改进贝叶斯分类的智能短信分类方法

杨 柳 殷 钊 滕建斌 王 衡 汪国平
(北京市虚拟仿真与可视化工程技术研究中心(北京大学) 北京 100871)

摘 要 随着移动通信技术的不断发展,手机的普及率在不断上升,而短信作为传统的移动通信服务,长久以来一直在人们的日常生活中占据着极为重要的位置。可以说,短信在一定程度上记录了人们生活的轨迹。但是,现有的短信管理系统仅对短信进行以联系人为特征分类、以时间为顺序显示的简单非智能化的管理,导致了用户手机中各类短信混杂不清,短信的管理效率极低。通过研究短信的特征,分析传统的基于文档频率的特征值提取方法和基于互信息的特征值提取方法的优势与不足,提出了一种适用于短信的基于词频和互信息的特征值提取方法,并结合短信长度实现了一种改进的贝叶斯分类算法。实验证明,算法在进行短信分类时可以得到相当可观的召回率和准确率。

关键词 短信智能管理,文本分类,特征值提取,贝叶斯分类
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.007

Intelligent SMS Classification Method Based on Improved Bayes Classification Algorithm

YANG Liu YIN Zhao TENG Jian-bin WANG Heng WANG Guo-ping
(Beijing Engineering Technology Research Center of Virtual Simulation and Visualization (Peking University), Beijing 100871, China)

Abstract With the development of the mobile communication technology, the number of mobile phone users is increasing continuously. As a traditional mobile communication service, SMS occupies a very important position in people's lives. SMS messages record the track of one's life to a certain extent. However, the existing SMS management systems only manage our messages in an unintelligent way—classifying by contacts and showing in the order of sending time. As a result, different kinds of messages mix together and are hard to be managed. By studying the characteristics of SMS messages and analyzing the shortages of the traditional algorithm based on word frequency and the algorithm based on mutual information, we proposed a new feature selection algorithm for SMS messages based on both word frequency and mutual information and improved the accuracy of the Bayes classification algorithm using more features including the length of SMS messages. In the experiments, it is proved that this new algorithm can get a very good recall rate and accuracy rate when processing SMS messages.

Keywords Intelligent SMS management, Text classification, Feature selection, Bayes classification algorithm

1 引言

在信息化高度发达的今天,人们的交流也变得愈加快捷和频繁。在移动通信领域,短信凭借着它独特的优势,在人们的生活中占据着重要的位置。相比于其他通讯手段,短信具有普及性、实时性、异步性等优势。可以说,短信记录了人们生活的点点滴滴,一次约会、一次聊天、一次活动的通知、一则生日的祝福,都被一条条的短信记录下来。对于忙碌的现代人,他们已经没有时间像往常一样通过日记来记录下生活的琐事,而短信就是他们信息化时代下新的日记本。

然而,如今的手机短信管理软件看似琳琅满目,实际上大同小异,基本都是对短信进行以联系人为特征分类、以时间为顺序显示的简单非智能化的管理,这样就导致了人们的短信成为了时间线而不是日记本,因为谁都不会将自己的工作日记和情感日记写到一起。

对短信智能管理的核心是能够对短信进行准确的分类。但是,对正常用户短信的分类困难重重,主要体现在以下几个方面:

(1)类别区别不明显:不同的人有着不同的分类习惯,有的人喜欢按心情好坏,有的人喜欢按关系疏远,但是对于自然语言处理的方法来说,很多人为区分的类别对计算机来说却没有特别大的区别度。

(2)没有大规模训练集:短信是私人数据,基本没有人愿意共享,所以获得大规模训练集是很难的一件事情。

(3)文本太短:由于短信文本很短,这就导致特征值会多而散,非常不明显,很多的分类算法面对这种情况很难达到预

到稿日期:2013-03-20 返修日期:2013-06-20 本文受863计划重点项目(2011AA120301),国家自然科学基金项目(60925007,61173080,61232014)资助。

杨 柳(1990—),男,主要研究方向为人机交互,E-mail:yangliu713@126.com;殷 钊(1989—),男,硕士生,主要研究方向为人机交互;滕建斌(1989—),男,硕士生,主要研究方向为人机交互;王 衡(1960—),女,博士,副教授,主要研究方向为人机交互、图像处理与图形学等;汪国平(1964—),男,博士,教授,主要研究方向为计算机图形学、人机交互与虚拟现实等。

