第35卷 第24期 计 算 机 工 程 2009年12月
Vol.35 No.24 Computer Engineering December 2009
·软件技术与数据库· 文章编号：1000—3428(2009)24—0069—03 文献标识码：A 中图分类号：TP391

基于KNN的不良文本过滤方法

王洪彬，刘晓洁
(四川大学计算机学院，成都 610065)

摘 要：不良文本过滤是当前的一个研究热点。通过对 χ^2 统计量的具体分析，证明 χ^2 统计量在2类文本特征项提取过程中特有的优势。提出正面文本阈值 δ，并从理论上推断出该值的大小。在此基础上改进KNN算法，消除了KNN算法中 N 的不确定性，彻底实现了无参性，大幅减少了分类所用的时间。实验证明，该算法符合Web实时在线分类的要求。

关键词：KNN算法；不良文本过滤；χ^2 统计量

Reactionary Text Filtering Method Based on K-Nearest Neighbor

WANG Hong-bin, LIU Xiao-jie
(School of Computer, Sichuan University, Chengdu 610065)

【Abstract】 Reactionary text filtering is a hot research now. This paper proves that statistics χ^2 has the unique advantages in the features extraction of the two types of texts based on statistics χ^2 analysis. It proposes the threshold δ of the positive texts and infers the value of it in theory, and the K-Nearest Neighbor(KNN) algorithm is improved. This algorithm eliminates the uncertainty of KNN algorithm, realizes no reference, and reduces the time used in the text categorization. Experimental results show that the algorithm meets the real-time online text categorization.

【Key words】 K-Nearest Neighbor(KNN) algorithm; reactionary text filtering; statistics χ^2

互联网的迅速发展大大方便了人们对信息的获取。但是网络上也充斥了不少不良信息，对社会稳定构成极大危害。如何过滤掉不良信息，是一个研究热点[1-2]。不良文本的过滤本质上是一个两类文本的分类问题，文本分类是指将一篇文本自动指定到一个或几个预定义的文本类别中。在文本分类中，基于SVM的(K-Nearest Neighbor, KNN)方法被证明是效果最好的分类方法之一[3]。但是KNN算法也存在缺陷：它是一种懒惰的学习方法，每一篇测试文本要和所有的训练文本做相似度比较，导致很大的计算量。本文的工作是基于KNN算法的改进与应用。

1 KNN文本分类算法

1.1 特征值提取

在对文本进行了分词和去停用词处理以后，需要对所有的训练文本进行特征项提取，以选取对于文本分类贡献最大的特征项。进行特征项提取的目的有2个：(1)降维，减少运算量；(2)提高分类的准确度。常见的特征项提取方法包括信息增益IG、互信息MI、统计量CHI[4]等。

1.2 文本权重计算

特征项对于分类越有利，就赋给它越高的权值。一般采用词频和倒文档频度方法(TF-IDF)来计算权重，表达式为

$$w(t,\bar{d}) = tf(t,\bar{d}) \times \lg(N / N_r) \quad (1)$$

其中，$w(t,\bar{d})$ 表示的是词 t 在文本 $\bar{d}$ 中的权重；$tf(t,\bar{d})$)表示的是词 t 在文本 $\bar{d}$ 中出现的次数；N 为训练文本总数；N_r 为训练文本中出现词 t 的文本数。

1.3 KNN算法

KNN算法[5-8]的本质是：把待测试文本和所有的训练文本都在一个向量空间模型中表示出来，找出与待测试文本最接近的 K 个训练文本，分析这 K 个训练文本中属于哪一类的较多，然后把测试文本归属到哪一类中，以此来确定待测试文本的分类。

KNN可以很好地实现文本的分类，但是KNN算法也有很多不足的地方。由于测试文本要与每一个训练文本比较，因此它的计算量很大。计算相似度时，特征向量维数高，各维权数相同，使得特征向量之间的距离计算不够准确，影响分类精度。基于以上这些问题，改进了KNN文本分类算法。

2 KNN改进算法分析

2.1 特征值提取方法

证明在不良文本分类中，CHI方法可以精确地提取出代表每一类文本的特征词。

证明：CHI方法认为词和类别之间符合 χ^2 分布。χ^2 统计量体现了词和类别之间的相关性，其值越高，词和类别之间的独立性越小，相关性越强，词对该类别的贡献越大，也暗示着包含该词的文档属于该类别的概率越大，χ^2 值为0表示两者不相关。

