短信文本分类技术的研究

王文霞,王春红
(运城学院 计算机科学与技术系,山西 运城 044000)

摘 要:短信作为一种重要的交流手段,发挥着越来越重要的作用。但伴随着短信的广泛使用,垃圾短信则严重影响着人们的生活,因此文中基于短信文本特征词对短信进行分类研究。其中,TF-IDF 特征词权重计算方法是对文本词汇权重计算的一种经典算法,得到了广泛应用。但此方法为了简化计算,忽略了词语之间的相互关系。针对此问题,依据同一短信文本中的词汇之间存在的相互关系,文中对权重计算法进行了调整,提出了基于模糊 K 均值的短信文本分类算法。即先将短信文本集用 TF-IDF 算法处理,得到词汇-文本集,再用模糊 K 均值算法对得到的词汇-文本集进行处理。最后通过实验,验证了基于模糊 K 均值的短信文本分类算法,其分类结果的查全率和查准率都较高,有效辨别了垃圾短信。

关键词:短信文本分类;向量空间模型;模糊聚类;模糊 K 均值
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)05-0145-04
doi:10.3969/j.issn.1673-629X.2016.05.031

Research on Text Classification Technology for Message

WANG Wen-xia, WANG Chun-hong
(Department of Computer Science and Technology, Yuncheng University, Yuncheng 044000, China)

Abstract: As an important means of communication, SMS plays an increasingly important role. But along with the extensive use of SMS, SMS spam seriously influences people's lives. Therefore, the classification of SMS is researched based on the keywords in this paper. TF-IDF weight calculation method is a classical algorithm to calculate the text word weight, which is widely used. But in order to calculate simply, this method ignores the mutual relations between words. Aiming at this problem, based on the same relationship between words in the text messages, in this paper, the weighting method is used for adjusting, it puts forward the text classification based on fuzzy K-means algorithm. The text set is processed by TF-IDF algorithm, getting a vocabulary-text set. Then fuzzy K-means algorithm is used to get a vocabulary-text set. Finally, through the experiment to verify the text classification based on fuzzy K-means algorithm, the classification results of recall and precision is high.

Key words: text categorization; vector space model; fuzzy clustering; fuzzy K-means

0 引 言

短信业务作为目前的一种重要通信手段,具有短小、迅速、简便、便宜等诸多优点。据中国新闻网统计,到 2010 年,中国的手机用户数量达到近 7.4 亿,2009 年短信发送量日均达到了 21 亿条,全年各类短信发送量达到 7 840.4 亿条[1]。根据中国互联网协会 2008 年年初发布的一项调查,中国手机用户平均每周收到的垃圾短信竟然多达 8.29 条,每周收到 40 条以上的居然达到了 6.25%。在飞速的发展过程中,短信业务在给广大使用者带来方便的同时,也出现了很多问题,比如泛滥的垃圾短信、诈骗短信、谣言短信等等。这些垃圾短信给手机用户带来了很大的危害,因此需对垃圾短信进行过滤。

文中将自然语言文本处理运用到手机短信的分类研究[2-5]中。通过对短信文本特点的分析,实现对短信文本的分类。利用文本分类算法对短信信息进行分类,常用的分类算法有:决策树、支持向量机[6-9]、粗糙集和贝叶斯算法[10]。由于短信内容较少,依据同一短信文本中的词汇之间存在的相互关系,文中通过对经典的 TF-IDF 权重计算法的调整,并采用了模糊聚类算法,实现对短信文本的分类,达到了提高短信文本分析准确性的效果。

收稿日期:2015-07-22 **修回日期**:2015-11-05 **网络出版时间**:2016-03-22
基金项目:国家自然科学基金资助项目(11241005);山西省高等学校教学改革研究项目(J2012098);运城学院教学改革研究项目(JG201418)
作者简介:王文霞(1979-),女,讲师,硕士,研究方向为数据挖掘及算法分析;王春红,教授,研究方向为信息检索及算法分析。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160322.1522.092.html