想的效果。

本文将自然语言处理运用到手机短信分类[1-3]研究中，通过分析短信的特点，综合用户习惯，合理运用近似和假设，详细研究适合短信的特征提取方法并改进贝叶斯分类算法，得到一个高效、准确的手机短信分类算法，提出了一种高效可行的短信智能管理方案。

2 文本的特征值提取

文本分类的一个重要的问题就是特征维度太高，且高维度的特征会在分类算法中带来很多问题。首先，维度过高会导致各维度之间的独立性变差，这会严重影响算法的准确率。其次，高维度的特征值会使算法效率下降，引入大量冗余的计算量。最后，高维度的特征值也会引入很多不必要的噪声。

所以，利用合理的特征值提取算法对特征空间降维是必要的。传统的文本特征值提取方法包括文档频率特征值提取方法和互信息特征值提取方法。

2.1 文档频率特征值提取方法

词条的文档频率(Document Frequency)是指训练集中出现该词条的文档数。基于文档频率的特征值提取方法基于一个假设：当一个词条的 DF 值小于某个阈值时，它并不具备或者很少具备类别区分的能力，需要将其作为噪声词去掉，从而达到降维的目的[4]。

设词条 t 的文档频率为 $P(t)$，阈值设为 th，则最终的特征值为：

$$Characteristic=\{t|P(t)>th\} \tag{1}$$

文档频率的特征值提取方法的优势在于该算法具有相对于训练集规模的线性复杂度，所以在大规模语料计算中很受欢迎。对于本文涉及到的手机短信处理，该算法的复杂程度和移动端的计算能力是相匹配的。

但是，文档频率的特征值提取方法的最大问题在于：对于那些“整体分散，类内集中”的训练集，该算法无法公平地取到所有类别的特征值。例如，如果将文本分为“社会”和“体育”两个类别，由于社会的词汇包罗万象，词汇多而分散，但是体育词汇则非常集中，因此最后使用该算法提取出来的特征值将有很多是体育词汇而社会词汇较少。对于本文所涉及的短信分类，这个问题将会愈加突出，因为短信没有大规模的语料库进行训练，而且词条“整体分散，类内集中”这个特点更加明显。

2.2 互信息特征值提取方法

互信息(Mutual Information)[5]是一个在统计语言模型中被广泛采用的概念。它表示了一个词条和一个类别的相关程度。词条 t 和文档类别 C 的互信息表示为：

$$MI(t,C)=\log\frac{F(t,C)}{F(t)\cdot F(C)} \tag{2}$$

其中，$F(t)$表示语料中出现词条 t 的文档频数，$F(C)$表示语料中类别 C 的数目，$F(t,C)$表示在类别 C 中出现 t 的文档频数。

一个词条在整个语料中的互信息值为：$MI(t)=\max_{i=1,\cdots,m} MI(t,C_i)$。

互信息特征值提取法成功地解决了文档频率特征值提取法中存在的问题，当一个词越多地出现在一个类而在其他类很少出现时，这个词条越能代表这个类，相应的互信息的值也就越大。

然而，互信息计算时完全没有考虑不同文档频率的词条对类别的判定能力的差异，导致该算法会向最终的特征值中引入大量的噪声。这样，一个极端的事情会在本文涉及的短信文本分类中出现，即当词条比较分散而训练集比较小时，如果一个单词 t 仅仅在 C_1 中出现，即使它的文档频率小到只出现了一次，因为这时 $F(t,C_1)=F(t)$，那么互信息值 $MI(t)=MI(t,C_1)=\log\frac{F(tC_1)}{F(t)\cdot F(C_1)}=\log\frac{1}{F(C_1)}$。这个结果显然是 C_1 与所有词条的互信息中的最大值，但是文档频率只有 1 的词条显然无法代表这个类别。当词条分散而训练集比较小的时候，这种词条会有很多，显然更好的做法是按照文档频率特征值提取法的思想，将其作为噪声处理掉。

3 短信分类方法

3.1 中文分词

中文分词是指将汉字序列分成单独的词的过程，分词是一切中文自然语言处理的基础。我们对不同的平台采用了不同的分词工具。

对于训练集的分词，可以在 PC 机上进行，因此我们采用中科院的汉语分词系统 NLPIR(又名 ICTCLAS2013)[6]进行分词。NLPIR 可同时进行分词和词性标注，并有较高的准确率。

