文章编号:1672-3961(2012)05-0087-04 DOI:10.6040/j.issn.1672-3961.2012.05.015

基于词贡献度的垃圾短信分类方法

张永军[1],刘金岭[2],于长辉[3]
(淮阴工学院计算机工程学院,江苏 淮安 223003)

摘要:针对垃圾短信分类问题,提出了一种以词贡献度为基础的分类方法。该方法引入词贡献度的概念表达词在不同短信分类中的权重差别,通过构建词贡献度——分类矩阵和计算矩阵行均方差来实现降维,以词贡献度为基础计算短信隶属于短信分类的隶属度,并通过比较隶属度密度的方法解决分类冲突问题。实验结果表明,该方法在分类效果和实时性方面优于其他常用垃圾短信分类方法。

关键词:垃圾短信;文本分类;词贡献度;方差;降维
中图分类号:TP311 **文献标志码**:A

A spam short message classification method based on word contribution

ZHANG Yong-jun[1], LIU Jin-ling[2], YU Chang-hui[3]
(Faculty of Computer Engineering, Huaiyin Institute of Technology, Huai'an 223003, China)

Abstract: A classification method based on word contribution was proposed to classify spam short messages. The concept of word contribution was introduced for representing weight difference of a word in different categories, the word contribution-classification matrix was constructed, then the mean square deviation of each row in the matrix was computed to reduce dimensionalities. To determine the classification a short message belongs to, short message-category membership degrees were calculated based on word contribution. Furthermore if category candidates were more than one, the classification conflict problem could be resolved by comparing the densities of short message-category membership degree. The experimental results showed that the proposed method was superior to other classification methods in the classification result and real-time.

Key words: spam short message; text classification; word contribution; variance; dimensionality reduction

0 引言

短信已经成为一种重要的交流手段和信息传播的载体,伴随着短信的广泛使用,通过短信传播包含广告、欺骗、色情、诅咒等垃圾信息内容的现象也越来越严重。因此,建立一种高效和准确的垃圾短信分类方法显得极为重要。国内外对文本分类工作进行了众多的研究,目前主流的是以文本向量空间为模型的分类方法[1-2]。在该方法中,将文本表示为由词权重组成的向量,通过计算文本和样本空间中样本文本距离来计算文本所属类别。词权重通常用词频 TFIDF[3-4]来计算。采用以上方法来对短信进行分类,存在着以下问题:(1) 短信文本表示为词向量时,向量只有极少数维值不为0,特征缺失问题较为严重;(2) 短信中词重复现象几乎不存在,因此采用 TFIDF 计算词权重时,TF 的值通常为定值1;(3) 在短信样本空间较大的情况下,计算待分类短信向量和样本空间中的短信文本向量距离较为耗时,不能满足实时性的要求。国内外关于垃圾短信分类的研究较少,刘金岭利用查询词扩展来解决垃圾短信检索问题,该方法考虑了某些词在标示垃圾短信时

收稿日期:2012-05-20
基金项目:江苏省教育厅高校哲学社会研究资助项目(2012SJD87001)
作者简介:张永军(1978 -),男,江苏扬州人,讲师,硕士,主要研究方向为文本数据挖掘. E-mail:13511543380@139.com

的显著作用，以这些词为中心来实现垃圾短信检索问题[5]。刘金岭还提出了基于上下文的短信文本分类方法，该方法通过词共现来计算短信距离[6]。龚才春在考虑了词性、词字长、词交集的基础上提出了一种计算短信相似度的方法[7]。该方法主要用于计算短信的相似度，未考虑同一词性的词在计算分类过程中所起的作用不同。国外对垃圾短信过滤和分类方法的研究主要集中在以贝叶斯为核心的分类算法[8-17]。一些学者还对采用支持向量机方法解决垃圾短信分类问题进行了研究[10,18-19]。

本研究考虑不同特征词在分类中的差别，采用一种全新的方法来计算垃圾短信的分类，该方法的主要优点在于：(1) 考虑了同一个词对不同垃圾短信分类的所起的作用不一样；(2) 不需要和样本空间的样本短信进行比较计算，降低了计算的复杂度，提高了实时性；(3) 一定程度上解决了短信文本的特征缺失问题，算法的分类效果较好。

1 词贡献度及分类方法

1.1 基本概念

定义垃圾短信样本空间 $S=\{\mathrm{sm}_1,\mathrm{sm}_2,\cdots,\mathrm{sm}_N\}$，其中 sm_i 是样本空间中的一条短信文本，样本空间包含的短信文本数量为 N。设 S 中的样本短信被分为了 K 类，定义 $S_j=\{\mathrm{sm}_i\mid \mathrm{sm}_i$ 为分类 j 中一个短信文本$\}$。显然有 $S_1,S_2,\cdots,S_K$ 是 S 的一个划分。定义 $|S_j|$ 为 N_j，即分类 j 中包含了 N_j 条样本短信。

