ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.36, December 2009, pp.10626-10628

E-mail: xsjl@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

文本分类算法中词语权重计算方法的改进

赵小华,马建芬

(太原理工大学 计算机与软件学院,山西 太原 030024)

摘要:在自动文本分类中,TFIDF 公式是常用的词语权重计算公式。该方法简单易行,但仅仅考虑了特征词出现的频率,而忽略了特征词对区分每个类的贡献。针对这个不足,该文提出了 TFIDF-CHI,来修正各个特征词的权重,重新调整每个特征词对各个类别的区分度,并用 KNN 分类器来验证其有效性。实验证明该方法优于原来的 TFIDF 算法,表明了改进的策略是可行的。

关键词:文本分类;特征权值;TFIDF;TFIDF-CHI

中图分类号:TP312 **文献标识码**:A **文章编号**:1009-3044(2009)36-10626-03

Modify the Method of Feature´s Weight in Text Classfication

ZHAO Xiao-hua, MA Jian-fen

(Dept. of Computer and Software College, Taiyuan University of Techonology, Taiyuan 030024, China)

Abstract: In auto text classification, TFIDF is often used when the weight of a term is calculated.The method is easy, only considers the frequency of the feature and ignores the feature´s contribution to each class. Aiming at this shortage, we put forward the TFIDF-CHI and use it to modify each feature´s weight, read just each feature´s differentiation to each class. Then the KNN classifier is used to check its validity. The method is better than traditional TFIDF and proves that the TFIDF-CHI method is feasible.

Key words: text classification; feature weight; TFIDF; TFIDF-CHI

现在,政府、工业、商业和其他机构的大部分信息都以文本数据库的形式电子地存储,同时电子出版物、各种电子文档、电子邮件和万维网等文本数据库也正在快速的增长。随着各种电子形式的文本文档以指数级的速度增长,有效的信息检索、内容管理及信息过滤等应用变得越来越重要和困难。文本自动分类是一个有效的解决办法,已成为一项具有实用价值的关键技术。文本分类的主要步骤为:文本预处理、训练分类模型、测试分类模型。近年来,多种统计理论和机器学习方法被用来进行文本的自动分类,掀起了文本自动分类研究和应用的热潮。

本文通过研究发现传统的文本特征权值表示方法 TFIDF 的不足:它忽略了特征词和类别之间的相关性。本文认为特征词和类别之间没有绝对的独立性,针对这个不足,提出了 TFIDF-CHI 算法,并用实验加以证明。

1 TFIDF 算法及其改进

1.1 χ^2 统计量

χ^2 统计量(chi-square statistic, CHI)特征选择方法又被称作开方拟合检验,这个概念来自列表检验,它可以用来衡量特征 x 与类别 c 之间的统计相关性。χ^2 方法认为特征 t 与文本类别之间的没有独立性,它们之间的关系类似于具有一维自由度的 χ^2 分布,χ^2 统计量的值越高,词汇和类别之间的独立性就越小。它基于如下假设:在指定类别 Ci 的文本中出现频率高的词语和在其它类的文本中出现频率高的词语,对判断文章是否属于类别 Ci 都有帮助。其计算公式如下:

$$CHI(F)=\sum_i P(C_i)X^2(t,C_i)=\sum_i P(C_i)\frac{N*(AD-BC)^2}{(A+B)(B+D)(A+B)(C+D)} \quad (1)$$

式中,A 是特征 t 和第 i 类文档共同出现的频度;B 是特征 t 出现而第 i 类文档不出现的频度;C 是第 i 类文档出现而特征 t 不出现的频度;D 是第 i 类文档和特征都不出现的频度;N 为总共的文本数,且 N=A+B+C+D,同时要求满足 A*D>B*C。文献[1]中指出,CHI 算法综合考虑了特征与类别出现的各种可能性,在文本数量逐渐增多的过程中,稳定性很好;与其他方法相比,CHI 大约减少 50%的词汇,分类效果好。