1 垃圾短信概述

1.1 垃圾短信的概念、特点、分类

没有经过接收者允许而收到的,内容具有违法性、欺骗性或广告性,并且侵犯了人们的合法权益,这样的短信被称之为垃圾短信。垃圾短信具有以下特点:骚扰性,未经接收者同意发布且具有广告性质,具有违法犯罪的内容等等。垃圾短信一般分为商业广告信息、非法制作各种票或证的信息、诈骗信息、赌博信息等。诈骗短信已成为危害社会治安秩序的一大公害。

目前,我国出现的诈骗短信共有三类:

1)手机费诈骗。

(1)通过赠送话费来骗取手机费:利用人们贪图小便宜的心理,使用户上当;

(2)通过朋友点歌或接收彩信来骗取手机费:人们往往以为是自己的朋友为自己点歌,所以就会毫无防备地回消息,造成手机费被骗;

(3)以冒充老朋友的身份骗取电话费:这种短信的迷惑性相当大,人们很容易上当受骗;

(4)以听取心里话的方式诈骗手机费:主要利用用户的好奇心理,诱使用户受骗。

2)银行卡诈骗。

一般是团伙作案,犯罪分子先利用短信群发器发送消息,对于上当的人,他们假扮银行工作人员、警察、银行管理中心人员等,让上当者成功地将钱存入其他账户;这种短信主要是利用用户对自身财产安全关心的心理。

3)现金诈骗。

(1)以谎称办假证、走私军火、售枪支弹药、招嫖或者提供其他违法服务或物品的方式诈骗现金:主要利用用户想走捷径的心理,将钱骗走;

(2)以谎称中奖骗取现金:这种短信利用用户贪小便宜心理,当用户联系时他们会要求先交一部分个人所得税等一系列费用,然后卷着钱财逃之夭夭[11]。

1.2 垃圾短信的危害

伴随着智能移动设备的普及,短信业务迅猛发展,垃圾短信也日益猖獗,已严重扰乱了人们正常的工作和生活,非常不利于社会稳定与和谐,主要表现如下:

(1)影响人们的正常工作和生活。无论接收者是否愿意,垃圾短信都会不分时段地发到接收者的手机。接到一条短信后,用户最少要花 10 s 来判断是不是垃圾短信,一天收到十几条,就需要花几分钟来查看,严重浪费了用户的时间。不管你看不看短信,都会收到短信铃声的骚扰,让用户苦不堪言,严重影响用户的工作和生活。

(2)扰乱社会秩序。垃圾短信为办假学历、假证件、出售黑车等非法行为提供了一种安全、廉价的业务促进方式,使社会秩序被严重扰乱。甚至有些垃圾短信包含着低级下流、污染社会风气的内容,直接影响青少年的身心健康[12]。

(3)垃圾短信已成为犯罪分子实施诈骗的载体。一些不法分子利用手机散布谣言,散布邪教和封建迷信的思想,煽动民众,造成民族关系紧张,影响社会稳定。不法分子通过抓住人们的心理,群发一些迷惑性短信,骗取信任,获得资金。

(4)影响正常通信。垃圾短信一般都是群发,数量极大,传输时会占用大量的通讯资源,严重的甚至会导致堵塞,使通信中断。

1.3 垃圾短信的处理

垃圾短信采用文本形式表示信息,首先需要把它转变成计算机可识别的形式。文中采用的是空间向量模型即 VSM。下面介绍一些关于 VSM 的基本概念:

(1)特征项:指文本中能够代表该文本特点的基本语言单位。

(2)特征项权值:指特征项代表文本的能力的大小。特征项权值计算方法有很多,例如:布尔权重计算、平方根权重计算、TF-IDF 权重计算等,其中 TD-IDF 权重计算最为常用。文中对于文本集的加权计算采用这种方法。

(3)文本向量:设文本集合中共有 m 个不同的特征项,分别计算出文本特征项的权值,由这些特征项权值所构成的向量称为文本向量[13]。

接下来详细介绍一下 TF-IDF 权重计算:

TF-IDF 是一种基于统计分析的方法,用以获取字词在一个文件集或一个语料库中某文本的重要程度。TF-IDF 权重计算的出发点是字词的重要性会随着它在文本中出现的次数增加,但同时会随着它在语料库中出现的频率下降[14]。其主要思想是:如果某个词或短语在某个文本中出现的频率高,而在其他文本中又很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。

TF-IDF 计算方法中有两个重要参数:

(1)TF 词频。

它是指特征项在文本中出现的频率,计算公式为:

$$\mathrm{tf}_{ik} = \text{特征项 } t_k \text{ 在文档 } d_i \text{ 中出现的频率} \quad (1)$$

(2)IDF 反文本频率。

它是对特征词在文本集中分布情况的量化,用于衡量该特征词区分不同文本的能力,常用计算公式为:

$$\mathrm{idf}_k = \log(N/n_k + 0.01) \quad (2)$$

其中:N 代表文本集所有文本的个数;n_k 代表文本集中出现特征词的文本数。

TF-IDF 权重计算方法,是 Salton 和 McGill 基于香农信息理论提出的一种方法。该方法已成为目前文本

聚类和分类中最常用的方法。它是将词频和反文档频率两方面因素相结合来得到特征词的权重值,计算公式为:

$$w_{ik} = \mathrm{tf}_{ik} \times \mathrm{idf}_k = \mathrm{tf}_{ik} \times \log(N/n_k + 0.01) \quad (3)$$

2　模糊 K 均值

Bezdeck 等提出了模糊 K 均值算法。模糊 K 均值算法将模糊原理与经典 K 均值算法相结合,是一种非监督聚类算法。其基本思想是按照一定的模糊隶属度将每个数据对象分配到某个聚类中,使得不同类中的数据对象具有较低的相似性,同一个类中的数据对象具有较高的相似性。该算法将分好的簇看做是模糊集合,一个簇对应一个模糊集合,用隶属度函数度量每个数据属于某个簇的可能性,然后依据最大隶属度原则将数据分配到隶属度最大的簇中。

2.1　算法基本思想

模糊 K 均值算法是基于最小化以下目标函数[15]:

$$J(U,V) = \sum_{i=1}^{K}\sum_{j=1}^{N} (\mu_{ij})^m d^2(x_j, v_i), 2 \leqslant K \leqslant N \quad (4)$$

其中:参数 m 是 μ_{ij} 的加权指数,是任意一个大于1的正数,并且 $\forall j, \sum_{i=1}^{K}\mu_{ij} = 1$,它控制聚类的模糊性; x_j 是第 j 个数据点; v_i 是第 i 个聚类中心; μ_{ij} 是 x_j 对于聚类 i 的隶属度; N 是聚类数据的数目; K 是聚类数目; $d^2(x_j, v_i)$ 是 x_j 与 v_i 之间的距离,一般常用的距离为 Euclidean 距离: $d^2(x_j, v_i) = \| x_j - v_i \|^2$。

2.2　算法描述

模糊 K 均值算法描述如下:

(1)给定 K 个类别,参数是 m,允许误差范围是 $l \in (0,1)$。

(2)初始化聚类中心 v_i, $i = 1,2,\cdots,K$,一般从 N 个数据点中任意选择 K 个数据点作初始聚类中心。

(3)根据式(5)计算所有聚类数据点对于每一个聚类中心的隶属度。

$$\mu_{ij} = \frac{(1/d^2(x_j, v_i))^{1/(m-1)}}{\sum_{i=1}^{K}(1/d^2(x_j, v_i))^{1/(m-1)}} \quad (5)$$

(4)根据式(6)修改聚类中心,并且更新隶属度 μ_{ij} 为 $\hat{\mu}_{ij}$。

$$\hat{v}_i = \sum_{j=1}^{N}(\mu_{ij})^m x_j \Big/ \sum_{j=1}^{N}(\mu_{ij})^m \quad (6)$$

(5)若 $\max_{i,j}|\mu_{ij} - \hat{\mu}_{ij}| < l$,则停止,否转第 4 步。

3　基于模糊 K 均值的短信文本分类算法

利用模糊 K 均值实现短信文本分类算法描述如下:

(1)输入文本集合中的特征项,建立特征项库。

(2)将文本内容输入数据库,建立文本信息库以及文本段信息库。

(3)对每个文本段信息利用 TF-IDF 权重计算公式算出每一个特征项的权值,构造文本向量信息库。

(4)用模糊 K 均值算法对文本向量进行处理。需要明确要处理的样本数、每一行的特征项个数、要分的类别数、迭代的次数、聚类的精度等等。

(5)输出一个隶属度矩阵,获得文本分类结果。

基于模糊 K 均值的短信文本分类算法的基本思想是首先收集待处理的短信文本集,接着要对短信文本进行分词;然后建立特征项集,利用 TF-IDF 对每个特征项进行加权计算,得到文本向量,构建“词汇-文本”矩阵;最后用模糊 K 均值算法对“词汇-文本”矩阵进行处理,输出一个隶属度矩阵。具体的算法设计如图 1 所示。

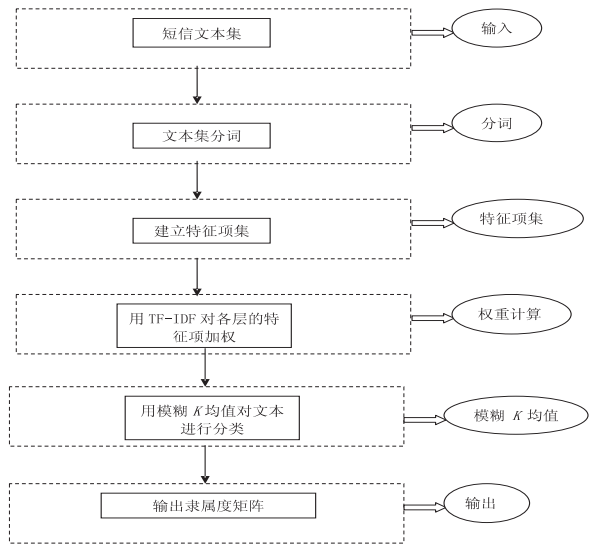

图 1　算法流程图

4　实验结果及分析

根据文本检索的度量标准,文中定义了两个评估指标,即查准率(Precision)和查全率(Recall),对基于模糊 K 均值的短信文本分类算法进行了有效性验证。

其中:查准率 p 是指实际相符的文本占属于类别 C_i 的所有文本的比例;查全率 r 是指正确归类的文本占专家判定的应属于类别 C_i 的所有文本的比例。两项指标分别定义如下:

$$p = \frac{\text{正确聚类的文本数}}{\text{实际聚类的文本数}} \quad (7)$$

$$r = \frac{\text{正确聚类的文本数}}{\text{类中应有的文本数}} \quad (8)$$

基于从互联网上收集的商业广告型短信、诈骗短信、非法制作各种票或证的短信、赌博类短信四方面的大量文本,分别从中各随机选取 10 个文本,共 40 个。

这 40 个文本分别按商业广告型短信、诈骗短信、非法制作各种票或证的短信、赌博类短信的次序排列,并对其进行预处理,进而基于模糊 K 均值聚类算法实现了文本分类。实验结果如表 1 所示,列出了 10 个文本的隶属度矩阵,商业广告型短信和诈骗短信各 2 个,非法制作各种票或证的短信和赌博类短信各 3 个;表 2 给出了每个文本所属的类。

表 1 输出的隶属度矩阵

样本	商业广告	诈骗	非法制作票证	赌博
1	0.144 369	0.070 017	0.751 968	0.032 046
2	0.022 047	0.008 520	0.966 048	0.003 385
3	0.020 407	0.007 160	0.969 744	0.002 689
4	0.573 843	0.102 105	0.296 756	0.027 296
5	0.893 444	0.061 707	0.034 107	0.010 742
6	0.036 764	0.922 644	0.007 510	0.033 082
7	0.048 314	0.912 504	0.009 002	0.030 101
8	0.048 296	0.172 401	0.016 330	0.762 973
9	0.002 895	0.008 368	0.001 103	0.987 634
10	0.015 781	0.050 898	0.005 671	0.927 650

