

基于改进 KNN 算法实现网络媒体信息智能分类

柴春梅, 李翔, 林祥

(上海交通大学信息安全工程学院, 上海 200240)

摘要:在互联网资源迅速膨胀的今天,面向重要网络媒体海量发布信息实现智能分类,能在很大程度上解决目前网上信息杂乱的现象,对于网络信息监管、舆论引导工作有着深远的意义。鉴于此,基于改进 KNN 算法实现网络媒体信息智能分类,并进一步验证改进算法的有效性。实验结果表明改进 KNN 算法能对网络媒体信息进行有效分类,算法性能指标达到网络监管工作关于信息分类的业务需求。将改进 KNN 算法实现网络媒体信息智能分类是可行、有效的。

关键词:智能分类;KNN 算法;查全率;查准率

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2009)01-0001-04

Implementation of Information Intelligence Classification on Internet Media Based on Improved KNN Algorithm

CHAI Chun-mei, LI Xiang, LIN Xiang

(School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: Today, the resources of network inflate quickly, it has momentous significance for the task of the surveillance and management of network and leading the public to carry out the intelligence classification of the massive amount of information that released by the important network medium. To a large extent, it can solve currently the information of disorderly phenomenon. Owing to this, mainly implements the information intelligence classification on network media, based on the improved KNN algorithm, and then validates the validity of this improved algorithm. The result of the experiment indicates that the improved KNN algorithm can realize an effective intelligence classification to information that network medium released, the index of algorithm's performance can achieve the business demands which the work of surveillance and management of network requires the information intelligence classification. So it's feasible and effective to apply the KNN algorithm to the intelligence classification of the massive amount of information that the network medium released.

Key words: intelligence classification; KNN algorithm; recall; precision

0 引言

中国互联网络信息中心(CNNIC)2008年1月17日发布的《第21次中国互联网络发展状况统计报告》^[1]中显示,截止2007年12月31日,我国网页数为84.7亿个,年增长率达到89.4%,网上信息资源的增长速度非常迅猛。从网页长度看,网站总字节数已达到198 348GB,平均每个网页的字节数为23.4kB。从网页内容看,仍是文本居多,占网页总数的87.8%。因此,在当前互联网已经成为重要网络媒体、网上海量资源迅速膨胀的形势下,基于自然语言理解领域,传统文本分类算法实现互联网媒体信息智能分类,对于互

联网信息监管、舆论引导工作拥有广泛而深远的意义。

1 文本分类及其研究现状

1.1 文本分类概念

文本自动分类^[2](Autonomous Text Categorization, ATC)是指在给定的分类体系下,根据文本的内容用计算机程序确定文本所属类别的过程。一般采用机器学习的方法进行自动文本分类,即:基于训练集的文本自动分类。文本分类的目的是根据给定的已知训练样本求取对系统输入输出之间依赖关系的估计,使它能够对未知输出做出尽可能准确的预测。

文本分类在其本质上就是一个模式识别的过程,在对其设计开发时可分为如下四个主要环节^[3]:

(1)文本预处理,主要包括对训练和测试语料库的分词和对所分单词词频统计的过程。

(2)文本特征描述,主要包括模型确立、特征选择

收稿日期:2008-05-01

基金项目:国家自然科学基金项目(60502032);上海市科技计划项目(065115020)

作者简介:柴春梅(1983-),女,浙江衢州人,硕士研究生,研究方向为互联网内容安全;李翔,副教授,研究方向为网络内容安全。

和文本特征表达等内容。

(3) 文本分类算法训练阶段。

(4) 新文本实时分类阶段。

1.2 文本分类研究现状

传统分类算法类型非常之多,如简单 Bayes(NB)算法,决策树/决策规则分类算法(Decision Tree/Rules algorithm)和 KNN 算法等。但将传统算法直接应用于互联网媒体海量发布信息智能分类时,算法实际效果普遍不佳或者性能不稳定^[4]。鉴于此,文中工作是基于改进 KNN 算法实现重要互联网媒体海量发布信息智能分类,并验证改进算法的有效性。选用的文本分类算法是 KNN 算法,原因在于 KNN 算法是目前分类性能最好的算法之一,实现非常简单有效。同时算法分类效率高,适用于海量互联网文本信息分类处理。

