# 文本分类中 TF-IDF 方法的改进研究

覃世安 李法运
(福州大学公共管理学院 福州 350108)

【摘要】针对 TF-IDF 在待分类文本类的数量分布不均时提取特征值效果差的问题,提出使用特征值在类间出现的概率比代替特征值在类间出现的次数比以改进 TF-IDF 算法。实验证明利用改进后的 TF-IDF 方法提取网页文本特征值,并配合简单累加求和的分类器,使得网页文本分类的准确率有明显提高,且分类速度加快。

【关键词】概率 TF-IDF 网页 文本分类
【分类号】TP391

## Improved TF-IDF Method in Text Classification

Qin Shian Li Fayun
(School of Public Administration and Policy, Fuzhou University, Fuzhou 350108, China)

【Abstract】When the count of one class is much more than another class's, the result of IDF in TF-IDF goes the wrong way according to its design idea. This paper solves the problem by using probability to change TF-IDF algorithm. In the end, the experiment proves that the solution mentioned above is good at classifying webpage text through a simple way to cumulative sum the value of characteristic words and the speed is faster and the accuracy rate is promoted.

【Keywords】Probability TF-IDF Webpage Text classification

## 1 引 言

目前有很多中小网站面临信息发布不及时、发布的信息缺乏时效性等问题。这多是因为网站实力不雄厚,无法及时获得用户所关注领域的信息。针对这种情况,网站设计者采用了使用网络爬虫爬取与本站关注领域相关信息的策略。然而面对互联网中海量的信息,如何对网页文本正确地分类成为一个难题。文本分类是指对未知类别的文档进行自动处理,根据文档内容判断其属于预定义类别集合中的哪一个或哪几个类别[1]。要进行网页文本分类首先要对网页进行预处理,然后提取文本特征词。文本特征词提取运用最广泛的方法是 TF-IDF。

## 2 国内外研究现状

TF-IDF 最早应用在信息检索领域,由于其计算方法简单实用,在业界得到普遍的应用,但是它也存在缺点,国内外学者对其进行了大量的改进工作,主要是围绕 IDF 计算方法的改进。

学者鲁松等[2]认为词语在文本集合中的分布比例量上的差异也是词语表达文本内容的一个重要因素,而 TF-IDF 无法描述这个差异,因此他们提出 tf. idf. IG 方法来解决这个问题。在 tf. idf. IG 中词语的信息增益作为一个文本表示的因子,用来衡量词语在文本集合中分布的量上的差异,通过实验证明 tf. idf. IG 的分类效果优于 TF-IDF 方法。罗欣等[3]将互信息量和信息增益两种思想同时用来改进 TF-IDF 的分类思想,提出基于词频差异的特征选取(WFDBFS)方法。实验结果表明基于词频差异的特征选取方法的分类效果优于单独应用互信息量的分类效果

收稿日期:2013 - 06 - 17
收修改稿日期:2013 - 08 - 29

和单独应用信息增益的分类效果。张保富等[4]针对传统 TF-IDF 方法将文档集作为整体来处理并没有考虑到特征项在类间分布不均的情况,提出一种结合信息熵的 TF-IDF 改进方法。该方法采用结合特征项在类间和类内信息分布熵来调整 TF-IDF 特征项的权重计算,避免了那些对分类没有贡献的特征项被赋予较大权值的缺陷,能更有效地计算文本特征项的权重。

国外学者 Forman[5]用概率统计方法度量并比较关于类别分布的显著性,提出用二元正态分隔(Bi - Normal Separation)计算方法替代 TF-IDF 中的 IDF 计算方法,实验结果表明这种新的 TF-IDF 在文本特征词的选取中表现优异。Lan 等[6]用相关性频率(rf)去重代替传统 IDF 计算方法,给出 rf = log(1 + ni/ni-)计算公式,通过实验证明用 TF-RF 进行文本分类具有更好的区分能力。Oren[7]为了寻找出更好的 TF-IDF 计算方法,利用遗传编码的方式并使用不同适应度函数去寻求一个新的 TF-IDF。利用这种方法进行文本分类准确率有了相应的提高。

虽然还有一些学者用其他的方法改进 TF-IDF 的计算方法,但在实际应用中应用最多的还是以上几种改进方法[5]。但对信息增益、互信息量、信息熵、二元正态分隔、相关性频率等计算量很大,计算复杂度高。针对这种情况,本文提出一种提取文本分类中特征词的简单高效的改进方法,并通过实验证明了该改进方法的有效性。

## 3 TF-IDF 及其改进

### 3.1 TF-IDF 的思想及其存在的问题

TF-IDF[8](词频率 - 逆文档频率)的主要思想是:特征词在文档中的权重为特征词在文档中出现的频数反比于包含该特征词的文档数目。TF 表示特征词 m 在文档 D 中出现的频率,IDF 表示所有文档中出现特征词 m 的文档数目。

