ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 516－522
DOI:10. 16337/j. 1004-9037. 2017. 03. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 一种面向文本分类的特征迁移方法

赵世琛　王文剑

(山西大学计算机与信息技术学院,太原,030006)

**摘　要**:传统的文本分类方法假设训练集与测试集中的特征词服从相同的概率分布,但在实际应用中,以上假设存在偏差,会影响到最终的分类结果。针对这一情况,本文采用迁移学习,通过计算特征词的迁移量对训练集中向量空间模型进行修正,最终使训练集与测试集中特征词的分布概率趋于一致。将提出的方法应用于中文垃圾邮件过滤与中、英文网页分类中,在CHI统计特征选择基础上进行特征迁移,实验结果表明新方法可以有效消除特征词分布的差异性,使文本分类的各项指标明显提高。
**关键词**:文本分类;迁移学习;迁移量;向量空间模型
**中图分类号**:TP18　　**文献标志码**:A

## Feature Transfer Learning for Text Categorization

Zhao Shichen, Wang Wenjian

(School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China)

**Abstract**: Traditional text classification methods assume that feature words in the training set and test set follow the same probability distribution. Nevertheless, deviations exist in a practical application, which can affect the final classification results. To solve the problem, a feature transfer learning algorithm for text categorization is proposed. By calculating the transfer volume and amending the vector space model in the training set, the distribution probability of feature words can be reconciled for the training set and test set. Experiments on Chinese spam filtering and web page classification data sets demonstrate that the proposed method can eliminate the dissimilarity of distributions of feature words, and improve the various indexes of test classification evidently.
**Key words**: text categorization; transfer learning; transfer volume; vector space model

## 引　　言

传统文本分类问题存在一个前提假设,即训练集与测试集服从相同的特征空间和相同的特征分布概率。然而在解决实际问题中,例如网页分类和邮件过滤,由于部分新词的使用或部分特征词词义的迁移,需要根据新增的网页或邮件的特征分布情况不断重新构建分类模型,使得以上假设不能满足实际应用情况,并且会导致文本分类精度的降低。在文本分类问题中,不同时期文本的特征空间不同。例如在中文网页分类问题中,互联网络的普及与发展为社会提供了更为宽广的言论和观点的交

**基金项目**:国家自然科学基金(60975035,61273291)资助项目;山西省回国留学人员科研基金(2012008)资助项目。
**收稿日期**:2015-05-15;**修订日期**:2015-07-07

流平台,随之不断涌现出的网络热词、新词成为汉语言的重要组成部分。教育部、国家语委发布的《2011 年中国语言生活状况报告》中指出[1],2006～2010 年共新增词语 2 977 条,其中 40%广泛被大众使用,如博客、微博、动车和保障房等。这样 2010 年中文网页特征空间维度将远高于 2006 年,即使经过特征选择,其特征空间仍会存在一定差别,如果以 2006 年的特征空间为基础构造空间向量模型,对 2007～2010 年内的文本进行分类,结果将会存在一定偏差。文本分类问题中特征词的分布概率同样会随着时间而发生变化。例如在邮件过滤问题中,中国互联网协会公布的《反垃圾邮件状况调查报告》中显示[2-3],2005 年垃圾邮件按内容主要分为"网上购物"、"网上赚钱"和"情趣用品"3 类,而到 2013 年,垃圾邮件主要分布在"网站推广""金融保险""教育培训"和"欺诈"等 10 个方面。这表明在 2005 年,涉及金融保险、教育培训的特征词较多地分布在正常邮件中,然而到 2013 年,这部分特征词可能更多地出现在垃圾邮件中。

迁移学习早期被广泛地用于教育学领域中,例如人们学会 C 语言,那么通过对比学习,将会更容易地掌握 Java。在机器学习中,迁移学习可以通过一系列的手段或方法消除新旧环境中学习模型的差异,从而使从旧环境中学习到的模型,可以较好地完成新环境中的学习任务。因此对于解决文本分类中特征词分布存在差异的问题,迁移学习是一种可行的方法。本文提出一种面向文本分类的特征迁移算法,通过计算特征词的迁移量,对训练集中向量空间模型进行修正,以有效消除训练集与测试集中特征词分布的差异性。