系统首先对样本空间中的短信进行分词处理，在分词过程中已经通过停用词表将虚词和常用词进行了过滤。分词后，系统还进行了以下的预处理，包括：(1) 同义词替换，将同义词用一个语义代表词来替换；(2) 词替换，例如将“2010-10-8”替换为词“日期”，“13253456687”替换为“电话”。

定义 W 为对 S 经过分词和预处理后剩余的词集合。令

$$f(w,\mathrm{sm})=\begin{cases}1, & \text{如果短信 sm 包含词 } w,\\ 0, & \text{如果短信 sm 不包含词 } w,\end{cases} \tag{1}$$

$$V_i=\sum_{\mathrm{sm}\in S} f(w_i,\mathrm{sm}), \tag{2}$$

$$V_{ij}=\sum_{\mathrm{sm}\in S_j} f(w_i,\mathrm{sm}), \tag{3}$$

其中 V_i 表示在样本空间中包含了词 w_i 的短信数目。V_{ij} 表示在分类 j 中包含词 w_i 的短信数。

定义词 w_i 对分类 j 的贡献度，如果词 w_i 在某个短信 sm 中出现，短信 sm 归类于分类 j 的可能性。用记号 p_{ij} 表示词 w_i 对分类 j 的贡献度，取

$$p_{ij}=\frac{V_{ij}}{V_i}, \tag{4}$$

其中 V_i 和 V_{ij} 的计算和含义如式(2)、(3)所示。在笔者的实验样本空间中，词“降价”和“礼包”在垃圾短信分类“广告促销”中的贡献度分别为 0.43 和 0.57。

以词贡献度为基础，构建矩阵 $(\boldsymbol{P}_{ij})$，矩阵的每一行为词 w_i 对不同分类的贡献度，每一列为某个分类中不同词的贡献度，见表1。

表1 词贡献度－分类矩阵

Table 1 Word Contribution-classification matrix

样本空间中的词	垃圾短信分类			
	1	2	…	K
W_1	p_{11}	p_{12}	…	p_{1k}
…	…	…	…	…
W_M	p_{M1}	p_{M2}	…	p_{Mk}

观察该矩阵，可以发现：从词的角度考虑(行)，如果词在不同分类中的贡献度差异大，则说明词的分类能力较强，该差异可以通过方差来体现。从分类的角度考虑(列)，贡献度大的词对该分类的分类效果好，因此这些词应该作为分类的特征词。

定义分类 j 的特征词集为 $\{w_i\mid p_{ij}>\theta_j\}$，其中 θ_j 为分类 j 的特征词贡献度阈值。该定义表明了只有词贡献度大于阈值 θ_j 的词才是分类 j 的特征词。记 W_j 为分类 j 的特征词集合。定义函数

$$f(i,j)=\begin{cases}1, & w_i\in W_j;\\ 0, & w_i\notin W_j。\end{cases} \tag{5}$$

为分类 j 的特征词函数。

1.2 降维

词集合 W 包含了较多的元素，需要进一步降维，设降维后的词集合为 W^*，W^* 的元素应具有以下特征：

(1) $\forall w_i\in W^*$，有 $VA_i=1/k\sum_{j=1}^{k}(p_{ij}-\overline{p_i})^2>\beta$，

其中 $$\overline{p_i}=1/k\sum_{j=1}^{k}p_{ij}。 \tag{6}$$

(2) $\exists$ 样本 j 使得 $f(i,j)=1$。

上述条件可以描述为词 w_i 在不同分类中的贡献度方差大于阈值 β，且是某个分类的特征词才可能在降维后被保留。

1.3 分类

1.3.1 短信隶属度计算及候选分类的确定

短信隶属度是一个用于表明短信属于某个分类的程度的数值。设短信 $\mathrm{sm}(\mathrm{sm}\in S_j)$ 进行分词后得到的词集合为 W'，记 $W''=W'\cap W^*\cap W_j$。则短信 sm 对分类 j 的隶属度为

$$c(\mathrm{sm}, j) = \sum_{w_i \in W''} p_{ij}。\quad (7)$$

对于样本空间中的某个分类 j,定义

$$\min c_j = \min_{\mathrm{sm} \in S_j}(c(\mathrm{sm}, j))。\quad (8)$$

$$\max c_j = \max_{\mathrm{sm} \in S_j}(c(\mathrm{sm}, j))。\quad (9)$$

若短信 sm 的隶属度 $\min c_j <= c(\mathrm{sm}, j) <= \max c_j$,则认为分类 j 是短信 sm 的候选分类。如果 sm 不存在候选分类,则判定 sm 不属于垃圾短信。实验表明,多数垃圾短信的候选分类唯一。当短信的候选分类有多个时,可以通过计算隶属度密度的方法来确定短信最终分类。