不良文本分类本质上就是一个两类文本的分类问题，设正面文本类为 $C1$，不良文本类为 $C2$。则得到某特征项 t 的 $\chi^2(t,C1)$ 中的 A,B,C,D。其中，$A = P(t \mid C1)$；$B = P(t \mid \overline{C1})$；$C = P(\bar{t} \mid C1)$；$D = P(\bar{t} \mid \overline{C1})$。$t$ 在 $\chi^2(t, C2)$中的 $A' = P(t \mid C2)$，$B' = P(t \mid \overline{C2})$，$C' = P(\bar{t} \mid C2)$，$D' = P(\bar{t} \mid \overline{C2})$，而 $\overline{C1} = C2$，

基金项目：国家自然科学基金资助项目(60573130, 60502011)；国家“863”计划基金资助项目(2006AA01Z435)；教育部新世纪优秀人才计划基金资助项目(NCET-04-0870)

作者简介：王洪彬(1983－)，男，硕士，主研方向：网络安全，人工智能；刘晓洁，副教授

收稿日期：2009-07-10 **E-mail：**wanghb280197846@yahoo.com.cn

$C1=\overline{C2}$。由此得到以下结果：$A=B'$, $B=A'$, $C=D'$, $D=C'$。因而进一步得到：

$$A+C=B'+D' \tag{2}$$

$$B+D=A'+C' \tag{3}$$

$$\frac{\chi^2(t,C1)}{\chi^2(t,C2)}\equiv 1 \tag{4}$$

此时，得出结论：在不良文本过滤中，任何一个特征项对正反两类文本的分类贡献是一致的。得出和要证明的内容完全相反的结论。

即使是相同的 χ^2 值，特征词也只能代表其中一类。在 $\chi^2(t,C1)$ 中，A 表示在文本类 $C1$ 的文本中包含特征项 t 的文本数；D 表示在文本 $C2$ 中不包含特征项 t 的文本数。A 和 D 表示的是特征项 t 和文本类 $C1$ 的相关程度。A 和 C 越大，表示 t 越能代表 $C1$；而 B 和 C 表示的是特征项 t 和文本类 $C1$ 的远离程度，B 和 C 越大，意味着该特征词 t 越不能代表 $C1$ 类。所以只有当 $AD>BC$ 时，才认为 t 和 $C1$ 类是相关联的，否则，t 和 $C2$ 类相关联。

根据式(2)~式(4)，推出 $AD>BC$ 和 $A'D'>B'C'2$ 个公式中只能有一个成立。

由此得到特征项 t 中较大的 CHI 值只能代表其中的一个类别。

证毕。

2.2 特征项的权值选取

通过特征项提取方法的性质，可以提取到最能够代表每个类别的特征项。在这些特征项中，χ^2 值越大，和该类的相关性越强，就越能够代表该类，应该被赋予更高的权值。由此，采用归一化思想定义了一种新的赋予权值的方法：

$$w_k=\frac{\chi^2(t_k,C)}{\sqrt{\sum_{i=1}^{n}\chi^2(t_i,C)^2}} \tag{5}$$

其中，w_k 表示特征项 t_k 在类 C 中的权值；n 表示的是能够代表类 C 的特征项的维数。

2.3 正面文本的阈值 δ

在不良文本过滤过程中，大部分文本都是正面文本，并且在实际应用中，测试文本的数量要远远大于训练文本的数量。所以设定一个阈值，让正面文本能够很迅速地通过，这是 Web 实时在线分类的要求，很有必要的，也是可以实现的。

设定一个阈值 δ，如果待分类文本与能够代表正面文本的特征项集合的亲和力大于 δ，则直接让该文本通过。以下是 δ 的公式：

$$\delta=\max(\theta,\lambda) \tag{6}$$

$$\theta=\max(\sum_{k=1}^{m}w_k\times E(t_k,C_{2i}))(i=1,2,\cdots,q) \tag{7}$$

$$\lambda=\min(\sum_{j=1}^{m}w_j\times E(t_j,C_{1i}))(i=1,2,\cdots,p) \tag{8}$$

其中，$C1$ 代表正面文本；m 为正面文本特征项集合的维数；p 为 $C1$ 的文本数；$C2$ 代表反面文本；q 为 $C2$ 的文本数；$E(t_j, C_{1i})$表示正面文本特征项 t_j 在文本类 $C1$ 的第 i 篇文本中出现的次数；w_j 是先前为正面特征项 t_j 所定义的权值大小。同理，$E(t_k,C_{2i})$表示正面文本特征项 t_k 在文本类 $C2$ 的第 i 篇文本中出现的次数； w_k 是先前为正面文本特征项 t_k 所定义的权值大小。