## 1 直推式迁移学习

根据数据集的差异,迁移学习可以分为归纳式迁移学习、直推式迁移学习和无监督式迁移学习,3 类迁移学习的区别如表 1 所示[4]。在归纳式迁移学习中,可以利用少部分有标记的测试集数据,对从训练集中构造出的学习模型进行不断的修正,从而提高迁移学习效果,例如戴文渊等提出的 TrAdaBoost 算法[5],文献[6]提出利用辅助数据来提高 SVM 的分类精度。然而在实际文本分类问题中,不能提前得到有标记的测试集数据,因此本文采用直推式迁移学习方法解决文本分类中的问题。直推式迁移学习最早由 Arnold 等提出[7],认为直推式迁移学习同半监督式学习类似,测试集相当于半监督式学习中的辅助训练集,在训练时即可得到所有测试数据,并用以提高决策函数的性能,Arnold 等认为虽然训练集和测试集的特征分布存在差异,但是具有相同的类标签。与 Arnold 等观点相补充的是一种类似于多任务学习的直推式迁移学习[8],其中训练集和测试集中特征词服从相同的特征分布,但特征所属的类别发生了变化,例如在引言中提到的邮件过滤问题,特征所属的类别从 2005 年的 3 类增加到 2013 年的 10 类。本文研究背景同 Arnold 一样,主要目的是消除特征分布间的差异,而假设训练集和测试集数据具有相同的类标签,据此对文本分类问题中直推式迁移学习问题定义如下:给定训练集 $D_S$ 和对应的类标签 $L_S$、测试集 $D_T$ 和对应的类标签 $L_T$,其中 $D_S \neq D_T$、$L_S = L_T$,$f_T(\cdot)$为通过对 $D_S$ 和 $L_S$ 学习得到的预测函数,直推式迁移学习是通过消除 $D_S$ 和 $D_T$ 间的差异提高 $f_T(\cdot)$在 $D_T$ 中的分类精度。

**表 1 迁移学习的分类**

**Tab. 1 Settings of transfer learning**

| 迁移学习 | 标记的训练集 | 标记的测试集 |
|---|---|---|
| 归纳式迁移学习 | 有,无 | 有 |
| 直推式迁移学习 | 有 | 无 |
| 无监督式迁移学习 | 无 | 无 |

## 2 文本分类中的迁移学习

影响文本分类结果主要包括文本的特征选择和文本分类方法。文献[9]对基于统计的特征选择方

法进行了较为深入的研究，得出的 CHI 统计(Chi-square)和信息增益方法效果都较好，本文选择 CHI 统计方法进行特征选择。目前常用的文本分类方法有朴素贝叶斯[10]、支持向量机(Support vector machine, SVM)[11]和决策树等方法，SVM 分类速度快、精度高，可以有效避免“维数灾难”，本文采用 SVM 作为文本分类器。

## 2.1 文本分类中的特征迁移分析

本节通过中文垃圾邮件过滤实验，验证文本分类里存在特征词分布迁移的现象。采用中国教育和科研计算机网紧急响应组公布的 5 600 封电子邮件数据集[12]，实验共分为 4 组，每组训练集由 500 封正常邮件和垃圾邮件组成，测试集由 300 封正常邮件和垃圾邮件组成。特征词典的维度为 600 维。邮件过滤率(Correct rate, CR)作为评价指标，则

$$\mathrm{CR}=\frac{N_{\mathrm{ham}}+N_{\mathrm{spam}}}{N_{\mathrm{mail}}} \tag{1}$$

式中：$N_{\mathrm{ham}}$为正常邮件分类正确的数量；$N_{\mathrm{spam}}$为垃圾邮件分类正确的数量；$N_{\mathrm{mail}}$为测试集总邮件数。实验 1 假设训练集和测试集服从相同的特征分布。为达到实验 1 假设条件，将训练集和测试集作为整体进行特征选择，这样构造出的特征词典可以很好地代表训练集和测试集中特征词的分布情况。实验 2 根据实际垃圾邮件过滤情况。由于在实际应用中，不可能在训练时得到测试集的类标记，因此实验 2 在训练阶段只对训练集进行特征选择，这样构造出的特征词典只能代表训练集中特征词的分布情况。实验 1 和实验 2 的结果对比如图 1 所示，从图 1 中可以看出，实验 1 在 4 组数据集上的表现均明显高于实验 2：其中差距最大的是第 3 组，相差 6.8%，最小的是第 4 组，相差 2.1%。通过对比实验可以得出，正是由于实验 2 中的特征词典无法很好地表示测试集中特征词的分布情况，从而导致了垃圾邮件过滤精度降低。