然而，经过试验证明，NLPIR 在手机端分词并不合适。在开发原型系统的 Android 平台上，我们改进了轻量级的 Jcseg 开源分词系统来对手机短信进行分词。Jcseg 开源分词系统是使用 Java 开发，使用流行的 mmseg 算法实现的一个中文分词器。根据 Jsceg 开源分词系统官方 wiki 公布，其分词准确率达到了 98.41%[7]。

另外，出于效率和准确率的折中考虑，对分词结果，本文仅保留名词、动词和形容词 3 类作为特征词的候选词汇。

3.2 适用于短信内容的特征值提取方法

文档频率特征值提取方法过分强调文档频率而忽略了其与类别之间的关系，而互信息特征值提取方法又过分强调词条与类别的关系，忽略了文档频率的重要性。尽管这两种方法在各个领域广为应用并且取得了不错的效果，但是对于像短信这样的词条分散小训练集的文本显然不能适用。本文综合考虑文档频率和互信息两个因素，提出了一种适用于短信内容的特征提取方法。

我们引入一个新的概念“重要性(Importance)”，来表示一个词条在一个类别中的地位。词条 t 在类别 C 中的重要性用 $I(t,C)$来表示。

首先，显然地，如果一个词条在这个类别中出现的概率越高，则其重要性越大，因此，$I(t,C)\propto\frac{F(t,C)}{F(C)}$。

其次，如果一个类别中的平均文档越长，则说明一个词条对文档类别的重要性会越小，用 $avgLen(c)$来表示类别 C 中文档的平均长度，因此，$I(t,C)\propto\frac{1}{avgLen(C)}$。

最后，如果一个类别中的单词越多，则一个词条对文档的重要性会越小，用 $termNum(C)$来表示类别 C 中的总词条数，因此，$I(t,C)\propto\frac{1}{termNum(C)}$。

综合以上分析，我们得到一个词条在一个类别中的重要性推导公式：

$$I(t,C)=\frac{F(t,C)}{F(C)}\cdot\frac{1}{avgLen(C)}\cdot\frac{1}{termNum(C)} \tag{3}$$

为了避免 $I(t,C)$ 等于 0，我们采用拉普拉斯概率估计：

$$I(t,C)=\frac{1+F(t,C)}{2+F(C)}\cdot\frac{1}{avgLen(C)}\cdot\frac{1}{termNum(C)} \tag{4}$$

最终的重要性用对数来表示：

$$I(t,C)=\log(\frac{1+F(t,C)}{2+F(c)}\cdot\frac{1}{avgLen(C)}\cdot\frac{1}{termNum(C)}) \tag{5}$$

同样，一个词条在整个语料中的重要性为：

$$I(t)=\max_{i=1,\cdots,m} I(t,C_i) \tag{6}$$

最终，提取 $I(t)$ 最大的前 N 个词条作为整个语料库的特征词集合 $T=\{t_1,t_2,\cdots,t_N\}$。

3.3 改进的朴素贝叶斯分类方法

朴素贝叶斯分类器(Naive Bayes Classifier，NBC)[5]基于一个基本的假设：不同的特征值对于给定的类别的影响是相互独立的。

如果将训练样本分为 m 类，则对于每个新样本 d，其属于类别 C_i $(i=1,\cdots,m)$ 的概率是 $P(C_i\mid d)$，这样，具有最大 $P(C_i\mid d)$ 的类别 C_i 就是文档最终的分类结果。

具体的 $P(C_i\mid d)$ 计算方法如下：

根据贝叶斯定理，$P(C_i\mid d)=\frac{P(d\mid C_i)\cdot P(C_i)}{P(d)}$。

由于 $P(d)$ 对于所有的类别 $C_i(i=1,\cdots,m)$ 均相等，因此有，$P(C_i\mid d)\propto P(d\mid C_i)P(C_i)$。

由于 $P(C_i)$ 可能为 0，因此使用拉普拉斯概率估计：

$$P(C_i)=\frac{1+F(C_i)}{m+n} \tag{7}$$

其中，$F(C_i)$ 代表类别 C_i 的文档个数，n 表示文档总数。

文档 d 可以用其中提取的特征词集合表示 $T_d=\{t_j\mid t_j\in T,j=1,\cdots,k\}$，基于特征独立性假设，有：

$$P(d\mid C_i)=\prod_{j=1}^{k}P(t_j\mid C_i) \tag{8}$$

对于上式中 $P(t_j\mid C_i)$ 的计算，传统上有文档型计算公式和词频型计算公式。而在本文中，我们已经基于文档频率提出“重要性”的概念，因此以词条 t_j 进行类别 C_i 中的计算，即：$P(t_j\mid C_i)=I(t_j,C_i)$。

