广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(4):436~438
Vol. 26, No. 4 November 2010

基于多质心的不良文本快速过滤方法
A Method of Illegal and Harmful Text Fast Filter Based on Multi-Centroid Vector

黄家裕,刘连芳
HUANG Jia-yu, LIU Lian-fang

(南宁市平方软件新技术有限责任公司,广西南宁 530007)
(Pingsoft New Technology Co. Ltd. of Nanning, Nanning, Guangxi, 530007, China)

摘要:针对 Rocchio 容易受到类别样本分布及噪声影响的而导致错误扩大类别范围的问题,提出对训练样本进行聚类,使用聚类形成的多个簇的质心向量替代单个质心向量作为过滤判定向量组的方法。该方法既能保证过滤效率,又比单质心的 Rocchio 过滤法具有更高的召回率和准确率。

关键词:不良文本 快速过滤 多质心向量 Rocchio K-means

中图法分类号:TP391.3 **文献标识码**:A **文章编号**:1002-7378(2010)04-0436-03

Abstract: Aiming at the defect in Rocchio that classification range could be easily mis-extended due to distribution of classification samples and noises, a filtering method is presented in this paper, in which a vector of single centroid is substituted by a vector group of centroids at multiple clusters formed by clustering trained samples and used as a deciding vector group for filtering. This method is characterized by lossless filtering efficiency. Recalling rate and accuracy of this method is higher than that of the single centroid-featured Rocchio Filtering.

Key words: illegal and harmful text, fast filter, multi-centroid vector, Rocchio, K-means

互联网已经成为一个大型信息库,是一个快速信息交流的平台,它对人们的工作、生活、学习都产生了巨大的影响。然而,互联网的开放性导致它提供大量有价值信息的同时,还充斥着大量的不良信息,比如:色情、反动、暴力、邪教、迷信等等,对社会造成极大危害。互联网不良信息过滤是当前互联网环境下的一个重要研究课题[1]。

从信息过滤的主体看,互联网信息过滤可分为两类,一类是用户自主过滤,另一类是第三方(国家管理部门或营运企业)过滤。从信息过滤技术看,互联网信息过滤可以分为基于数据库(IP 库、URL 库)、基于关键词以及基于内容过滤等三种方式[2]。根据过滤的时效性看,互联网信息过滤又可以分为快速过滤和非快速过滤等。互联网信息过滤的本质是二类分类问题。目前文本分类的主要方法有 Rocchio、朴素贝叶斯(NB)、K 最近邻(KNN)、支持向量机(SVM)、神经网络(NNet)、决策树(Decision Tree)、关联规则(Association Rules)等等[3,4]。国内外研究者对这些分类方法进行了很多评测[1,5~7],SVM 和 KNN 的分类效果具有一定的优势,但是由于测试集合和测试条件的差异,所获得的结果仅能作为方法效果的参考。Rocchio 方法的分类效果,在文献[5]中较弱,文献[7]则显示在测试集相对充分时其准确率高于 SVM,文献[1]的实验结论则是与 SVM 效果相当。KNN 方法训练简单,只需要记住各个训练样本,但是在分类时,需要计算待分类文档和所有训练文档之间的相似度,时间复杂度为 $O(m)$,m 为训练文档个数,在 m 较大时,不适合快速过滤的场合。传统的 SVM 方法通过常规二次规划求解,训练时间复杂度为 $O(m^3)$,当 m 较大时训练代价太大,甚至难以实现。Rocchio 方法的算法简单,训练时间复杂度为 $O(m)$,分类时间复杂度为 $O(1)$,算法效率高,特别适合于快速过滤和有反馈信息情况下的过滤。但是由于 Rocchio 最终使用一个质心代表类别,丢失了训练样本的分布细节,容易

