基于语义的文本特征加权分类算法*

张国栋[1,2]，张化祥[1,2]

(1. 山东师范大学 信息科学与工程学院，济南 250014；2. 山东省分布式计算机软件新技术重点实验室，济南 250014)

摘 要：文本分类存在维数灾难、数据集噪声及特征词对分类贡献不同等问题，影响文本分类精度。为提高文本分类精度，在数据处理方面提出一种新方法。该方法首先对数据集进行去噪处理，结合特征提取算法和语义分析方法对数据实现降维，再利用词语语义相关度对文本特征向量中每个特征词赋予不同权重；并利用经过上述处理的文本数据学习分类器。实验结果表明，该文本处理方法能够有效提高文本分类精度。

关键词：语义分析；降维；语义相关度；分类

中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2012)12-4476-03

doi:10.3969/j.issn.1001-3695.2012.12.018

Classification algorithm based on semantics and text feature weighting

ZHANG Guo-dong[1,2], ZHANG Hua-xiang[1,2]

(1. *School of Information Science & Engineering, Shandong Normal University, Jinan* 250014, *China*; 2. *Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan* 250014, *China*)

Abstract: Text categorization faces the problems of dimensionality curse, noise data and different classification contributions for different feature words. In order to improve text classification accuracy, this paper presented a new approach to data processing. The approach first removed the noise data, and then employed feature extraction algorithms and semantic analysis methods to implement dimensionality reduction. Different weights were assigned to different text features based on a semantic similarity evaluation. The processed data were used to construct classifiers. Experimental results show that the text processing method can effectively improve the accuracy of text classification.

Key words: semantic analysis; dimensionality reduction; semantic correlation; classification

目前互联网资源呈现海量增长趋势，应用信息检索、数据挖掘技术有效组织和利用海量信息逐渐成为备受关注的问题。其中，文本分类在自然语言处理、信息组织和管理等方面都有着广泛应用[1]。文本分类技术是信息检索和文本挖掘的重要基础，其主要任务是通过训练分类器，利用该分类器对未标记样本进行分类。现有用于文本分类的方法 SVM[2]、最大熵模型[2]、模糊理论[3]以及基于等效半径的方法[4]等都是基于统计的机器学习方法，其面临的主要困难在于维度灾难。为实现降维，常用的方法有特征选择和特征抽取。特征选择主要方法有 DF、信息增益、互信息 χ^2 统计量[5]等，另一种降维方式是利用向量聚合技术[6]等特征抽取方法，从原始特征集产生新的、更少的合成特征集合。这些方法实现简单，但大多基于统计理论。以上文本分类方法并没有考虑文本内部字、词、句等间的内在语义关系。事实上，一个类别的文本特征词间有很大的关系，即词与词之间存在相关关系、同义关系等。因此，在本文分类中一种可行的方法是对文本中的词进行部分语义分析，从而弥补统计方法缺乏语义信息的弊端。基于这一点，本文首先对数据集进行降噪处理，然后在数据处理上利用同义词词林(http://ir.hit.edu.cn/phpwebsite/index.php)的词频合并方法对特征词进行语义上的分析，并进行词频合并实现降维。同时本文结合知网(http://www.keenage.com)的词语语义相似度计算方法对测试文本权重向量再处理，从而提高文本分类精度。

1 文本数据处理技术

1.1 特征词权重改进计算

向量空间模型(vector space model, VSM)[7]使用向量表示文本，文本间差异性表示为向量之间的相似度。如文档 T：$(W_1, W_2, \cdots, W_n)$，其中 W_i 代表第 i 个特征项的权重，n 表示文档中特征项个数，一般选取文档中的词作为特征项。当文档被表示为空间中的向量时，就可以通过计算向量间的相似性衡量文档间相似度。常用的相似性度量方式是余弦距离。最初向量表示模型使用布尔函数[8]，即

$$\begin{cases} 1 & W \text{出现} \\ 0 & W \text{不出现，其中 } W \text{代表特征词} \end{cases} \tag{1}$$

但这种表示方法无法表示该词在文档中的作用，于是逐渐被更准确的词频代替。目前普遍使用 TFIDF 公式计算词频[9]，即

$$w(t,d) = \frac{tf(t,d) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t=1}^{n} [tf(t,d) \times \log(N/n_t + 0.01)]^2}} \tag{2}$$

收稿日期：2012-04-29；**修回日期**：2012-05-30 **基金项目**：国家自然科学基金资助项目(61170145)；国家教育部高等学校博士点专项基金资助项目(20113704110001)；山东省自然科学基金资助项目和科技攻关计划项目(ZR2010FM021,2008B0026,2010G0020115)

作者简介：张国栋(1988-)，男，山东泰安人，硕士研究生，主要研究方向为机器学习、数据挖掘(guodongZ_1988@163.com)；张化祥(1966-)，男(通信作者)，教授，博导，主要研究方向为机器学习、模式识别及 Web 挖掘等.