常用的计算方法如下:

$$TF = \frac{m}{M} \tag{1}$$

其中,m 表示文档 i 中特征词出现的次数,M 表示文档 i 的总的单词数目。

$$IDF = \log(\frac{N}{n} + 0.01) \tag{2}$$

其中,N 为总文档数,n 为包含某项特征词的文档总数。

$$TF\text{-}IDF = TF \times IDF \tag{3}$$

TF-IDF 是信息检索领域常用的方法,其涵义是如果一个特征词在某个文档中多次出现,且其他的文档包含该特征词较少,则该特征词能够很好地表示该文档。TF 为某一特征值在文档中出现的频数,反映文档的内部特征,IDF 为某一特征值在整个文档集合中的分布情况,反映文档间的特征。

TF-IDF 将文档集作为整体处理,特别是 IDF 的计算,在文本分类中存在明显的缺陷:在公式(2)中令 $n = n_1 + n_2$,$n_1$ 表示 $c_i$ 中包含特征词 m 的文档数目,$n_2$ 表示其他类中包含特征词 m 的文档数目。当 $n_1 >> n_2$ 时,在总的文档数 N 一定的情况下,IDF 的值很小。然而实际情况是特征值 m 在类 $c_i$ 中出现的频率远远大于在其他类中出现的频率,特征词 m 应该有很好的区分能力,而这里却与期望的结果恰好相反。

表 1 $m_1$、$m_2$ 的特征分布

| 类别 | $m_1$ | $m_2$ |
|---|---|---|
| $C_1$ | 9 | 5 |
| $C_2$ | 1 | 5 |

以表 1 为例,有 $C_1$, $C_2$ 两个类,$m_1$,$m_2$ 为两个特征值词。在 $C_1$, $C_2$ 中包含 $m_1$,$m_2$ 特征词的文档数分别为 9 篇、5 篇、1 篇、5 篇。$m_1$,$m_2$ 在类 $C_1$ 中 IDF 值分别为 $IDF_1$,$IDF_2$,则 $IDF_1 = \log(10/9 + 0.01) = 0.0496$,$IDF_2 = \log(10/5 + 0.01) = 0.303$。$IDF_1$,$IDF_2$ 的值表明 $m_2$ 比 $m_1$ 具有更好的分类效果。但是据实际观察 $m_1$ 分布不均匀而 $m_2$ 分布很均匀,表明 $m_1$ 比 $m_2$ 具有更好的类别区分能力。

### 3.2 TF-IDF 方法的改进

现有待分文档类别的集合 $S = \{C_1, C_2, \cdots, C_j\}$,在 $C_i(C_i \in S)$ 中有文档集合 $D = \{d_1, d_2, d_3, \cdots, d_n\}$,n 为文档数目。特征词的集合为 $M = \{m_1, m_2, m_3, \cdots, m_k\}$,k 为 $C_i$ 中所有出现的词语个数。针对 3.1 节提出的问题,在计算 IDF 时可以用特征值在类间出现的概率比代替特征值在类间出现的次数比来解决。根据大数定律[9]的特性,假定某类文档的作者在书写该类文档时用到哪些词组是一个随机事件,因此可以用 $P(m_k)$ 表示词组 $m_k$ 在类 $C_i$ 中出现的概率,$count(m_k)$ 表示词组 $m_k$ 在 $C_i$ 中出现的次数。则:

$$P(m_k)=\frac{count(m_k)}{count(m_1)+count(m_2)+count(m_3)+\cdots+count(m_k)}$$

一个类的文档的特征词应该有很好的代表该类文档的特征信息，用 $P(m_k)$ 表示词组 $m_k$ 在类 $C_i$ 中出现的概率，$P(m_k)'$ 表示词组 $m_k$ 在 S 中除了 $C_i$ 以外的类中出现的概率之和。令：

$$IDF=\log\left(\frac{P(m_k)}{P(m_k)+P(m_k)'}\right)$$

在 $\frac{P(m_k)}{P(m_k)+P(m_k)'}$ 中，当 $P(m_k)$ 很大，IDF 的绝对值反而小，则对它取反，根据 log 函数的特性，自变量要大于 0，IDF 要为正值，最后修正 IDF 得：

$$IDF=-\log\left(1-\frac{P(m_k)}{P(m_k)+P(m_k)'}\right)=\log\left(1+\frac{P(m_k)}{P(m_k)'}\right)$$