收稿日期:2010-09-28
修回日期:2010-10-18

作者简介:黄家裕(1975-),男,软件工程师,主要从事中文信息处理、机器翻译、计算机通信研究工作。

歪曲样本分布状况,错误扩大类别范围;另外质心法平等对待每个样本,对噪声敏感,容易受到边缘噪声样本的影响,使质心偏移,同样也会错误扩大类别范围。这些情况将导致过滤时大量补类对象被错误过滤掉,而如果人为缩小类的半径,又势必造成漏检,即将该类对象误认为是补类对象。本文改进 Rocchio 方法,通过聚类获得训练样本的多个质心,使用多个质心向量作为不良文本快速过滤的判定向量组,在保证过滤效率和训练效率的情况下提高基于内容的自主快速文本的过滤效果。

1 基于多质心的不良文本快速过滤方法

将训练样本按其密度划分成多个簇,然后计算每个簇的质心,构成质心组;在过滤时计算待过滤文档 d 与各个质心的相似度,当存在 $Sim(c_i,d) > t_i$ 时,$i=1,2,\cdots,k$,t_i 为簇 i 的相似度阈值,将 d 判定为正例,即待过滤文档与某个质心的相似度大于阈值时,将被过滤掉。

多质心方法的目的是尽最大程度拟合样本分布情况,在一定程度上克服 Rocchio 法完全丢失训练样本分布细节以及平等对待每个样本的缺陷,解决因为多模态类或样本存在噪声而导致类范围错误扩大的问题。

多质心方法最好时间复杂度为 $O(1)$,最坏时间复杂度为 $O(k)$,k 为簇个数;因此可以根据实际应用需求,通过控制簇的个数来控制过滤的效率。多质心方法的训练效率则依赖于簇的划分算法,为了能够保持 Rocchio 方法的训练效率优势,应为多质心过滤法选择一个较高效率的簇划分算法。将训练样本按其密度划分成多个簇是一个聚类过程。目前的聚类算法可分为:层次法、划分法、基于密度的方法等等[8]。其中划分法中的典型算法 K-means 的效率较高,而且该算法在聚类过程中就不断地计算簇的质心向量,在聚类结束时就可以直接获得质心向量组,而不必额外计算,所以我们选择 K-means 算法作为多质心的不良文本快速过滤的算法。

2 算法设计

2.1 训练算法

输入:
k:簇的个数
t:迭代次数阈值
D:m 个样本的训练样本集
输出:k 个簇
方法:
从训练样本中任选 k 个样本作为初始簇质心;
while(迭代次数小于 t && 质心发生改变)
{
for(i=1 至 k)
{
for(j=1 至 m)
{
计算簇 i 的质心与样本 j 的距离;
}
}
for(j=1 至 m)
{
将样本 j 划分到与它最接近的质心的簇中;
}
for(i=1 至 k)
{
计算簇 i 的新质心;
计算簇 i 新质心与原质心的距离;
}
}
return {簇 1,簇 2,...,簇 k};

2.2 过滤算法

输入:
d:待过滤文档
{簇 1,簇 2,...,簇 k}:k 个簇
输出:是否是不良信息
方法:
IsHarmful = false;
for(i=1 至 k)
{
计算簇 i 的质心与待过滤文档 d 的相似度 Sim;
if(Sim >= 簇 i 的相似度阈值)
{
IsHarmful = true;
break;
}
}
return IsHarmful;

3 实例验证

实验在 AMD Athlon 3600+、1GB 内存、80GB

硬盘、Windows 2000 profession 环境下进行，实验程序使用 C++开发，分别对法轮功和色情两个不良信息专题的文本进行实验。所有文本均从互联网下载，设置如下两组训练集和测试集。