在求出阈值 δ 以后，可以分析出：如果是 $\theta<\lambda$，则对于训练文本而言，所有的正面文本都能通过，反面文本则需要进一步检测才可以通过；如果是 $\theta>\lambda$，则可以让一大部分正面文本通过。在上面的 2 种情况下，都不会放训练文本中的任何一个反面文本通过，反面文本还需要作进一步检查。在提高了效率的同时，不会降低准确率。

同时，可以给 δ 添加一个修正值 α。如果要求实时性更好，使 α=1；如果为了减少阈值 δ 过低导致不良文本没有被检测出来，而使其被错误地分类，可以适当上调 α，使 α=2 或者 α=3。在实验中，取 α=2。

3 算法设计

KNN 算法来源于文本分类算法，而不良文本的过滤本质上是一个两类文本分类的问题。在本文中，把经过 χ^2 特征提取的正面文本向量集合和反面文本向量集合以及它们各自的权值都计算出来。然后看待分类文本是和正面文本的亲和力大还是和反面文本的亲和力大，依此判定待分类文本的类别。由于事先已经确定了每一个特征项的权值，而且不用和每一个训练文本都作比较，因此提高了文本的过滤速度。

定义待分类文本 d 和代表一类文本 C 的特征向量集合的相似度为亲和力 $Affinity(d,C)$。$Affinity(d,C)$ 值越大，表示该文本 d 和类 C 越接近。

$$Affinity(d,C)=\sum_{k=1}^{n}w_k\times tf(t_k,d) \tag{9}$$

其中，n 为代表 C 类文本的特征项维数；$tf(t_k, d)$表示特征项 t_k 在文档 d 中出现的频数。

不良文本过滤流程如图 1 所示。

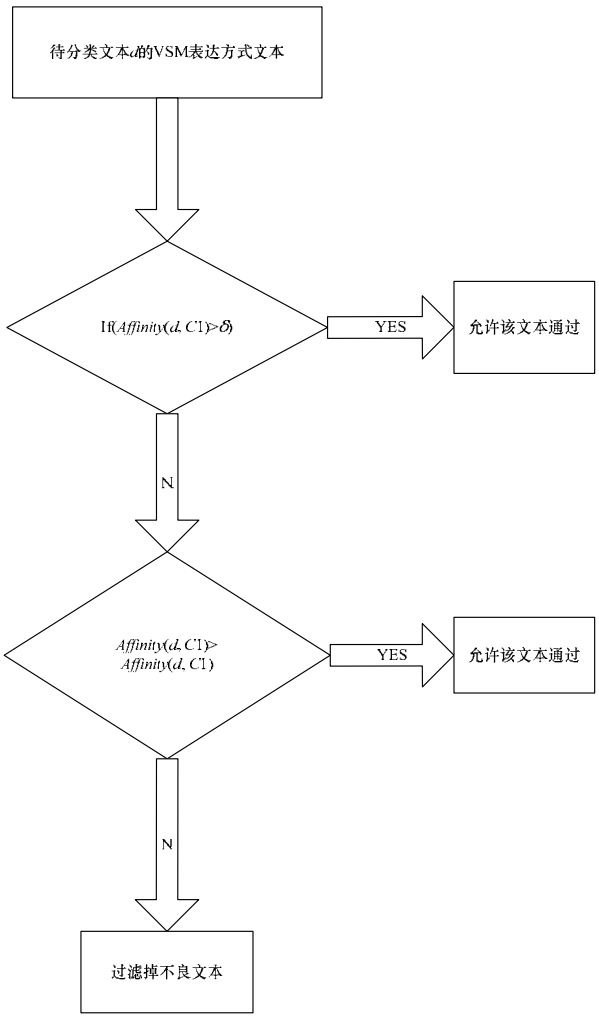

图 1 不良文本过滤流程

4 仿真实验与结果分析

4.1 实验验证

以某项不良主张相关文献为实验，检验改进的 KNN 算法以及不良文本过滤模型的可行性。

搜索到 217 篇反对的文章，再从某些不良网站搜索到 338 篇支持的文章，一共 555 篇作为训练文本，然后随机抽取 275 篇文章，其中支持 167 篇，反对 108 篇。用经典的 KNN 方法，N 值取 51，特征向量维数取为 1 000，作为第 1 种方法。采用 CHI 方法进行正面文本和反面文本特征项的提取，并利用在本文中设计的模型进行不良文本过滤。此时，为了对比，做了 3 种情况，分别是不对正反特征项进行归一化操作，对正反特征项进行归一化操作，以及在不对正反特征项进行归一化的前提下，根据正反特征项维数的比例，设权值 α=4，加权以后对这 3 种情况做实验。正面文本和反面文本的特征项维数和为 1 000。实验结果如表 1 所示。

