个方面的考虑，得到一个混合的信息增益的文本情感分析方法。这种方法的目的是选择出集中出现在某一特定类、并且在该类别中每篇文档中具有代表性的情感特征词。其计算式如下：

$$\mathrm{HIG} = \mathrm{IG} \times (by+1)\left(a\chi^2 + (1-a)\alpha\beta\right) \tag{11}$$

其中：a、b 都为相应的权重系数，且 $0 \le a,b \le 1$，y 取值为 0 和 1，当特征词为情感词时，y 为 1，不为情感词时，则为 0。

在改进的方法中，引入类间频度和逆文档频率测试指标使得计算的信息增益值考虑了词频对特征元素预测能力的影响，更加准确地反映了特征元素分布的比例情况，也使得那部分集中出现在某类文档中而且在该类文档中分布均匀的特征元素获得更高的权重；引入卡方统计量系数，使得低频情感词也得到了有效地利用，这也纠正了一个问题，即传统方法中错误地提高了在一个类别中出现次数不多而在其他类别中经常出现的特征元素的权重的问题。

4. 实验结果分析

本文实验分别采用来自于中国科学院院士谭松波教授的酒店管理评论语料集以及汽车销售评论语料集来进行实验，数据语料集总共 8000 条评论数据，其中积极评论为 4000 条，消极评论为 4000 条。本文分别采用信息增益，卡方统计量以及混合信息增益 3 种不同的特征选择方法，并计算各自的准确率、召回率和 F1 值。

4.1. 实验评价标准

实验实在 Inter(R)Core(TM)i7-7500U CPU @ 2.70GHz 2.90GHz，内存 12 G，硬盘 120 G 和 win10 操作系统下进行的，使用 python 语句编程。为验证算法的有效性，并采用交叉实验的方式，将评论数据中的 80%作为训练集，其余 20%作为测试集。在文本情感分类中普遍使用的评价标准有准确率、召回率以及 F1 值，其中精确率是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本；召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了；F1 值是是统计学中用来衡量二分类模型精确度的一种指标。三种评价标准具体定义式如下：

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + TN} \tag{13}$$

$$Fl = \frac{2 \times P \times R}{P + R} \tag{14}$$

其中，TP 表示被分类器正确分类的正例数据，TN 表示被分类器正确分类的负例数据，FP 表示被错误地标记为正例数据的负例数据，FN 表示被错误地标记为负例数据的正例数据。

4.2. 实验步骤

1) 对数据集进行预处理，对其进行标注，将正负文档进行合并，并对其进行 jieba 分词和去停用词处理。

2) 在分好词的基础上，对其分别进行信息增益(IG)、卡方统计量(CHI)以及混合信息增益(HIG)方法的特征选择。

3) 对于第(2)步结果接着进行特征项表示，运用词袋模型(BOW)进行文本特征表示；并将文档转换为向量，然后对其进行计算分析；其中词袋模型 Bag-of-words (简称 BoW)是一种从文本中提取特征的方法，用于从文档中提取特征，并进行表示。

4) 采用支持向量机的分类方法对训练样本集进行训练以及测试，并对三种特征选择方法实验结果进行分析对比。

4.3. 实验结果

本文采用信息增益、卡方统计量、混合信息增益三种不同的特征选择方式，以及支持向量机的分类算法，确定算法的权重系数，并计算准确率、召回率、F1 值。实验结果如表 2 所示。然而在此算法中，如何确定算法的权重系数极为关键，往往不同的权重系数会导致特征选择出现极大的差距。经过多次实验发现，当 $a = 0.2$ 和 $b = 0.4$，情感分类效果最佳。从表 2 可看出，本文所提出的混合信息增益特征选择方法相对于信息增益，卡方校验特征选择方法在不同维度下情感分类的准确率 P、召回率 R 和 F1 值均有所提高。

Table 2. Comparison of classification results of three feature selection methods
表 2. 三种特征选择方法的分类结果比较