令：$TF=P(m_k)$ 表示特征词在某类文本中出现的概率，和公式(1)表示的意义一致。

最后得到改进后的提取某类文本特征值的 TF-IDF 公式如下：

$$TF-IDF=P(m_k)\times\log\left(1+\frac{P(m_k)}{P(m_k)'}\right)\qquad(4)$$

根据公式(4)再次计算表 1 中特征值 $m_1$，$m_2$ 的 IDF。$IDF_{m_1c_i}=\log\left(1+\frac{9/14}{1/6}\right)=0.552$，$IDF_{m_2c_i}=\log\left(1+\frac{5/14}{5/6}\right)=0.155$，$IDF_{m_1c_i}>IDF_{m_2c_i}$，表明词组 $m_1$ 的分类效果好于 $m_2$，这与事实相符。

## 4　实验设计及结果分析

### 4.1　文本表示方法

用改进后的 TF-IDF 方法来提取训练文档集中每个类的特征词，并把特征词按照所计算的 TF-IDF 值从大到小的顺序排列。从有序的特征词集合中选择前 m 个 TF-IDF 值较大的特征值词组成特征向量。对于每一个文档 $d_i$，先进行切词处理，统计文档 $d_i$ 中每个词组出现的次数。对照 $d_i$ 所属的类 $c_i$ 的特征向量进行标记。如果 $d_i$ 中的词组 $word_1$ 在特征向量中出现，则标记为 1，没有出现则标记为 0。

例如文档集合类 $c_i$ 的特征词组成的特征向量为｛汽车、前窗、奥迪｝，文档 $d_i$ 分词后的结果为：“汽车，玻璃，前窗，汽车，颜色，轮胎”，则文本文档 $d_i$ 表示为“101100”的字符串。

### 4.2　文本分类方法

文本分类的方法很多，分类器实现的复杂程度各不相同。为了追求计算的高效性并验证特征词提取的准确性，本文使用文档与文档所属的类的相似度进行分类。文档 $d_i$ 与类 $c_k$ 之间的相似度 $S(d_i,c_k)$ 计算公式如下：

$$S(d_i,c_k)=\sum_{j=1}^{m}v_j\quad(v_j=1\text{ 或 }0)\qquad(5)$$

当文档 $d_i$ 和类 $c_k$ 的相似度值大于 $d_i$ 与其他类的相似度值时，文档 $d_i$ 被划分到类 $c_k$，例如文本文档 $d_1$ 相对于文档类别 $c_1$，$c_2$，$c_3$ 的表示字符串向量依次为：$c_1$：111001，$c_2$：100100，$c_3$：000100，按照公式(5)就有：

$$S(d_1,c_1)=1+1+1+0+0+1=4$$
$$S(d_1,c_2)=1+0+0+1+0+0=2$$
$$S(d_1,c_3)=0+0+0+1+0+0=1$$

$d_1$ 与 $c_1$ 的相似度最大，证明它们的相似程度最高，因此 $d_1$ 被划分到 $c_1$。

### 4.3　评价指标

对于文本分类系统的性能评估测试，国际上有通用的评估指标，包括查全率(Recall)、查准率(Precision)和 F1 评估值三项[10]。本实验是为了测试改进的 TF-IDF 方法对文本分类精度影响，因此，采用查准率作为评价改进后的 TF-IDF 方法提取特征值好坏程度的指标。另外，错误率为查准率的互补，用来检验程序是否正常运行，查准率和错误率之和等于 1。

### 4.4　实验描述及结果分析

数据集来自自编爬虫下载的新闻网页，首先去掉网页中标记，然后按类别储存。数据集一共有 6 个文本类，每类有 4 000 篇文章，依次为汽车、文化、医药、军事、体育和经济。随机在每个类中挑取 100 篇文章作为测试集，其他 3 900 篇文章作为训练集。采用的分词系统是中国科学院计算技术研究所的 ICTCLAS[11]。实验采用 Java 自编程序实现，Java 的 JDK 为 JDK7，运行硬件为装有 XP 系统、主频为 3.3GHz、内存为 4GB 的联想台式机。经过多次实验测得，选取每个类中使用改进后的 TF-IDF 方法所计算的特征值按从大到小排序后的前 200 个特征词组成该类的特征向量时，分类效果最佳，具体实验结果如表 2 所示：

表 2　不同个数特征词对应的整体查准率

| 特征词个数 | 查准率 |
|---|---|
| 100 | 92.67% |
| 150 | 94.67% |
| 200 | 95.5% |
| 250 | 95.16% |

为了证明改进后的 TF-IDF 方法的有效性,本文把对 TF-IDF 具有相似改进思想的文献[12]提出的方法用于该数据集,实验结果如表 3 所示:

表3 改进 TF-IDF 实验结果和文献[12]方法实验结果对比

| 指标 \ 类别 | 汽车 | 文化 | 医药 | 军事 | 体育 | 经济 |
|---|---|---|---|---|---|---|
| 查准率 | **86%** | **98%** | **98%** | **96%** | **98%** | **97%** |
| 查准率 | 85% | 87% | 92% | 96% | 90% | 92% |
| 错误率 | **14%** | **2%** | **2%** | **4%** | **2%** | **3%** |
| 错误率 | 15% | 13% | 8% | 4% | 10% | 8% |
| 耗时 | | | **1 026 毫秒/600 篇** | | | |
| 耗时 | | | 1 038 毫秒/600 篇 | | | |

表 3 中颜色较深的行显示的是本文提出的改进 IF – IDF 方法的实验结果,颜色较浅的行显示的是文献[12]提出方法的实验结果。结果表明用基于概率的 TF-IDF 方法在进行网页文本特征词提取后,在进行汽车、军事文档分类时分类的准确率与文献[12]的方法相当,在进行文化、医药、体育、经济类文档分类时分类准确性均优于文献[12]的方法。在对相同的 600 篇文档分类测试中,本文提出的改进方法耗时较短。

## 5 结 语

本文根据大数定理修改了 TF-IDF 中 IDF 值的计算方法,使得 IDF 的计算方法更符合其设计思想。并用实验证明这种改进是有效的且分类精度较高,加之其运算过程简单快速,因此适合企业网站采用。

TF-IDF 方法的改进工作还需进一步完善:当两个类别相近时,本文提出的改进方法的分类效果不好,比如实验中容易将报道汽车的文章划分为报道经济的文章。

## 参考文献:


[1] Sebastiani F. Machine Learning in Automated Text Categorization [J]. *ACM Computing Surveys (CSUR)*, 2002,34(1):1 – 47.

[2] 鲁松,李晓黎,白硕. 文档中词语权重计算方法的改进[J]. *中文信息学报*, 2000,14(6): 8 – 13. (Lu Song, Li Xiaoli, Bai Shuo. An Improved Approach to Weighting Terms in Text[J]. *Journal of Chinese Information Processing*, 2000,14(6): 8 – 13.)

[3] 罗欣, 夏德麟, 晏蒲柳. 基于词频差异的特征选取及改进的 TF-IDF 公式[J]. *计算机应用*, 2005, 25(9): 2031 – 2033. (Luo Xin, Xia Delin, Yan Puliu. Improved Feature Selection Method and TF-IDF Formula Based on Word Frequency Differentia [J]. *Journal of Computer Applications*, 2005, 25(9): 2031 – 2033.)

[4] 张保富, 施化吉, 马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. *计算机应用与软件*, 2011, 28(2):17 – 20. (Zhang Baofu, Shi Huaji, Ma Suqin. An Improved Text Feature Weighting Algorithm Based on TFIDF[J]. *Computer Applications and Software*, 2011,28(2): 17 – 20.)

[5] Forman G. BNS Feature Scaling: An Improved Representation over tf – idf for SVM Text Classification[C]. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, 2008: 263 – 270.

[6] Lan M, Tan C L, Low H B, et al. A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines[C]. In: *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. New York, NY, USA: ACM, 2005: 1032 – 1033.

[7] Oren N. Reexamining tf. idf Based Information Retrieval with Genetic Programming[C]. In: *Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*. Republic of South Africa: South African Institute for Computer Scientists and Information Technologists, 2002: 224 – 234.

[8] Aizawa A. An Information – theoretic Perspective of tf – idf Measures[J]. *Information Processing and Management*,2003,39(1):45 – 65.

[9] 梁之舜, 邓集贤, 杨维权,等. 概率论及数理统计[M]. 北京:高等教育出版社, 1988. (Liang Zhishun, Deng Jixian, Yang Weiquan, et al. Probability Theory and Mathematical Statistics [M]. Beijing: Higher Education Press, 1988.)

[10] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. *软件学报*,2006,17(9): 1848 – 1859. (Su Jinshu, Zhang Bofeng, Xu Xin. Advances in Machine Learning Based Text Categorization[J]. *Journal of Software*,2006,17(9):1848 – 1859.)

[11] Zhang H P, Yu H K, Xiong D Y, et al. HHMM – based Chinese Lexical Analyzer ICTCLAS[C]. In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics,2003,17:184 – 187.

[12] 张玉芳,彭时名,吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. *计算机工程*,2006,32(19): 76 – 78. (Zhang Yufang, Peng Shiming, Lv Jia. Improvement and Application of TFIDF Method Based on Text Classification[J]. *Computer Engineering*, 2006,32(19):76 – 78.)


(作者 E – mail:787186038@qq.com)