1.2 传统 TFIDF 计算方法

传统的 TFIDF 权重计算方法是由 Salton 在 1988 年提出的。指导思想是:在同一个文本中出现的频率较高,在不同文本中出现的频率较小的词应该赋予较高的权值。它主要考虑两个方面:词语在文本中出现的频率(TF),用于计算该词描述文档内容的能力;反文档频率(IDF),用于计算该词区分文档的能力。特征词条的权值与词条频率成正比,与文档频率成反比。

传统 TFIDF 权值计算公式:

$$w(t_k,d)=\frac{tf(t_k,d)*\log_2(N/n_k+d)}{\sqrt{\sum_{t\in d}[tf(t_k,d)*\log_2(N/n_k+d)]^2}} \quad (2)$$

收稿日期:2009-09-20

作者简介:赵小华(1982-),女,山东枣庄人,在读研究生,太原理工大学,计算机与软件学院,主要研究方向为自然语言处理,数据挖掘;马建芬(1967-),女,山西太原人,副教授,硕士研究生导师,太原理工大学,计算机与软件学院,主要研究方向为自然语言处理、语音信号处理、多媒体技术。

其中 tf(t,d)为特征 t 在文本 d 中的频数,n 为文本集中含有 t 的文本的数量,a 是一个常量(一般取 0.01),$\log_2(N/n_{tk}+a)$是逆文本频率函数,即 n 越大此值越小。分母是归一化因子。

但是传统 TFIDF 权值计算方法也有其不可避免的不足,IDF 的简单结构并不能有效地反映单词的重要程度和特征词的分布情况,使其无法很好地完成对权值调整的功能,所以 TFIDF 法的精度并不是很高。其权重计算的有效性和词条的分类能力就存在严重不足。而 CHI 算法却能够很好的弥补 TFIDF 算法的不足。

1.3 TFIDF–CHI 算法

因此,我们将 TFIDF 算法和 CHI 算法加以综合,用 CHI 算法的优点来弥补 TFIDF 的不足,提出了新的权值计算方法,TFIDF–CHI 算法。TFIDF–CHI 的计算公式为:

$$w(t_k,d)=\frac{tf(t_k,d)*\log_2(N/n_{tk}+a)*CHI}{\sqrt{\sum_{t_k\in d}[tf(t_k,d)*\log_2(N/n_{tk}+a)*CHI]^2}} \quad (3)$$

2 试验过程

2.1 实验环境与实验数据集

我们用 Visual C++ 6.0 实现本文的算法,在 Windows XP 的环境下进行试验。实验数据是是从中文自然语言处理开放平台网站获取李荣陆收集的新华社的新闻样本语料库。其中训练样本 2000 个,测试样本 815 个,共 2815 个样本。样本有 10 个类别,分别为 政治、艺术、医药、体育、军事、经济、教育、交通、计算机、环境;

2.2 评估方法

因为文本分类本质上是一个映射过程,所以评估文本分类系统的标志是映射的准确程度和映射的速度。映射的速度取决于映射规则的复杂程度,而评估映射准确程度的参照物是通过专家思考判断后对文本的分类结果(这里假设人工分类完全正确并且排除个人思维差异的因素),与人工分类结果越相近,分类的准确程度就越高。 本文中文本分类的评价方法主要有查准率(也称为准确度)、查全率(也称为召回率)。

准确率是所有判断的文本中与人工分类结果吻合的文本所占的比率。其数学公式为:

准确率(precision)=分类的正确文本数/实际分类的文本数

召回率是人工分类结果应有的文本中分类系统吻合文本所占的比率,其数学公式为:

召回率(recall)=分类的正确文本数/应有文本数

2.3 试验分析

本文的试验使用 KNN 分类器,其中 K 取 35。本文随机抽取了“收入”、“亚军”、“武器”和“青年”四个特征词,其中“亚军”和“武器”是在体育和军事类中常见的,而在其他类中则不常见,所以它们对类的贡献比较大。而“青年”可在多个类别中多次出现,所以其对类的贡献相对较小。通过我们对公式的改进,可猜想改进后的“亚军”和“武器”的权值应该增大,而“青年”的权值则应减小。