其中:$w(t,d)$表示词 t 在文本 d 中的权重;$tf(t,d)$表示词 t 在文本 d 中的词频;N 表示所有的训练文本个数;n_t 表示训练文本中含有词 t 的文本数。

上述方法在计算词语权重时只考虑了词的出现频率,忽略其他因素对词语权重的影响。为此本文对 $w(t,d)$的计算式进行改进,把与特征词相关的一些参数加入词语权重表达式中。经过研究发现,特征词的权重还和该词出现的位置、词的长度等有关,所以权值更改为

$$W = w(t,d) \times \alpha \times \sum_{i=1}^{k} \beta \tag{3}$$

其中:W 为该特征词的权重;α 为特征词基于词长的长度参数;β 为词语在第 i 种位置上出现次数的位置参数;$k=4$ 代表特征词出现的四个位置,包括标题、首段、正文、末段。这四个参数的获得方法是:首先通过大量文本进行测试,获得初始化参数,然后在实验中对这些参数不断调整测试,获得一组最优参数序列。

1.2　词频合并分析

作者为让一篇文章丰富且具有阅读性,常常对一个概念使用不同词语表达。同一个概念不同表达方法,不仅分散了概念,而且在提取特征词、统计词频时容易让人产生误导,往往导致对词语相近的文档划分到不同的类别,降低分类的准确性[10]。传统降维方法如互信息、信息增益、PCA 等没有考虑词语之间的语义关系,只是简单通过计算剔除一些向量,这样很有可能去掉一些出现次数低但对分类有用的词语。本文在传统降维基础上再利用同义词词林合并相近意义的特征词,在一定程度上强化特征词的权重,弱化非特征词的权重。

本文利用同义词词林(扩展版)[11]对提取的特征词进行语义分析。该词林收录词语共有 77 458 条,其保留同义词词林原有的三层分类体系,并在此基础上对词语继续细分类,增加两层,得到最终的五层分类体系。它把词汇分成大、中、小三类,大类有 12 个,中类有 97 个,小类有 1 428 个,小类下再划分词群。每个词群中的词语又进一步分成了若干个行,同一行的词语词义相同或有很强的相关性。通过对特征词中存在同义、相关、上位、下位等相关信息的词进行词频合并,就可以把对特征词的表示提升到概念层次。利用词林处理后,文本特征向量的维数得到进一步降低。

1.3　词语语义相关度计算

传统文本分类对测试文本只进行简单预处理,然后利用分类算法进行分类,这样处理对分类结果会产生误差。本文在此基础上利用知网中的词语语义相似度对特征词间的相似度进行计算,对测试文本的特征向量权重重新赋值,从而使得两文本间相似度的计算更加准确[12]。

知网是以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念间以及概念所具有的属性间关系为基本内容的常识知识库。它有两个主要的概念,即概念与义原。一个词语有 m 个概念,一个概念有 n 个义原。例如:教育 N affairs|事务,education|教育 V teach|教。这里“教育”就有两个概念,并且词性还不相同,一个为名词 N,一个为动词 V。其名词概念有两个义原。词语间的相似度为概念对相似度的最大值。例如:对于词语 W_1 和 W_2,如果 W_1 有 m 个义原 $w_{11},w_{12},\cdots,w_{1m}$,$W_2$ 有 n 个义原 $w_{21},w_{22},\cdots,w_{2n}$,那么 W_1 和 W_2 相似度是各个概念的相似度的最大值:

$$\mathrm{sim}(W_1,W_2) = \max_{i=1,\cdots,m,j=1,\cdots,m} \mathrm{sim}(w_{1m},w_{2j}) \tag{4}$$

其中义原间的相似度为

$$\mathrm{sim}(w_{1m},w_{2j}) = \frac{\lambda}{\lambda + \mathrm{dis}(w_{1m},w_{2j})} \tag{5}$$

$\mathrm{dis}(w_{1m},w_{2j})$是两个义原在树上的距离。如果两个义原不在同一棵树上,则相似度记为 0,实验中 $\lambda=1.6$。

测试文本预处理后,利用知网的词语语义相似度计算方法对测试文本特征词权重进行重新赋值。例如,训练文本和测试文本的特征词权重如表 1 所示。

表 1　特征词权重分布

训练文本	中华	人民	建造	结构	团结
	0.70	0.60	0.40	0.40	0.30
测试文本	中国	框架	合作	大众	建筑
	0.60	0.50	0.70	0.30	0.40