表 1 仿真实验结果

方法类别	准确率	召回率	$F1$ 值
经典 KNN 类别 1	0.925 2	0.916 7	0.920 9
经典 KNN 类别 2	0.946 4	0.952 1	0.949 3
归一化前类别 1	0.577 5	1.000 0	0.732 2
归一化前类别 2	1.000 0	0.526 9	0.690 2
归一化后类别 1	0.882 9	0.907 4	0.895 0
归一化后类别 2	0.939 0	0.922 2	0.930 5
加权值后类别 1	0.840 3	0.925 9	0.895 0
加权值后类别 2	0.948 7	0.886 2	0.916 4

4.2 数据分析

当用未归一化的特征项来进行分类操作时，发现 $F1$ 值都比较低，但是从数据中可以发现，所有的正面文本都被分辨出来了，只是对于反面文本的分辨率很低。归其原因，是因为在经过 CHI 运算得到的特征项中，属于正面文本的特征项的维数要远大于属于反面文本特征项的维数，大概是 4:1 的关系。这就是加上修正值 α=4 的原因。

通过实验数据，得到改进后的 KNN 算法的准确率大致上和经典的 KNN 算法的准确率相同，而所用的时间却要少很多。因为对于经典的 KNN 算法来说，当一篇测试文本到来的时候，需要计算它和所有训练文本的相似度，然后还要将计算出来的相似度从大到小进行排序，选取出来相似度最大的 N 个值，最后根据这 N 个值判断该文本的类别。对于动辄几千的训练文本来说，这是一个很大的计算量。而对于改进的 KNN 算法来说，当测试文本到来的时候，只需要将它和正反特征项集合进行 2 次相似度计算就可以了，然后看它更接近于哪一类文本，就把它归于哪一类，效率最起码提高了一个数量级。

通过对训练文本的分析，综合式(6)~式(8)，得到：θ=1.536 6>λ=0.210 8，设修正值 α=2，推出 δ=3.073 2。当把设置的正面文本过滤阈值 δ 用于对测试文本进行过滤的时候，没有出现错误过滤，正确判断正面文本的准确率是 28.7%，进一步提高了分类的速度。

5 结束语

本文利用 χ^2 统计量在两类文本分类中提取特征项的优势，提出了改进的 KNN 算法。实验证明：用改进的 KNN 算法来过滤不良文本，查全率和召回率大致和改进前的 KNN 算法相等，但是改进后的算法拥有较高的效率，大大缩短了分类所用的时间，满足了实时在线分类的需求。并且在新的算法中不用像在传统的 KNN 算法中那样人为地设置 N 的大小，这样就减少了由于经验值带来的误差。下一步工作主要集中在如何进一步提高算法的准确率和召回率。

参考文献

[1] 李 强, 李建华. 基于向量空间模型的过滤不良文本方法[J]. 计算机工程, 2006, 32(10): 4-8.

[2] Hanani U, Shapira B, Shoval P. Information Filtering: Overview of Issues, Research and Systems[J]. User Modeling and User-adapted Interaction, 2001, 11(3): 203-259.

[3] He Ji, Tan Ah-Hwee, Tan Chew-Lim. A Comparative Study on Chinese Text Categorization Methods[C]//Proc. of the International Workshop on Text and Web Mining. Melbourne, Australia: [s. n.], 2000: 24-35.

[4] 王秀娟, 郭 军, 郑康锋. 文本分类中一种新的特征选择方法[J]. 计算机应用, 2005, 25(3): 661-663.

[5] Kuncheva L I. Fitness Functions in Editing KNN Reference Set by Genetic Algorithms[J]. Pattern Recognition, 1997, 30(6): 1041-1049.

[6] Li Ying, Zhang Xiaohui, Wang Huayong, et al. Vector-combination-applied KNN Method for Chinese Text Categorization[J]. Mini-Micro Systems, 2004, 25(6): 993-996.

[7] 杨丽华, 戴 齐, 郭艳军. KNN 文本分类算法研究[J]. 微计算机信息, 2006, 22(21): 269-270.

[8] Wang Yu, Wang Zhengou. A Fast KNN Algorithm for Text Categorization[C]//Proc. of the 6th International Conference on Machine Learning and Cybernetics. Hong Kong, China: [s. n.], 2007: 3436-3441.

编辑 顾逸斐