表 1 特征词权重

		收入	亚军	武器	青年	分类准确率
TFIDF	权重	3.388722	4.167415	3.474268	3.206953	0.787595
TFIDF-CHI	权重	3.297835	4.423568	3.584629	2.986537	0.806135

表 2 各个类别的分类准确率及召回率

		政治	艺术	医药	体育	军事	经济	教育	交通	计算机	环境
TFIDF	准确率	58.683%	92.857%	90.476%	93.333%	82.609%	55.932%	95.082%	98.000%	100.000%	91.429%
	召回率	93.333%	79.592%	70.370%	93.333%	57.576%	88.000%	82.857%	76.563%	78.000%	64.000%
TFIDF -CHI	准确率	60.568%	94.652%	92.634%	94.412%	82.609%	56.758%	96.012%	98.096%	99.965%	92.986%
	召回率	94.234%	80.125%	71.542%	94.826%	59.576%	89.965%	83.623%	77.546%	79.060%	64.452%

表 1 列出了针对四个不同的特征词,分别采用 TFIDF 和 TFIDF–CHI 两种不同的计算方法所得到的权值和的分类准确率。实验中具有代表性的词使用 TFIDF–CHI 算法后的权值明显比用 TFIDF 算法所得的权值大。从表中可看出计算结果与我们所猜想的结果基本一致。表 2 通过列出常用的度量标准,准确率和召回率对两个算法进行对比。由表 2 可知通过对 TFIDF 的改进,重置特征词的权重,使文本分类的准确率和召回率都明显得到改善,凸显了特征词与类别之间的关系。

3 试验总结

综上所述, 从实验的图表中我们可以看出不同的权值计算方法对分类的准确率有着一定的影响。与 TFIDF 相比, 改进后的 TFIDF 更能够反映特征词与类别之间的关系,进一步提高了文本分类的准确率。近年来,一些研究者针对 TFIDF 权重函数提出了大量的改进算法。文献[2–4] 在 TFIDF 的基础上结合了文本语义、频率等多方面信息,提出了新的改进算法。文献[5–11]针对 TFIDF 没有考虑特征向在文本集上的分布比例,而对其改进,将 TFIDF 和互信息,信息增益等方法进行了融合。文献[12–13]将 Gini index 与 TFIDF 相结合,提出了新的改进算法。文献[14]在 TFIDF 的基础上引入了遗传算法。文献[15]考虑了当训练文本属于同一类别时,文本特征的权值计算。文献[16]通过对 TFIDF 的改进考虑到了特征项在类间的分布情况。

本文将 TFIDF 和互信息综合,各取其优点,将分类的准确率和召回率进一步提高。同时由于 CHI 算法综合考虑了特征与类别出现的各种可能性,在文本数量逐渐增多的过程中,稳定性很好;与前面其他各种改进方法相比,CHI 大约减少 50%的词汇,分类效果好。

4 结束语

特征权重计算方法的选择对文档分类的精确度有很大的影响。本文研究了传统的 TFIDF 算法,并在其基础上提出了新的改进方法 TFIDF–CHI。由于 CHI 方法考虑了特征词对类别的贡献,所以理论上改进后的 TFIDF–CHI 方法能够通过重置权值,更好的优

化分类效果。实验证明,新的改进方法 TFIDF-CHI 在分类的精确度上确实有更好的表现,因此,实验说明了,改进后的 TFIDF 算法——TFIDF-CHI 是一种高效可行的特征权重计算方法。虽然 TFIDF 是很经典的算法,但是 TFIDF 还有很多方面值得我们研究,例如虽然 TFIDF 权值计算方法可以和很多的特征选择方法进行综合,更改权值的计算方法,以提高分类效率,但是他 TFIDF 和哪种选择方法综合后,分类效率最高,适用的场合也最广则还有待研究。