训练集 1：宣扬法轮功文本 400 篇。测试集 1：宣扬法轮功文本 200 篇；揭批法轮功文本 200 篇；其它文本 200 篇。

训练集 2：色情文本 400 篇。测试集 2：色情文本 200 篇；其它文本 200 篇。

上述的“其它文本”包含经济、军事、体育、音乐、计算机等领域。

实验 1：用 Rocchio 法和多质心法分别对训练集 1 进行训练，用测试集 1 分别对两种方法进行过滤测试。

实验 2：用 Rocchio 法和多质心法分别对训练集 2 进行训练，用测试集 2 分别对两种方法进行过滤测试。

本实验的向量维数为 18588 维，包含从文本中提取的法轮功专题词汇 48 个，色情专题词汇 119 个。根据经验，多质心法的质心数取 5，迭代次数阈值为 10；所有记录的用时都不包括从文档构造向量的时间；过滤效果使用召回率和准确率衡量：召回率＝被正确过滤掉的文档数÷应该被过滤掉的文档数，准确率＝被正确过滤掉的文档数÷被过滤掉的文档数。

实验中，实验 1 的 K-Means 实际迭代次数为 2，实验 2 的 K-means 实际迭代次数为 3。实验结果（表 1）显示，多质心法的平均训练速度是 Rocchio 的 2.8 倍，平均过滤速度是 Rocchio 的 5 倍，这样的效率完全能满足快速训练、快速过滤的要求。多质心法的查全率比 Rocchio 平均高出 6.75%，查准率比 Rocchio 平均高出 4.14%，过滤效果提高明显。

表 1 实验结果

方法	训练总用时(s)	训练速度(秒/篇)	过滤总用时(s)	过滤速度(秒/次)	召回率(%)	准确率(%)
实验 1 Rocchio	6.27	0.016	0.8594	0.001	83.5	87.43
Multi-Centroid	16.47	0.041	3.219	0.005	90.5	91.88
实验 2 Rocchio	6.33	0.016	0.5522	0.001	86.5	89.64
Multi-Centroid	19.43	0.049	1.717	0.004	93	93.47

由于宣扬法轮功和揭批法轮功两类文章有大量相同的词汇，更容易导致错分，而在色情文档中则没有类似的情况，所以实验 2 的查全率和查准率高于实验 1。另外测试集 1 和测试集 2 都有相同数量的不良文档，但测试集 1 的非不良文档比测试集 2 多 200 个，而几乎每个非不良文档都要跟 5 个质心进行比较，所以实验 1 的过滤效率比实验 2 要稍低。

4 结束语

本文针对 Rocchio 容易受到类别样本分布及噪声影响的而导致错误扩大类别范围的问题，提出对训练样本进行聚类，使用聚类形成的多个簇的质心向量替代单个质心向量作为过滤判定向量组的方法，并选择 K-means 作为聚类算法。实验表明，本方法的召回率和准确率都比 Rocchio 有明显的提高，而且训练效率和过滤效率都较高，适合用于不良文本的快速过滤。本文在选择 K-means 初始簇质心时采用的是随机选取的方式，我们进一步的工作是考虑优化初始簇质心的选择方法，提高聚类的效果和效率。

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)：4-5，8.

[2] 陈新淼. 不良信息过滤技术的法律思考[J]. 计算机安全，2004(5)：8-10.

[3] Christopher D Manning，Prabhakar Raghavan，Hinrich Schütze. Introduction to information retrieval[EB/OL]. (2010-09-23). http://nlp. stanford. edu/IR-book/.

[4] Jiawei Han，Micheline Kamber. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007.

[5] 苏金树，张博锋，徐昕. 基于机器学习的文本分类研究进展[J]. 软件学报，2006，17(9)：1848-1859.

[6] 谭金波，李艺，杨晓江. 文本自动分类的测评研究进展[J]. 现代图书情报技术，2005(5)：46-49，14.

[7] 周雪忠. 中文文本分类特征表示及分类方法比较研究[C]. Advances in Computation of Oriental Languages，北京：清华大学出版社，2003.

[8] 刘远超，王晓龙，徐志明，等. 文档聚类综述[J]. 中文信息学报，2006，20(3)：55-62.

（责任编辑：邓大玉）