特征维数	IG			HIG			CHI		
	P	R	F1	P	R	F1	P	R	F1
1000	0.75	0.78	0.76	0.79	0.82	0.80	0.72	0.80	0.76
2000	0.83	0.86	0.84	0.86	0.89	0.87	0.82	0.83	0.82
2500	0.88	0.89	0.88	0.89	0.92	0.90	0.87	0.88	0.89
3000	0.88	0.89	0.88	0.89	0.92	0.90	0.87	0.88	0.89

由上表的计算可知，基于混合信息增益的特征选择方法相比于其它两种传统方法在不同维度下分类准确率、召回率和 F1 值均有所提高。图 1 到 3 分别为在不同维度下的准确率、召回率和 F1 值的文本情感分类评估情况。从图 1 中可以看出，HIG 方法的分类准确率远高于其他 2 种特征选择方法。图 2 显示，在特征维数为 2500 时，HIG 方法的召回率最高，当特征维数超过一定数值时，召回率逐渐减少，但是召回率总体大于 0.85，相对其他特征选择方法要高。从图 3 中可以看出，HIG 的 F1 值远高于其他 2 种特征选择方法，这说明 HIG 方法能够提取高质量的特征词，有效区分 2 个类别的文本，提高文本情感分类的准确度。

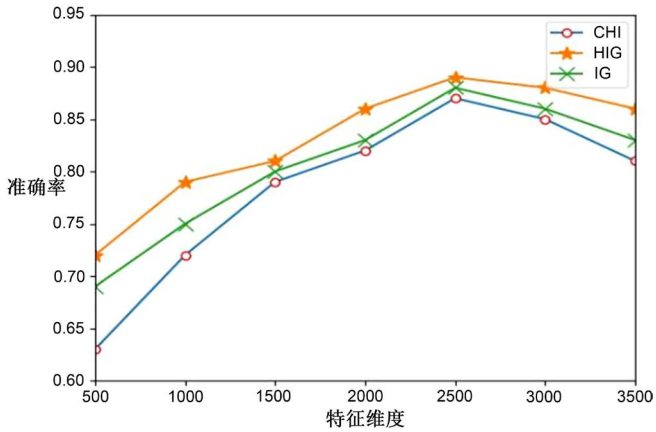

Figure 1. Comparison of classification accuracy of three feature selection methods in different dimensions
图 1. 不同维度下三种特征选择方法分类的准确率对比

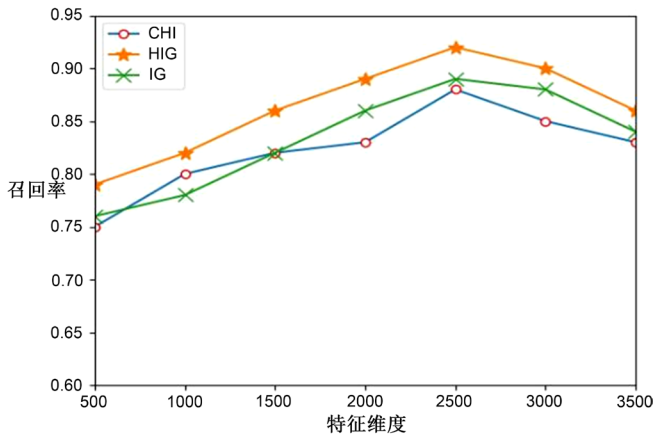

Figure 2. Comparison of recall rates of three feature selection methods in different dimensions
图 2. 不同维度下三种特征选择方法分类的召回率对比

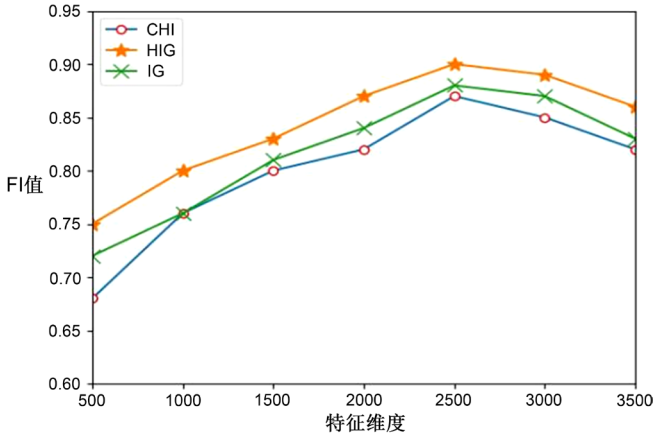

Figure 3. Comparison of F1 values of three feature selection methods in different dimensions
图 3. 不同维度下三种特征选择方法分类的 F1 值对比