1.3.2　最终短信分类的确定

*K*NN 是一种常用的文本分类算法,其核心思想是给定新文本后,考虑在训练文本集中与该新文本距离最近(最相似)的 k 篇文本,根据这 k 篇文本所属的类别判定新文本所属的类别[20]。借鉴该算法思想本研究提出了一种根据隶属度密度来确定短信最终分类的算法。该算法的中心思想是给定短信 sm,如果某分类 j 中,在 sm 的周围具有最为密集的样本短信,则认为短信 sm 的最终分类为 j。

给定分类隶属度区间$[l,u]$,定义

$$\mathrm{Locate}(\mathrm{sm}, j)_l^u = \begin{cases} 1, & l <= c(\mathrm{sm}, j) <= u, \\ 0, & c(\mathrm{sm}, j) > u \text{ 或 } c(\mathrm{sm}, j) < l。 \end{cases} \quad (10)$$

式(10)可以描述为如果短线 sm 在分类 j 的隶属度在区间$[l,u]$中,用值 1 表示;否则用 0 表示。

定义 sm 短信在 j 分类中的密度定义为

$$\mathrm{Density}(\mathrm{sm}, j) = 1/N_j \sum_{\mathrm{sm}^* \in S_j} \mathrm{Locate}(\mathrm{sm}^*, j)_{c(\mathrm{sm}, j) - \Delta_j}^{c(\mathrm{sm}, j)}, \quad (11)$$

其中 $\Delta_j = (\max c_j - \min c_j)/l$,实验中通常取 $l = 10$ 或者 $l = 20$。N_j 是分类 j 的样本短信数目。

计算短信在每一个候选分类 j 的隶属度密度,最大密度所对应的分类被认为是短信 sm 的最终分类。

2　实验结果及分析

实验算法采用 Visual 2010 和 SQL 2008 实现,实验环境如下:内存为 4 G,CPU 为 Intel Core 2 Duo P7350 45 nm 处理器,操作系统为 Windows 7。人工收集了 5 328 条垃圾短信来构造了样本空间,对 420 条短信进行了分类测试。实验中共采用了 5 种方法,将文献[6]中基于上下文的方法记为 Context,将文献[2]中的方法记为 TF,文献[8]中的贝叶斯方法记为 BA,支持向量机方法记为 SVM,本研究所描述的方法记为 Contribution(β 取值为 0.008, l 取 10)。

2.1　分类效果比较

采用指标分类查准率(P)和查全率(R)来评估分类算法效果。其中,

$$P = \frac{\text{分类准确文本数}}{\text{实际分类文本数}} \times 100\%,$$

$$R = \frac{\text{分类准确文本数}}{\text{应有文本数}} \times 100\%。$$

从表 2 可以看出,除了 TF 以外 4 个分类方法的分类效果较为接近,Contribution 方法效果最优的原因在于通过隶属度密度的计算可以二次确定分类。TF 方法采用改进的 TFIDF 方法来计算词权重,由于短信中 TF 的值几乎都为 1,因此分类效果差。

表 2　分类算法实验结果

Table 2　Experiment results of classification algorithms

分类算法	实验结果	
	R	*P*
Context	0.823	0.796
TF	0.732	0.692
Contribution	0.848	0.818
BA	0.826	0.811
SVM	0.832	0.809

2.2　分类时间比较

为了更好的反映算法的时间复杂度,在实验中记录了 100 条、200 条、300 条和 400 条的耗时,见图 1 所示。

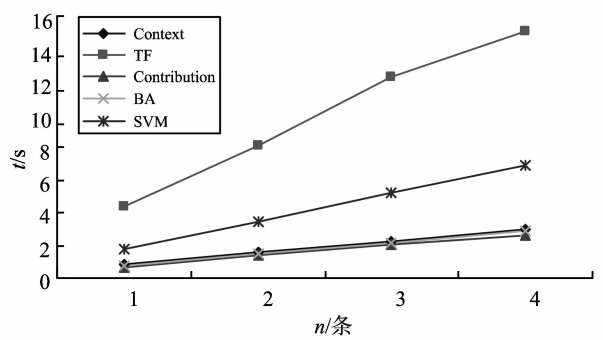

图 1　实验分类耗时

Fig. 1　Classification time cost of experiment

Context、Contribution 和贝叶斯方法性能较为接近,其原因在于分类时主要与每个垃圾短信分类进行比较,因此三者在耗时上非常接近。TF 方法需要与每个样本短信比较,最为耗时。

3　结语

本研究在考虑了词在不同分类中的分类差别,提出了词对分类贡献度的概念。并以此为基础实现了垃圾短信分类算法。实验结果表明:该方法在准

确度和时间复杂度都达到了较好的效果。但是本研究未考虑词性和词字数对词贡献度的影响，另外计算短信隶属度通过累加短信中关键词的分类贡献度来实现，未考虑词群对分类的影响，综合考虑词的属性和词群贡献度来实现短信分类是下一步研究方向。

参考文献：

[1] SALTON G, WANG A, YANG C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(5):613-620.