2 KNN 算法应用于网络文本分类具体过程

2.1 文本预处理

与之后的特征表达和分类算法的要求相应,较为复杂的文本预处理包括文本去噪、分词、单篇文本词频统计等处理过程。由于文中所应用的是以单词为基本特征描述单位的方法,在文本预处理阶段主要进行了分词与词频统计的处理。

分词过程即对待处理文本进行逐字扫描,对已有词库进行匹配而得出单词的过程,由于扫描过程的不同,该过程分为单向匹配和双向匹配等方法,单向匹配是指从文章开头开始,向后逐字读入字符进行匹配。与之相应,双向匹配还要加上从结尾开始向前逐字读入字符进行匹配检验的过程,以达到分词的准确与不遗漏。由于分词准确度双向配法比较有效,这里使用了双向匹配,并且在匹配过程中使用了二分查找的方法以提高匹配速度。

2.2 文本特征描述

2.2.1 文本模型表达

由于文本不能被分类算法所使用,所以在对文本进行特征描述前,首先要建立文本表示模型,主要的建模方法有布尔模型、概率模型和向量空间模型,其中 G. Salton 提出的向量空间模型较为常用,在之后分类算法的使用效率和计算复杂度上也有着良好的表现,它的基本思想是把文本 d 看作是向量空间中的一个 n 维向量 $(w(t_1), w(t_2), \dots, w(t_n))$, 其中 t_1, t_2, \dots, t_n 为表示文本的 n 个特征,在文中所选用的分类方法中表示经过降维处理后的关键词,而 $w(t_k), k = 1, 2, \dots, n$ 是第 k 个特征在该文本中的权重。

2.2.2 特征抽取

在以单词作为向量空间基本维度时,由于中文特

征数量大必然造成算法计算维数高、计算量大的后果,而如此高维的特征对之后的分类算法学习过程未必是有意义的^[5]。特征抽取在整个分类过程的效率与准确度上起着尤为关键作用,是一个分类算法成功与否的关键。而常用的特征评估方法也已经较为成熟,主要有如下几种:词频(DF, Document frequency)、信息增益(IG, Information gain)、互信息(MI, Mutual information)、开方拟合检验(Chi-square)和几率比(OR, Odds ratio)等。

由于几率比函数可以综合考察正样本(属于该类)和负样本(不属于该类)的情况,在分类过程中有着较好的表现,因此文中选用几率比函数,如式(1)所示。

$$OR(t_k, c_i) = \frac{P(t_k | c_i) \cdot (1 - P(t_k | \bar{c}_i))}{(1 - P(t_k | c_i)) \cdot P(t_k | \bar{c}_i)} \quad (1)$$

其中 $P(t_k | c_i)$ 表示在训练样本集属于 c_i 类文档中出现 t_k 特征的概率,而 $P(t_k | \bar{c}_i)$ 表示在不属于 c_i 类文档中出现 t_k 特征的概率。综上所述,通过计算 $OR(t_k, c_i)$,便可以得到特征项 t_k 对类别 c_i 的代表程度。当 $OR(t_k, c_i)$ 越大时,其代表程度越高。几率比算法的输出是对应 $OR(t_k, c_i)$ 值最大的前 N 个特征词汇,作为训练样本的特征库参与新文本实时分类阶段的运算,如式(2)所示。

$$OR_{\max}(t_k) = \max_{i=1}^{|C|} OR(t_k, c_i) \quad (2)$$

不过纯粹选取 $OR(t_k, c_i)$ 值最大的前 N 个特征词汇作为训练样本特征库的方法,往往存在训练样本“不可描述”的问题,即部分训练文本不包含任何选取出的特征项。这是由于简单根据 $OR(t_k, c_i)$ 值大小确定训练样本特征库时,选取出的特征词汇通常会落入容易使用特征词汇描述的部分类别,从而导致剩余类别特征词汇匮乏,2.3.1 节将着重解决该问题。