通过表 1 得到 10 个特征词:中华、人民、建筑、结构、团结、中国、框架、合作、大众、建筑,分别记为 $W_1,W_2,\cdots,W_{10}$。设定一个阈值,对于两词语间相似度大于该阈值进行特征词权重重新赋值,通过实验取得阈值为 0.80 是最佳值。

通过计算两词语间的相似度大于阈值分别有:

$\mathrm{sim}(W_1,W_6)=1.0$,$\mathrm{sim}(W_2,W_9)=0.9$,$\mathrm{sim}(W_3,W_{10})=1.0$

对词语间相似度大于阈值的特征词,测试文本中的特征词用训练文本中的相应词代替,权重重新取值为 $w \times \mathrm{sim}(w_i,w_j)$,这样对测试文本特征词重新设置权重,如表 2 所示。

表 2　重新设置后的特征词权重

类别	特征词						
	中华	人民	建造	结构	团结	框架	合作
训练文本	0.70	0.60	0.40	0.40	0.30	0.00	0.00
测试文本	0.60×1.0	0.30×0.9	0.40×1.0	0.00	0.00	0.50	0.70

文本经过处理不仅降低了特征向量维数,对于计算两文本之间的相似度更加精确。

2　数据降噪及算法描述

2.1　数据降噪

本文实验采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组的语料库(http://www.nlp.org.cn)。该语料库存在如下问题:a)语料库包含训练集(train)和测试集(answer)两部分,分别包含 9 000 多个文档,但分别有近 1 500 个文档重复;b)train 和 answer 中 C35-Law 中的部分文件已经经过分词处理,但分词结果很差;c)有些文章只有文章头部,没有实际内容,有的类别存在空文档;d)一些短小的文档由于分词后的关键词太少,对分类基本不起作用。

为了提高分类精度,本文对该语料库进行了降噪处理:删除 train 和 answer 文件下的重复文件;删除 train 和 answer 文件下的 C35-Law 文件夹;删除所有长度小于 500 的文档;对文本重新编号,便于分类的实现。

2.2　算法描述

1)预处理　对语料库中的文本进行分词,去除停用词,计算词语权重。

2)特征提取　采用信息增益[13]结合同义词词林中的语义分析方法进行特征词的提取,降低特征向量维数。

3)利用知网中的词语语义相似度计算特征词间的相似度,对测试文本中的特征词权重进行处理。

4)利用分类算法进行文本分类。

3 实验设计与分析

实验数据采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组的语料库。其中,train 为训练文本,answer 为测试文本,语料库包含 20 类。降噪处理前后文档类别分布如表 3 所示。

表 3 降噪处理前后文档类别分布

category	train		answer	
	处理前	处理后	处理前	处理后
art	740	513	742	510
literature	33	0	34	0
education	59	0	61	0
philosophy	44	0	45	0
history	466	440	468	450
space	640	470	642	464
energy	32	0	33	0
electronics	27	0	28	0
communication	25	0	27	0
computer	1 357	982	1 358	984
mine	33	0	34	0
transport	57	0	59	0
environment	1217	756	1 218	740
agriculture	1 021	816	1 022	822
economy	1 600	1 392	1 601	1404
law	51	0	52	0
medical	51	0	53	0
military	74	0	76	0
politics	1 024	750	1 026	754
sports	1 253	1 065	1 254	1 060

实验中用以上九类作为实验对象,文本分类效果的测评标准采用 precision、recall、F-score 三个值。分别用未进行数据处理的学习支持向量机算法(SVM)、朴素贝叶斯分类算法(NB)、神经网络分类算法(NN)、K-近邻算法(KNN)(取 $k=15$)与用处理后的数据学习这四种算法,在该语料库上进行实验,结果如表 4(表中分类效果较好的数据用粗体标出)所示。

表 4 各个分类算法分类效果对照

分类算法		precision	recall	F-score
SVM	处理前	0.901	0.823	0.860
	处理后	0.907	0.842	0.873
NB	处理前	0.793	0.764	0.778
	处理后	0.782	0.793	0.787
NN	处理前	0.798	0.753	0.775
	处理后	0.832	0.742	0.784
KNN	处理前	0.889	0.827	0.856
	处理后	0.903	0.854	0.877