此外，短信具有一个很重要的特征——长度。我们引入一个长度评估 $L(d,C)$，评估因子越大，表明长度特征越吻合，$P(C_i\mid d)$ 越大。即，$P(C_i\mid d)\propto L(d,C_i)$。

假设类别 C 的文档平均长度为 $avgLen(c)$，待分类文档 d 的长度为 $Len(d)$，则：$L(d,C_i)=\frac{1}{K^{|Len(d)-avgLen(C_i)|}}$ (9)

参数 K 代表了长度特征对类别的影响度，称之为长度影响因子，若 K 越大，则长度特征对最终的结果影响越大。

综合以上推导，得到改进的贝叶斯分类算法，对于类别 $C_i(i=1,\cdots,m)$，待分类文档为 d，则 d 属于类别 C_i 的概率为：

$$\begin{aligned}P(C_i\mid d)&\propto P(d\mid C_i)P(C_i)L(d,C_i)\\&=(\prod_{j=1}^{k}I(t_j\mid C_i))\frac{1+F(C_i)}{m+n}\frac{1}{K^{|Len(d)-avgLen(C_i)|}}\end{aligned} \tag{10}$$

4 实验评估

本文设计了实验来对特征词提取算法和分类算法进行评价。我们分别采用传统的文档频率特征值提取方法、互信息特征值提取方法和本文提出的适用于短信的结合文档频率和互信息的特征值提取方法来提取特征值，并都应用到本文改进的贝叶斯分类算法以比较 3 种特征值提取方法在短信分类中的优劣。

在特征值提取算法和分类算法中，存在两个关键的参数——特征词数量 N 和长度影响因子 K，本文通过实验来确定它们较为合理的取值。之后，为全面评价和分析算法的效果，我们首先进行二分类测试，然后进行更加复杂的多分类测试。

4.1 评价标准

特征词提取是为了接下来的分类，所以特征值提取算法的评价可以由分类的效果得出。另外，为了更直观地表示特征值的选取对不同的类别是否均衡，我们引入一个最直观的评价指标：均衡度。

设类别数为 m，最终提取的特征词数目为 N。对任意特征值 t，对于文档频率特征值提取方法，记 t 的文档频率最高的类别为 C_k；对于互信息特征值提取方法，由 $MI(t)=\max_{i=1,\cdots,m}MI(t,C_i)$，将与 t 互信息最大的类别记为 C_k；对于本文改进的特征值提取方法，由 $I(t)=\max_{i=1,\cdots,m}I(t,C_i)$，将 t 的重要性最大的类别记为 C_k。

无论哪一种算法，特征值 t 在 C_k 中的运算值是对结果影响最大的，则称特征值 t 更趋向于类别 C_k。对任意 $C_i(i=1,\cdots,m)$，将趋向于 C_i 的特征值数量记作 $n(C_i)$，则均衡度为：

$$Eq=1-(\sum_{i=1}^{m}|\frac{n(C_i)}{N}-\frac{1}{m}|) \tag{11}$$

对于分类算法，我们使用传统的准确率(Precision)和召回率(Recall)来评价其优劣。对于类别 C_i，设分类结果得到的总样本数目为 $NE(C_i)$，其中真实属于 C_i 的样本数目为 $EF(C_i)$，整个测试样本中真实属于 C_i 的样本数目为 $NF(C_i)$，则有下面的定义：

准确率是指一个分类结果中分类正确的样本数占该分类结果总样本数目的比率，召回率是指一个分类结果中分类正确的样本数占该类别所有样本数目的比率：

$$P(C_i)=\frac{EF(C_i)}{NE(C_i)},R(C_i)=\frac{EF(C_i)}{NF(C_i)} \tag{12}$$

F 值是综合考虑准确率和召回率的一个评估指标：

$$F(C_i)=\frac{2P(C_i)R(C_i)}{P(C_i)+R(C_i)} \tag{13}$$

4.2 算法参数确定

我们通过实验对特征值和分类算法中的两个参数进行测试和确定，第一个是特征词数量 N，第二个是长度参数影响因子。实验数据采用 4.3 节中二分类测试的测试集。

首先，我们对特征词数量分别为 100，200，500，1000，2000，5000 的时候进行测试，因为在 5000 以后很多词频已经降为 1，所以没有继续测试的意义。我们发现，当特征词数量取值 2000 时，分类的准确率达到最大值，且两个类别的召回率都比较高。当特征词数量太少时，不足以充分表达一个类的特征；当特征词数量太多时，则会有太多的噪音，这些都会影响分类的效果。因此，特征词数量取 2000 比较合理。