2.2.3 向量表达

特征抽取过程之后,就要对文档集合进行向量表达,对每个文档的特征项进行权值计算,在文中实现的过程中使用两种权值向量表示方法:1. 布尔(bool)表达;2. 词频表达。

在布尔表达中,一篇文档中出现某一特征项则其维度权值计为 1,否则计为 0。而在词频表达中则将该特征项的出现次数作为其维度权值。对于改进 KNN 算法实现互联网媒体信息智能分类时,应用哪种表达方式更有效,并没有明确的结论,文中将通过实验来证明哪种表达方式更有效。

2.3 特征选择算法(OR 算法)及 KNN 算法的改进

考虑到传统文本分类算法直接应用于互联网媒体海量发布信息智能分类时,算法实际效果普遍不佳或者算法性能不稳定^[4],论文主体工作是基于改进 KNN

算法实现重要互联网媒体海量发布信息智能分类,算法改进工作主要集中于特征抽取方法的调整,以及算法分类运算环节的改进。

2.3.1 OR 算法的改进

正如前文所述,简单根据 $OR(t_k, c_i)$ 值大小确定训练样本特征库时,往往存在训练样本“不可描述”的问题。这里首先对于分类算法特征抽取环节的 OR 算法进行相应调整,主要包括两方面:一是在计算 $OR(t_k, c_i)$ 值前去除训练样本分词结果中的低频词;二是对每个训练类别分别抽取 $OR(t_k, c_i)$ 值最大的前 N 个词汇共同构成训练样本特征库。

在 OR 算法实现过程中,首先在训练样本分词结果中去除词频低于 5 的低频词汇,进而在每个训练类别中抽取 $OR(t_k, c_i)$ 值最大的前 200 个词汇共同构成训练样本特征库。在对 OR 算法调整后得到的训练样本特征库对于训练文本的“可描述”性能显著提高,此时每篇训练样本不论长短都能包含至少 5 个特征词汇,训练文本特征库对于每个训练类别文档都能实现充分描述,如图 1,图 2 所示。

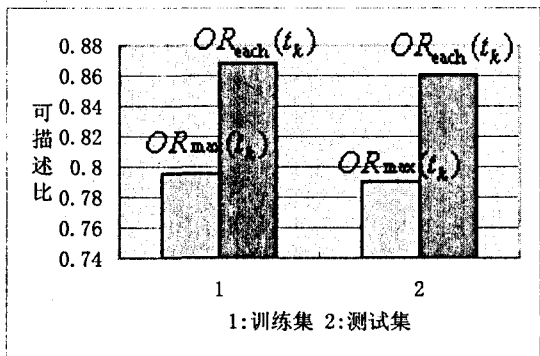


图 1 无超低频词汇处理

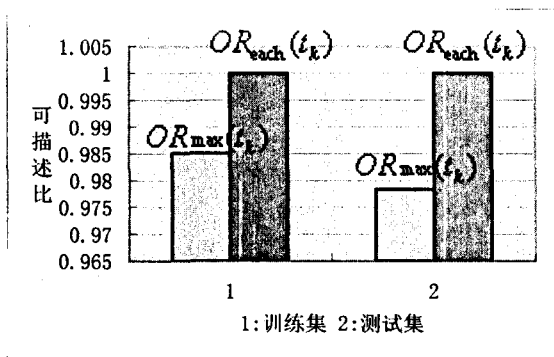


图 2 去除词频小于等于 5 的特征

2.3.2 KNN 算法的改进

在完成了文本预处理和文本特征描述后,就要确立分类训练算法以最终构建出文本分类算法。正如前文所述,在文本分类环节 KNN 算法性能佳,效率高,实现方法简单。传统 KNN 算法又称基于样例的分类

算法(Example - Based Learning),它并不在输入待分类文档之前,根据训练样本对向量空间进行分割,构成不同类别的区间。而是在输入待分类文档之后通过对每个训练样本的对比,找出与待分类文本最相似的 K 个文本。与待分类文本最相似的 K 个文本中选择包含最相似文本数最多的类别作为待分类文本类。而文中对 KNN 算法的改进是通过加权考虑最相似文本数和待分类文本与训练文本的相似度情况,最终确定待分类文本的类别。