5. 结束语

本文在分析传统信息增益特征选择方法的基础上，提出一种基于混合信息增益的特征选择方法，通过引入词频系数、逆文档频率系数以及卡方统计量系数来优化算法，最终实现文本情感分类。从实验结果可看出，本文提出方法在情感分类方面明显优于其它两类特征选择方法。下一步可将本文方法用于分析不同领域的评论语料，提高其在不同类型语料下的分类准确率。

基金项目

本课题获广东省自然科学基金研究项目(2016A030310003)、上海市信息安全综合管理技术研究重点实验室开放课题基金(编号：AGK2018006)资助。

参考文献

[1] Cherry, C. and Mohammad, S. (2012) Binary Classifiers and Latent Sequence Models for Emotion Detection in Suicide Notes. *Journal of Biomedical Informatics Insights*, **5**, 147-154. https://doi.org/10.4137/BII.S8933

[2] 梅莉莉, 黄河燕, 周新宇, 毛先领. 情感词典构建综述[J]. 中文信息学报, 2016, 30(5): 19-27.

[3] Zhai, S. and Zhang, Z.M. (2016) Semi-Supervised Autoencoder for Sentiment Analysis. In: *Thirtieth AAAI Conference on Artificial Intelligence*, AAAI Press, Palo Alto, CA, 1394-1400.

[4] Tang, J. and Zhou, S.G. (2016) A New Approach for Feature Selection from Microarray Data Based on Information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **13**, 1004-1015. https://doi.org/10.1109/TCBB.2016.2515582

[5] Bidi, N. and Elberichi, Z. (2016) Feature Selection for Text Classification Using Genetic Algorithms. 2016 8*th International Conference on Modelling, Identification and Control*, Algiers, Algeria, 15-17 November 2016, 806-807. https://doi.org/10.1109/ICMIC.2016.7804223

[6] Yang, Y.M. and Pederseno, J. (1997) A Comparative Study on Feature Selection in Text Categorization. In: *Proceedings of the* 14*th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, 412-420.

[7] 郭亚维, 刘晓霞. 文本分类中信息增益特征选择方法的研究[J]. 计算机工程与应用, 2012, 48(27): 119-122+127.

[8] 李海瑞. 基于信息增益和信息熵的特征词权重计算研究[D]: [硕士学位论文]. 重庆: 重庆大学, 2012.

[9] 蒲国林. 基于粗糙集与信息增益的情感特征选择方法[J]. 微电子学与计算机, 2016, 33(1): 96-99.

[10] 龚安, 费凡. 基于多特征融合的评论文本情感分析[J]. 计算机技术与发展, 2018, 28(8): 91-95.

[11] 曲炜. 信息论基础及应用[M]. 北京: 清华大学出版社, 2005.

[12] 李航. 统计学习方法第二版[M]. 北京: 清华大学出版社, 2019.

[13] 李平, 戴月明, 王艳. 基于混合卡方统计量与逻辑回归的文本情感分析[J]. 计算机工程, 2017, 43(12): 192-196+202.

[14] 徐明, 高翔, 许志刚, 刘磊. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用, 2014, 50(19): 113-117+142.

[15] 陈东亮, 白清源. 基于词频向量的关联文本分类[J]. 计算机研究与发展, 2009, 46(z2): 839-844.

[16] 马费成, 张勤. 国内外知识管理研究热点——基于词频的统计分析[J]. 情报学报, 2006, 25(2): 163-171.

[17] 邹娟, 周经野, 邓成, 等. 基于多重启发式规则的中文文本特征值提取方法[J]. 计算机工程与科学, 2006, 28(8): 78-79+104.