参考文献:

[1] 张俊丽,赵乃瑄,冯君.基于统计频率的文本分类特征选择算法研究[J].现代图书情报技术,2008(11):44-48.
[2] 沈志斌,白清源. 一种基于多重因子加权的文本特征项权值计算方法[J].南京师范大学学报:工程技术版,2008,8(4):80-83.
[3] 龚静,田小梅.基于文本表示的特征项权值计算方法[J].电脑开发与应用(总 133),21(2):46-48.
[4] 龚静,周经野.一种基于多重因子加权的文本特征项权值计算方法[J].计算技术与自动化,2007,26(1):80-83.
[5] 姚兴山.基于统计的中文文本分类研究[J].信息系统,2009,32(5):95-98.
[6] 陈国松,黄大荣.基于信息熵 TF IDF 文本分类特征选择算法研究[J].湖北民族学院学报:自然科学版,2008,26(4):402-404.
[7] 廖浩,李志蜀,王秋野,等.基于词语关联的文本特征词提取方法[J].计算机应用,2007,27(12):3009-3012.
[8] 白硕,王实.文档中词语权重计算方法的改进[J].中文信息学报,2007,14(6):8-13.
[9] 冯长远,普杰信.Web 文本特征选择算法的研究[J].计算机应用研究,2005,(7):36-38.
[10] 初建崇,刘培玉,王卫玲.Web 文档中词语权重计算方法的改进[J].计算机工程与应用,2007,43(19):192-198.
[11] 沈志斌,白清源.文本分类中特征权重算法的改进[J].南京师范大学学报:工程技术版,2008,8(4):95-98.
[12] Shang Wenqian,Qu Youli,Zhu Haibin,et al.An Adaptive Fuzzy kNN Text Classifier Based on Gini Index Weight [C].Proceedings of the 11th IEEE Symposium on Computers and Communications(ISCC'06)0-7695-2588-1/06,2006.
[13] Shang Wenqian,Huang Houkuan,Zhu Haibin,et al.An Adaptive Fuzzy kNN Text Classifier [M]//Alexandrov V N.ICCS 2006,Part III, LNCS 3993,2006:216-223.Springer-Verlag Berlin Heidelberg.2006.
[14] 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.
[15] 寇莎莎,魏振军.自动文本分类中权值公式的改进[J].计算机工程与设计,2005,26(6):1616-1618.
[16] 熊忠阳,黎刚,陈小莉,等伟.文本分类中词语权重计算方法的改进与应用[J].计算机工程与应用,2008,44(5):187-189.

(上接第 10623 页)

Agent 是嵌入被管网元设备内的代理,和被管网元设备的其他软件之间采用紧耦合方式进行通信,可共享全局变量或者调用其他模块的函数,它负责收集、保存网元的相关信息,并为 Manager 提供访问接口。Manager 通过访问被管网元设备的 MIB 树来获取网元信息,并把 GUI 对网元设备的操作命令也提交给 Agent,然后由 Agent 进行过滤、解析,进而操作网元设备。

3 结束语

本系统的功能完善,把所有的网元设备(数量可成百上千)呈现在 GUI 的拓扑视图中,管理员可以很容易的用图形化的方式配置、管理和维护网络中的设备,能够详细、全面地掌握网络的状态,大大提高了管理的效率和服务质量,达到最初的设计目标。

参考文献:

[1] 岑贤道,安常青.网络管理协议及应用开发[M].北京:清华大学出版社,1998.
[2] 韦乐平.SDH 及其新应用[M].北京:人民邮电出版社,1998:170-184.
[3] SERVA Software.TWaver™ Swing Components Developer Guide Version 2.0[S].2008:10-260.
[4] 唐宝民,张颖.电信网监控和管理计数[M].北京:人民邮电出版社,2006:138-168.