对于向量 D, Q ,使用向量的夹角余弦 Cosine 值定义相似度,如式(3)所示。

$$\text{Sim}(D, Q) = \frac{D \cdot Q}{\|D\| \times \|Q\|} = \frac{\sum_i (a_i \times b_i)}{\sqrt{\sum_i a_i^2 \times \sum_i b_i^2}} \quad (3)$$

在待分类文本 d_j 进行判别时,计算它与各类之间的相似度 $\text{CSV}(d_j, c_i)$,如式(4)所示。

$$\text{CSV}(d_j, c_i) = \sum_{d_z \in Tr_k(d_j)} \text{Sim}(d_j, d_z) \cdot \Phi(d_z, c_i) \quad (4)$$

其中 $Tr_k(d_j)$ 表示与 d_j 最相似的 k 个文本集合,而

$$\Phi(d_z, c_i) = \begin{cases} 1, & \text{当 } d_j \text{ 属于 } c_i \text{ 类时} \\ 0, & \text{当 } d_j \text{ 不属于 } c_i \text{ 类时} \end{cases}$$

原先算法只是考虑将文本分入含相似文本最多的预设类别,而论文的改进方法是计算每个预设类别所含相似文本的相似度和,并将文本分入相似文本相似度总和最高的预设类别,加权考虑最相似文本数和待分类文本与训练文本的相似度情况。

3 算法性能评估

3.1 系统框图

基于改进 KNN 算法构建网络媒体信息智能分类系统主要涉及训练样本特征库生成,训练样本表示,待分类文本与训练样本相似度计算和待分类文档所属类别的判定,如图 3 所示。

3.2 算法评估方法

文档分类中普遍使用的性能评估指标有查全率(Recall,简记为 r)、查准率(Precision,简记为 p),如式(5)所示,式中变量 a, b, c 的具体含义详见表 1。

表 1 二值分类问题的列联表

	真正属于该类的文档数	真正不属于该类的文档数
判断为属于该类的文档数	a	b
判断为不属于该类的文档数	c	d

$$r = \frac{a}{a + c} \quad p = \frac{a}{a + b} \quad (5)$$

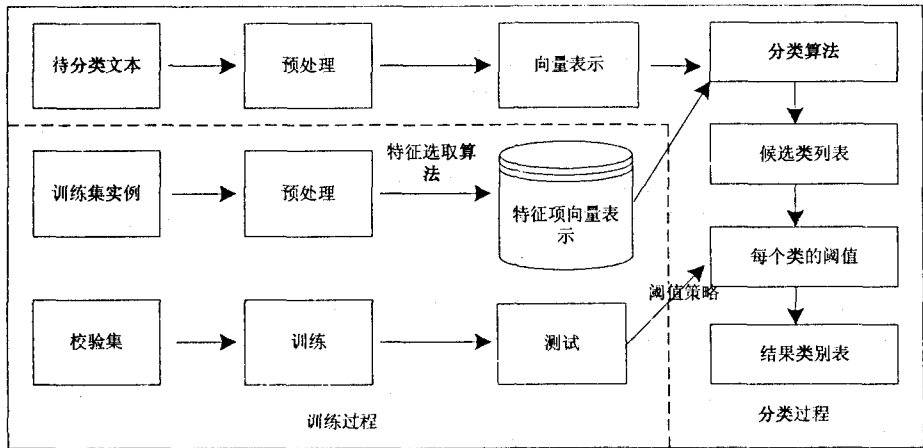


图 3 基于改进 KNN 算法构建网络媒体信息智能分类系统

宏观平均可以对单类赋值分类算法的性能进行统计计算,它先对每一个类按下式统计 r 、 p 值,然后对所有的类求 r 、 p 的平均值,如式(6)所示。

$$\bar{r} = (\sum_{c \in C} r_c) / |C|$$

$$\bar{p} = (\sum_{c \in C} p_c) / |C|$$

(6)