然后，将长度参数影响因子 K 控制在 1.01～1.20 范围内进行测试，结果如图 1 所示。尽管当 K 取值 1.09 时分类的准确率达到最大值 98.260%，但类别 A 的召回率相对较低一些；而当 K 取值 1.07 时，准确率达到次大值98.251%，但两个类别的召回率也都处于较高的水平。因此，我们将长度影响因子取值为 1.07。

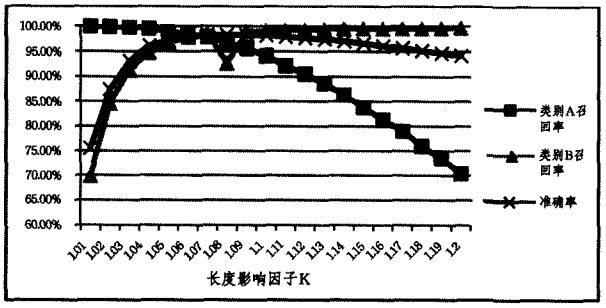

图 1　分类的各项评价标准与长度影响因子 K 的关系

通过图 1 的结果我们也可以看出，当长度影响因子越大时，长度特征对分类结果的影响也越大，一般长度波动较大的类别 A(普通短信)召回率总体逐渐减小，而一般长度波动较小的类别 B(祝福短信)召回率总体逐渐增大。

4.3　二分类测试

本文采用从互联网获得的祝福短信库和新加坡国立大学 2004 年收集的短信语料库(NUS SMS Corpus)[8]作为实验数据，分别随机取 70%作为训练集，30%作为测试集。

实验将短信分为普通短信和祝福短信两类，即式(10)中的 m 取 2，该测试集的特点是分类特征明显，所以最终的分类效果也比较好。训练集和测试集大小如表 1 所列。

表 1　二分类测试各类别训练集和测试集大小

	祝福短信(分类 A)	普通短信(分类 B)
训练集	4780	20769
测试集	2049	8901

由于短信是敏感数据，因此很难获得大规模的训练集。而且短信源于日常的会话，所以分类很难明确而具体。本文选择上面的数据集测试，是因为普通短信和祝福短信是很明确的两类，而且上万的数据量也足够具有代表性。在二分类测试中 3 种特征值算法的分类效果如表 2 所列。

表 2　二分类中 3 种特征值提取方法比较

	文档频率方法	互信息方法	本文方法
趋向于 A 的特征词数量	1242	0	918
趋向于 B 的特征词数量	758	2000	1082
均衡度	75.8%	0	91.800%
类别 A 召回率	99.951%	58.272%	97.852%
类别 B 召回率	35.193%	96.257%	98.342%
综合召回率	67.572%	77.265%	98.097%
综合准确率	47.344%	89.130%	98.251%
F 值	55.678%	82.758%	98.174%

表 3 是二分类测试中本文提出的特征值提取算法以及改进的贝叶斯分类算法对每个类别的测试结果。

表 3　二分类测试结果

	祝福	普通	综合
准确率	98.074%	98.428%	98.251%
召回率	97.852%	98.342%	98.097%

4.4　多分类测试

二分类测试中的测试集和分类方案是理想的——特征值明显，长度参数有一定影响。在实际的运用中，大部分对短信的分类效果则没有这么理想，分类的样本常常需要多分类，而且相互之间的特征值区分不明显。

我们选取一组来自互联网的测试集，将其分为 4 类：爱情、友情、节日和养生。同样，随即选取 70%作为训练集，剩余 30%作为测试集。各个类别的训练集和测试集大小如表 4 所列。