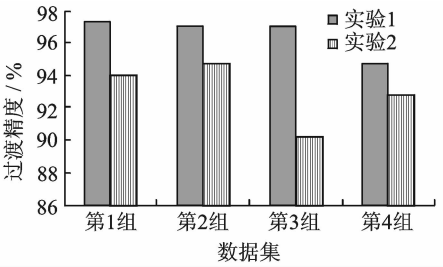



图 1 实验 1 和实验 2 垃圾邮件过滤率的对比

Fig. 1 Contrast between test 1 and test 2

## 2.2 迁移量方法

记文本数据集 $D=\{\boldsymbol{x},y\}$，其中 $\boldsymbol{x}$ 为文本向量；$y=p(\boldsymbol{x})$为文本类别；$D_S$，$D_T$ 分别为训练集和测试集。在直推式迁移学习中，根据经验风险最小化原理，希望得到的测试集中关于参数 $\theta^*$ 的最优化模型为

$$\theta^*=\underset{\theta\in\Theta}{\operatorname{argmin}}\sum_{(x,y)\in D_T}P(D_T)l(x,y,\theta) \tag{2}$$

式中：$P(D_T)$为测试集的边缘分布概率；$l(x)$为关于文本 $x$，类别 $y$ 和参数 $\theta$ 的损失函数。由于得不到测试集中有标记的数据，因此无法直接构造出损失函数。如果 $P(D_S)=P(D_T)$，可以直接利用训练集中的数据得到最优参数 $\theta^*$ 为

$$\theta^*=\underset{\theta\in\Theta}{\operatorname{argmin}}\sum_{(x,y)\in D_S}P(D_S)l(x,y,\theta) \tag{3}$$

但是在迁移学习中，由于 $P(D_S)\neq P(D_T)$，这样就需要对训练集中构造的模型进行修改，从而得到具有高泛化能力的模型为

$$\theta^*=\underset{\theta\in\Theta}{\operatorname{argmin}}\sum_{(x,y)\in D_T}\frac{P(D_T)}{P(D_S)}P(D_S)l(x,y,\theta)\approx\underset{\theta\in\Theta}{\operatorname{argmin}}\sum_{(x,y)\in D_S}\frac{P(D_T)}{P(D_S)}P(D_S)l(x,y,\theta)=$$

$$\underset{\theta\in\Theta}{\operatorname{argmin}}\sum_{i=1}^{n_S}\frac{P_T(x_{T_i},y_{T_i})}{P_S(x_{S_i},y_{S_i})}P_S(x_{S_i},y_{S_i})l(x_{S_i},y_{S_i},\theta) \tag{4}$$

式中:$n_s$ 为从训练集中构造出的特征词典的规模;$i$ 为特征词典中第 $i$ 个特征词;$P_S(x_{S_i}, y_{S_i})$,$P_T(x_{T_i}, y_{T_i})$分别为第 $i$ 个特征词在训练集和测试集中的边缘分布概率。这样可以通过$\frac{P_T(x_{T_i}, y_{T_i})}{P_S(x_{S_i}, y_{S_i})}$对训练集中的每个特征$(x_{S_i}, y_{S_i})$重新调整权重,使得 $P(D_S) \approx P(D_T)$,记 $\beta(x)$为文本分类中训练集的特征迁移量,定义为

$$\beta(x) = \frac{P_T(x_{T_i}, y_{T_i})}{P_S(x_{S_i}, y_{S_i})} \tag{5}$$

由于本文文本分类问题中不考虑存在类迁移的情况,即 $P(Y_T \mid X_T) = P(Y_S \mid X_S)$,问题进一步简化为

$$\beta(x) = \frac{P_S(x_{S_i})}{P_T(x_{T_i})} \tag{6}$$

在文本分类中,常用词频-反文档频率(Term frequency-inverse document frequency, TF-IDF)算法实现文本的数值化表示[13],每个特征词典中的词作为 1 个特征项,通过特征词的词频和文档频率计算其权重,由特征词组成的文本为 1 行向量,最终构成文本集的空间向量模型,记作 $\boldsymbol{W}(d,x)$,其中 $d$ 表示文档,$x$ 表示特征词。通过 TF-IDF 算法构造的空间向量模型可以较好地表示特征词的分布情况,因此

$$\beta(x) = \frac{P_S(x_{S_i})}{P_T(x_{T_i})} = \frac{\sum_{d \in S} \boldsymbol{W}(x_{S_i})}{\sum_{d \in T} \boldsymbol{W}(x_{T_i})} \tag{7}$$