由于查全率和查准率共同反映了自动分类的质量,两者必须综合考虑,不可偏废。因此通常使用 F1 测试值统筹评估分类结果,其中 F1 测试值的计算公式如式(7)所示。

$$\text{F1 测试值} = \frac{\text{查全率} \times \text{查准率} \times 2}{\text{查全率} + \text{查准率}} \quad (7)$$

3.3 实验结果

最后通过编程实现的分类算法,利用了两种向量表示方法即布尔和词频,并对布尔表示方法和词频表示方法进行了 k 从 5 到 75 的计算,其中查全率与查准率均为宏观平均值,最后结果如表 2 表示。

表 2 布尔和词频的查全率与查准率

K		10	20	30	40	50	60	65	70	75
加权	r	0.893	0.907	0.915	0.924	0.924	0.919	0.922	0.924	0.924
	p	0.846	0.864	0.859	0.863	0.860	0.858	0.860	0.861	0.861
F1		0.869	0.885	0.886	0.892	0.891	0.887	0.890	0.891	0.891
词频	r	0.842	0.879	0.887	0.888	0.906	0.907	0.912	0.916	0.916
	p	0.791	0.819	0.827	0.827	0.845	0.846	0.849	0.851	0.851
F1		0.816	0.848	0.856	0.856	0.874	0.875	0.879	0.882	0.882

由表 2 中可见,布尔表示方法在 $k = 40$ 时,效果最好,这是因为文中所用的训练文本集较小,在 k 取较大值时会引入更多的噪声文本。此时,布尔表达: $r = 0.924$, $p = 0.863$; F1 测试值为 0.892。词频表示方法则在 $k = 70$ 时,效果最好,此时的词频表达: $r =$

0.916, $p = 0.851$; F1 测试值为 0.882。从而可以得出结论:布尔表示方法比词频表示方法好。

从图 4 不难看出,在 k 值相同的情况下,使用 bool 的向量表示方法能比使用词频的向量表示方法获得更高的查准率和查全率,即 bool 向量表示方法更加适合于网络媒体信息智能分类。

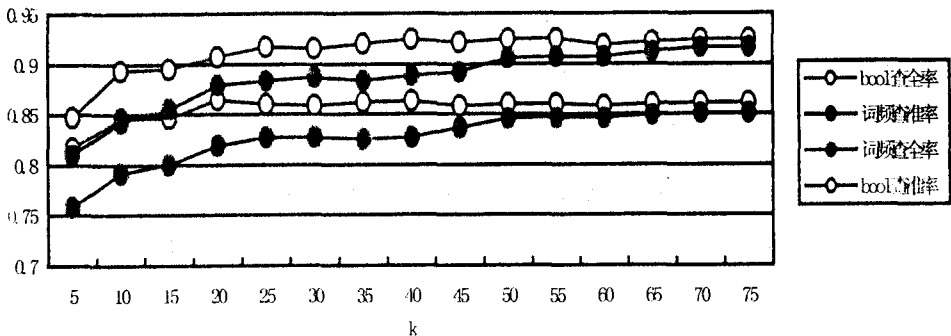


图 4 bool 表达和词频表达查全率与查准率比较

4 结束语

实验结果表明基于改进的 KNN 算法能对重要网络媒体海量发布的信息进行有效的智能分类,算法性能指标达到网络监管工作关于信息分类的业务需求。因此,将改进 KNN 算法实现互联网媒体信息智能分类是可行、有效的。

相信随着基于自然语言理解领域传统文本分类算法实现互联网媒体信息智能分类的不断发展,将对互联网信息监管、舆论引导工作起到广泛而深远的作用。

参考文献:

- [1] 中国互联网络信息中心. 第 21 次中国互联网络发展状况统计报告[DB/OL]. 北京:[出版者不详], 2008-01-17.
- [2] Lewis D D. Evaluating and optimizing autonomous text classification systems[C] // In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval. Seattle, US:[s.n.], 1995.
- [3] 徐威,董渊,白若鸢,等. 针对中文文本自动分类算法的评估体系[J]. 计算机科学, 2007, 34(18): 177-179.
- [4] 张宁,贾自艳,史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程, 2005, 31(8): 171-172.
- [5] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.