表 2 40 个样本的分类结果

类别	样本
商业广告	{1,2,4,5,6,7,8,9,10}
诈骗	{11,12,13,14,3,15,16,17,19,20}
非法制作票证	{21,22,23,24,25,26,27,29,30,31}
赌博	{18,28,32,33,34,35,36,37,38,39,40}

为了验证该算法的有效性,将该算法聚类分析结果与人工分类的结果进行了对比,如表 3 所示;并采用了聚类分析的两个评价标准—查准率和查全率对聚类结果进行量化分析,其结果如表 4 所示。从这两个表可以看出,基于模糊 K 均值对文本分类,其查准率和查全率都较高。

表 3 模糊 K 均值聚类分析最终结果

样本	最大的隶属度值	分类结果	人工分类
1	0.751 968	非法制作票证	非法制作票证
2	0.966 048	非法制作票证	非法制作票证
3	0.969 744	非法制作票证	非法制作票证
4	0.573 843	商业广告	商业广告
5	0.893 444	商业广告	商业广告
6	0.922 644	诈骗	诈骗
7	0.912 504	诈骗	诈骗
8	0.762 973	赌博	赌博
9	0.987 634	赌博	赌博
10	0.927 650	赌博	赌博

5 结束语

文中提出的基于模糊 K 均值的短信文本分类算法,很好地克服了经典 TF-IDF 权重计算中忽略了词语之间的相互关系的弊端。实验结果表明,该聚类算法大大地改善了短信文本聚类的效果,查全率和查准率都较高。

表 4 查准率和查全率

类别	查准率/%	查全率/%
商业广告	100	90
诈骗	90	90
非法制作票证	90.9	90
赌博	83.3	90

参考文献:

[1] 刘国香,张钧锋. 垃圾短信分类方式的探讨[J]. 沧州师范专科学校学报,2011,27(4):122-124.

[2] Patel D, Bhatnagar M. Mobile SMS classification: an application of text classification [J]. International Journal of Soft Computing and Engineering, 2011, 1(2): 47-49.

[3] Liu Wuying, Wang Ting. Index-based online text classification for SMS spam filtering[J]. Journal of Computers, 2010, 5(6): 844-851.

[4] Li Feng, Li Jigang. Studying of classification Chinese SMS message based on Bayesian classification[J]. Journal of Theoretical and Applied Information Technology, 2012, 44(1): 141-146.

[5] 杨 柳,殷 钊,滕建斌,等. 改进贝叶斯分类的智能短信分类方法[J]. 计算机科学,2014,41(10):31-35.

[6] 李 慧,叶 鸿,潘雪瑞,等. 基于 SVM 的垃圾短信过滤系统[J]. 计算机安全,2012(6):34-38.

[7] 冯欧鹏. 垃圾短信过滤中字特征与词特征对过滤效果的比较研究[D]. 北京:北京邮电大学,2011.

[8] 徐 易. 基于短文本的分类算法研究[D]. 上海:上海交通大学,2010.

[9] Lan Man, Tan C L, Su Jian, et al. Supervised and traditional term weighting methods for automatic text categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721-735.

[10] 张 兢,候旭东,吕和胜. 基于朴素贝叶斯和支持向量机的短信智能分析系统设计[J]. 重庆理工大学学报:自然科学,2010,24(1):77-81.

[11] 赵晓芳. 短信诈骗的类型、法律定性及应对策略[J]. 消费导刊,2008(2):125-125.

[12] 董月琴. 基于 Android 的垃圾短信处理系统的研究与设计[D]. 淮南:安徽理工大学,2011.

[13] 付克志,林鸿飞. 基于 N-Level VSM 在 Web 信息检索中的研究[J]. 计算机工程与应用,2006,42(19):158-160.

[14] 包金龙. 基于向量空间模型的信息检索系统的设计[J]. 情报杂志,2005,24(7):44-45.

[15] 叶吉祥,谭冠政,路秋静. 基于核的非凸数据模糊 K-均值聚类研究[J]. 计算机工程与设计,2005,26(7):1784-1785.