表 4　多分类测试各类别训练集和测试集大小

	爱情(分类 A)	友情(分类 B)	养生(分类 C)	节日(分类 D)
训练集	365	327	244	318
测试集	156	140	104	136

这组短信样本对我们的分类是相当不利的，首先是我们的长度影响因子基本无效，因为其长度大体相近。其次是它们的特征值并不明显，虽然通过人类的判断很容易识别是属于哪一个类别，但是对于计算机来说，它们是很相近的，比如下面一组短信：

爱情类：太阳的晨光一缕，是温暖的问候一句，早上好！阳光的午间播报，是时间的牵挂一句，还好吗？夕阳的剪影一抹，是真心诚意的告白一句，我爱你。

友情类：西伯利亚的寒流，冻不住朋友的牵挂；零摄氏度的大风，刮不走我的祝福；凌霜傲雪的腊梅，绽放吉祥甜蜜的祈盼；薄雾中晶莹的露珠，凝聚心间的温暖。

养生类：春天到，打个太极拳，健康“纵队”来组建；练个瑜伽操，快乐“海洋”来拓宽；勤加注意来保暖，幸福“核心”来提现；慢跑来做加餐，争取不做“宅男”。

节日类：父亲对儿女的爱，总是用严厉裹起来；儿女对父亲的爱，总是在心中紧埋；父亲节让父亲感受最温暖的表白。我们大声说出对父亲的爱，祝愿父亲们节日快乐！

在多分类测试中 3 种特征值算法的分类效果如表 5 所列。

表 5　多分类中 3 种特征值提取方法比较

	文档频率方法	互信息方法	本文方法
趋向于 A 的特征词数量	615	0	319
趋向于 B 的特征词数量	422	628	270
趋向于 C 的特征词数量	641	1372	952
趋向于 D 的特征词数量	322	0	459
均衡度	75.3%	0	54.8%
类别 A 召回率	87.821%	48.077%	85.897%
类别 B 召回率	88.571%	37.143%	87.857%
类别 C 召回率	91.346%	97.115%	93.269%
类别 D 召回率	58.823%	0	61.029%
综合召回率	81.640%	45.584%	82.013%
综合准确率	81.343%	42.537%	81.530%
F 值	81.491%	44.008%	81.771%

表 6 是多分类测试中本文提出的特征值提取算法以及改进的贝叶斯分类算法对每个类别的测试结果。

表 6　多分类测试结果

	爱情	友情	养生	节日	综合
准确率	86.292%	85.812%	90.691%	63.325%	81.530%
召回率	85.897%	87.857%	93.269%	61.029%	82.013%

4.5　结果分析

从上述两类测试的结果可以发现，无论是二分类测试还是多分类测试，本文提出的智能短信分类方法(包括特征值提取方法和改进的贝叶斯分类算法)都可以取得比较满意的效

果，准确率分别达到98.251%和81.530%，召回率分别达到98.097%和82.013%，各个类别的分类准确率和召回率也都较高。

在二分类测试中，互信息的特征值提取方法提取特征之后均衡度很差(0:2000)，而且该算法在多分类测试中结果也很差(准确率42.537%)。而在多分类测试中，文档频率的特征值提取方法提取特征之后的结果很好(准确率81.343%)，但是在二分类测试中结果却很差(准确率47.344%)。

导致这个结果的原因正是我们在前文中分析总结过的两种方法的缺点。

基于文档频率的特征值提取算法无法很好地处理“整体分散，类内集中”这种情况，而在上面的二分类测试中，祝福短信中的高频词很多，比如“幸福”、“快乐”等等，而普通短信中的词则较为分散，这样分类结果就很严重地偏向了祝福短信(召回率高达99.951%)，而普通短信的召回率只有35.193%。

基于互信息的特征值提取算法则因为没有考虑不同文档频率对类别的判别能力的不同，而夸大了某些低频率却单独在某个类别中出现的词的判别作用，这样的词在二分类中的普通短信和多分类的友情短信和养生短信中很多，这就导致了特征值选取严重偏向了这些类别。

结束语 本文着重研究了对手机短信的分类，通过综合考虑词频和互信息，改进了传统的特征值提取算法，以及对朴素贝叶斯分类进行了优化，得到了一种适用于短信的改进贝叶斯分类算法。从实验结果来看，本文的分类算法效果比较令人满意，即便是如多分类测试那样不如意的分类样本，也可以达到80%以上的准确率。此外，本文的算法可以合理地提取特征值，而且通过引入训练文本平均长度和词汇量因子，能够均衡非均衡训练样本带来的差异，从而不会出现结果偏向文本长或者词汇量小的类别。