[2] ESIN Y E, ALAN O, ALPASLAN F N. Improvement on corpus-based word similarity using vector space models [C]// 24th International Symposium on Computer and Information Sciences. Guzelyurt: Middle East Technical University Press, 2009: 280-285.

[3] LEWIS D. Feature selection and feature extraction for text categorization[C]// Proceedings of Speech and Natural Language Workshop. San Mateo: Morgan Kaulinann Press, 1992: 212-217.

[4] 张玉芳，彭时名，吕佳. 基于文本分类TFIDF方法的改进与应用[J]. 计算机工程，2006,32(19):76-78.
ZHANG Yufang, PEN Shiming, LÜ Jia. The improvement and application of text classification method based on TFIDF[J]. Computer Engineering, 2006, 32(19):76-78.

[5] 刘金岭. 基于查询词扩展的中文垃圾短信检索[J]. 计算机工程,2011,37(8):52-54.
LIU Jinling. The Chinese spam short message retrieval based on query words extension[J]. Computer Engineering, 2011, 37(8):52-54.

[6] 刘金岭. 基于上下文的短信文本分类方法[J]. 计算机工程，2011,37(10):41-43.
LIU Jinling. The SMS text classification method based on context[J]. Computer Engineering, 2011, 37(10):41-43.

[7] 龚才春. 短文本语言计算的关键技术研究[D]. 北京：中国科学院计算技术研究所，2008.
GONG Caichun. Research on short text language computing[D]. Beijing: Institute of Computing Technology, Chinese Academy of Science, 2008.

[8] BELEM D. Content filtering for SMS systems based on Bayesian classifier and word grouping[C]// Network Operations and Management Symposium (LANOMS), 2011 7th Latin American. Quito: IEEE Press, 2011: 1-7.

[9] UYSAL A. Detection of SMS spam messages on mobile phones[C]// Signal Processing and Communications Applications Conference (SIU), 2012 20th. Mugla: IEEE Press, 2012: 1-4.

[10] KHEMAPATAPAN C. Thai-English spam SMS filtering [C]// Communications (APCC), 2010 16th Asia-Pacific. Auckland: IEEE Press, 2010: 226-230.

[11] CAI Jie, TANG Yuezhong, HU Rile. Spam filter for short messages using winnow[C]// International Conference on Advanced Language Processing and Web Information Technology. Dalian:IEEE Press, 2008: 454-459.

[12] 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:自然科学版，2010,24(1):77-81.
ZHANG Jing, HOU Xudong, LÜ Hesheng. Development of an intelligent SMS analysis system based on naive Bayes and support vector machine[J]. Journal of Chongqing University of Technology: Natural Science, 2010, 24(1):77-81.

[13] VAHORA S, HASAN M, LAKHANI R. Novel approach: Naïve Bayes with vector space model for spam classification[C]// Engineering (NUiCONE), 2011 Nirma University International Conference. Ahmedabad Gujarat:Nirma University Press, 2011: 1-5.

[14] AMAKAMI A, ALMEIDA J. Evaluation of approaches for dimensionality reduction applied with naive Bayes anti-spam filters[C]// Machine Learning and Applications, ICMLA′09. International Conference [S. l.]. IEEE Press, 2009: 517-522.

[15] GUNAL S, ERGIN S, GUNAL E S. Detection of SMS spam messages on mobile phones [C]// 2012 20thSignal Processing and Communications Applications Conference (SIU). Mugla: IEEE Press, 2012: 1-4.

[16] KHALED S M, FARHAN K, ABDUR Rahman M. Modeling spammer behavior: Naïve Bayes vs. artificial neural networks [C]// Information and Multimedia Technology, ICIMT′09 International Conference. Jeju Island: IEEE Press, 2009: 52-55.

[17] HAN Kyoungsoo, RRIM Haechang, SUNG Hyon Myaeng. Some effective techniques for Naive Bayes text classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11):1457-1466.

[18] YU Qiang, JIAN Wei. A Chinese anti-spam filter approach based on support vector machine[C]// Management Science and Engineering, ICMSE 2007 International Conference. Harbin:Harbin Institute of Technology Press, 2007: 97-102.

[19] TAN Chewlim, SU Jian, LU Yue. Supervised and traditional term weighting methods for automatic text categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4):721-735.

[20] WANG Hui. Nearest neighbors by neighborhood counting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(6):942-953.

（编辑：陈燕）