从表 4 可以看出,相对于传统未进行数据处理的算法,对文本数据进行处理后,算法在 precision、recall、F-score 三个指标上都有所提高。

未进行数据处理的 KNN 算法与进行数据处理后的 KNN 算法(取 $k=15$)结果比较如表 5 所示。

表 5 未进行数据处理的 KNN 算法与进行数据处理后的 KNN 算法(取 $k=15$)结果比较

类别	precision		recall		F-score	
	处理前	处理后	处理前	处理后	处理前	处理后
art	0.939	0.949	0.919	0.922	0.929	0.935
history	0.781	0.761	0.682	0.667	0.728	0.711
space	0.807	0.858	0.722	0.806	0.762	0.831
computer	0.921	0.938	0.910	0.927	0.915	0.933
environment	0.932	0.908	0.855	0.830	0.892	0.867
agriculture	0.838	0.853	0.827	0.849	0.833	0.851
economy	0.917	0.955	0.756	0.856	0.829	0.903
politics	0.875	0.910	0.851	0.890	0.863	0.900
sports	0.987	0.991	0.916	0.942	0.950	0.966
average	0.889	0.903	0.827	0.854	0.856	0.877

从表 5 可看出,在九个实验类别中,history 和 environment 效果不如处理前 KNN 算法分类效果(表中分类效果较好的数据用粗体标出)。由于 history 类别文本数量相对较少,而本文算法对于有大量文本的类别分类效果更好,因为文本越多,权重计算更加准确;而对于 environment 类别,由于其主题包括各个方面,导致提取的特征词不明显,所以分类效果不是很好。但在其他七个类别中,本文算法在 precision、recall、F-score 值上都优于未进行数据处理的 KNN 算法,average-precision、average-recall、average-F-score 也都高于未进行数据处理的 KNN 算法。

4 结束语

本文针对文本分类中的维数灾难问题,提出了利用信息增益结合同义词词林中的语义分析方法实现降维,利用知网中的语义相似度计算特征词间的相似度,对测试文本特征向量各维赋予不同权重。实验表明本文提出的方法能够有效地提高文本分类精度,但本文在处理文本时忽略了文本分类的速度,所以,在提高分类精度的基础上如何提高效率是下一步要研究的问题。

参考文献:

[1] ZAKARIA E, ABDELATTIF R, MOHAMED A. Using WordNet for text categorization[J]. The International Arab Journal of Information Technology, 2008, 5(1): 16-24.

[2] KAZAMA J, TSUJII J. Maximum entropy models with inequality constraints: a case study on text categorization [J]. Machine Learning, 2005, 60(1-3): 159-194.

[3] WEI C P, LIN Yen-ting, YANG C C. Cross-lingual text categorization: conquering language boundaries in globalized environments [J]. Information Processing & Management, 2011, 47(5): 786-804.

[4] WANG Jian-hui, WANG Hong-wei, ZHAN Shen, *et al*. A simple and efficient algorithm to classify a large scale of text [J]. Computer Research and Development, 2005, 42(1): 85-93.

[5] SUJEEVAN A, YOUNES B. Semi-structured document categorization with a semantic kernel [J]. Pattern Recognition, 2009, 42(9): 2067-2076.

[6] FERNANDO F, KSENIYA Z, WOLF-GANG M. Text categorization methods for automatic estimation of verbal intelligence [J]. Expert Systems with Applications, 2012, 39(10): 9807-9820.

[7] MANNE S, KOTHA S K, FATIMA S S. Text categorization with K-nearest neighbor approach [C]//Proc of InConINDIA. Berlin: Springer Verlag, 2012: 413-420.

[8] GUO Yi, SHAO Zhi-qing, HUA Nan. Automatic text categorization based on content analysis with cognitive situation models [J]. Information Sciences, 2010, 180(5): 613-630.

[9] YAHIA M E. Arabic text categorization based on rough set classification [C]//Proc of Computer Systems and Applications. [S. l.]: IEEE Press, 2011: 293-294.

[10] WANG Zi-qiang, XU Qian. Text categorization based on LDA and SVM [C]//Proc of Computer Science and Software Engineering. Washington DC: IEEE Computer Society, 2008: 674-677.

[11] LI Yan-jun, HSU D F, CHUNG S N. Combining multiple feature selection methods for text categorization by using rank-score characteristics [C]//Proc of the 21st IEEE International Conference on Tools with Artificial Intelligence. Washington DC: IEEE Computer Society, 2009: 508-517.

[12] CAI Deng, HE Xiao-fei, HAN Jia-wei. Document clustering using locality preserving indexing [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(12): 1624-1637.

[13] ZHANG Yun-liang, ZHU Li-jun, QIAO Xiao-dong, *et al*. Flexible K-NN algorithm for text categorization by authorship based on features of lingual conceptual expression[C]//Proc of WRI World Congress on Computer Science and Information Engineering. Washington DC: IEEE Computer Society, 2009: 601-605.