短文本分类是一个复杂的方向，本文的方法在很多方面可以进一步提升。本文并没有对语义进行扩展，仅仅使用词作为特征进行文本分类研究，如何能够自动地得到文本中所带的语义信息，并将其合理地利用到分类算法中，这方面的研究是非常有意义的。另外，由于短信分类有时候很难找到合适的训练样本，因此这些训练就只能完全依赖于用户，但是用户能接受的训练操作次数是比较少的，这时如何完全利用用户的操作中所包含和隐含的有用信息，最大限度地获取该类别的特征信息，将是一个值得研究的问题。

最后，虽然本文研究了基于短信的智能分类，但短信只是一个出发点，如今层出不穷的IM(即时通讯)软件以及各类社交网络中的文本性质与短信大同小异，在进行管理和分类时遇到的问题也有类似之处，本文所关注和努力解决的问题，也正是这些人们日益依赖的平台所共同面临的问题。

参考文献

[1] Patel D, Bhatnagar M. Mobile SMS Classification: An Application of Text Classification[J]. International Journal of Soft Computing and Engineering, 2011, 1(1): 47-49

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

[3] Li Feng, Li Ji-gang. Studying of Classifying Chinese SMS Message Based on Bayesian Classification[J]. Journal of Theoretical and Applied Information Technology, 2012, 44(1): 141-146

[4] 陈艳秋. 有效特征值提取的快速中文文本分类[D]. 天津：南开大学，2007

[5] 李静梅，孙丽华，张巧荣，等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报，2003，24(1)：71-74

[6] 自然语言处理与信息检索共享平台[OL]. [2013-08-12]. http://www.nlpir.org

[7] Jcseg开源中文分词组件[OL]. [2013-08-12]. https://code.google.com/p/jcseg

[8] Chen Tao, Kan Min-yen. Creating a live, public short message service corpus: The NUS SMS corpus[J]. Language Resources and Evaluation, 2013, 47(2): 1-37

(上接第18页)

[16] Schatzmann J, Weilhammer K, Stuttle M, et al. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies[J]. The Knowledge Engineering Review, 2006, 21(2)

[17] Eyben F, Wollmer M, Graves A, et al. On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues [J]. Journal on Multimodal User Interfaces (JMUI) Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots, 2010, 3(1/2): 7-12

[18] Lee C, Jung S, Kim K, et al. Recent Approaches to Dialog Management for Spoken Dialog Systems[J]. Journal of Computing Science and Engineering, 2010, 4(1): 1-22

[19] Carolis B D, Pelachaud C, Poggi I, et al. Behavior planning for a reflexive agent[C]//Proceedings of the International Joint Conference on Articial Intelligence (IJCAI'01). Seattle, 2001

[20] Cerekovic A, Pejsa T. Pandzic Igors RealActor: Character Animation and Multimodal Behavior Realization System[J]. IVA: 2009, 5773: 486-487

[21] Van Welbergen H, Reidsma D, Ruttkay Z M, et al. A BML Realizer for continuous, multimodal interaction with a Virtual Human [J]. Journal on Multimodal User Interfaces, 2010, 3(4): 271-284

[22] Kipp M, Heloir A, Gebhard P. Schroeder, Realizing Multimodal Behavior: Closing the gap between behavior planning and embodied agent presentation[C]//Proceedings of the 10th International Conference on Intelligent Virtual Agents. Springer, 2010

[23] Tao Jian-hua, Mu Kai-hui, Che Jian-feng, et al. Audio-Visual Based Emotion Recognition with the Balance of Dominances[C]// International Conference on Artificial Intelligence (ICAI1010). Oct. 2010: 100-110

[24] EMMA: Extensible MultiModal Annotation markup language [OL]. [2013-08-21]. http://www.w3.org/TR/emma/

[25] Speech Synthesis Markup Language(SSML) Verison 1.1[OL]. [2013-08-21]. http://www.w3.org/TR/speech-synthesis11/

[26] Tao Jian-hua, Xin Le, Yin Pan-rong. Realistic Visual Speech Synthesis based on Hybrid Concatenation Method [J]. IEEE Transactions on Audio, Speech and Language Processing, 2009, 17(3): 469-477

[27] 3D CHARACTER ANIMATION LIBRARY [OL]. [2013-8-21]. http://home.gna.org/cal3d/