## 3 实验结果

将本文提出的方法分别用于中文垃圾邮件过滤与中、英文网页分类中,在 CHI 统计特征选择方法构造出空间向量模型的基础上,利用基于迁移量的迁移学习方法对空间向量模型进行修正,并将修正前后的文本分类结果进行对比,以检验算法的有效性。

### 3.1 中文垃圾邮件过滤

本节采用 2.1 节中文电子邮件数据集,实验结果如表 2 所示。实验分为 4 组,每组训练集由 500 封正常邮件和垃圾邮件组成,测试集由 300 封正常邮件和垃圾邮件组成,特征词典的维度为 600 维。采用 SVM 作为学习分类器,其中核函数采用粒度高斯核,正则参数 $C$ 取 100,核参数 $\sigma$ 取 1.0。采用邮件过滤率(Correct rate, CR),正常邮件通过率 (Normal mail rate, NMR) 和正确过滤率(Correct rejection rate, CRR)作为评价指标[14],指标值越大越好,分别定义为

$$\mathrm{CR} = \frac{N_{H \to H} + N_{S \to S}}{\text{邮件总数}} \tag{8}$$

$$\mathrm{NMR} = \frac{N_{H \to H}}{N_{H \to S} + N_{H \to H}} \tag{9}$$

$$\mathrm{CRR} = \frac{N_{S \to S}}{N_{S \to H} + N_{S \to S}} \tag{10}$$

式中:$N_{H \to H}$为正常邮件被识别为正常邮件的数目;$N_{H \to S}$为正常邮件被识别为垃圾邮件的数目;$N_{S \to H}$为垃圾邮件被识别为正常邮件的数目;$N_{S \to S}$为垃圾邮件被识别为垃圾邮件的数目。通过观察表 2 可以发现,在垃圾邮件过滤准确率上,利用迁移量方法对模型进行修正后,其实验结果在 4 组数据集上均高于原有的 CHI 统计方法,其中相差最大的为第 3 组数据集,差值为 6.1%,相差最小的为第 4 组数据集,差值为2.8%。在正确过滤率上,新算法在 4 组数据集上均明显高于 CHI 统计方法,其中相差最大的为第

3 组数据集，差值为 13%；在正常邮件通过率上，两种算法则相差不大。表2～4中数值越大，性能越好。图 2 为 3 个评价指标在 4 组数据集中平均值的对比。从图 2 可以看出，利用迁移量方法进行修正后，整体的邮件过滤率提高了近 4%，并且在指标 CRR 中，较迁移前精度提高了 8.4%。为检验迁移学习算法的有效性，对文档频率特征选择方法构造的空间向量模型进行修正，实验结果如表 3 所示。从表 3 可以看出，在指标正常邮件通过率上，迁移量算法在 4 组数据集中均提高了文档频率特征选择方法的分类精度。但在指标正确过滤率上，4 组数据集的分类精度略有降低，因此在整体邮件过滤率上，只有第 4 组数据集的实验结果较好。通过从算法、过程和数据等多个方面，与基于 CHI 统计方法的实验对比发现，通过 CHI 统计方法构造的特征词典可以充分考虑特征词与邮件类别的相关关系，可以很好地代表训练集和测试集中特征词分布规律；因此对 CHI 统计方法构造的空间向量模型进行迁移后，可以进一步提高算法的精确率。而文档频率方法在特征选择过程中主要考虑特征词在训练集中出现的文档频率，以此选择出的特征词有一部分在测试集中基本不出现或出现频率较低，并且这种情况主要出现在从垃圾邮件中选择出的特征词，因此导致进行迁移学习算法后，垃圾邮件过滤精度降低。

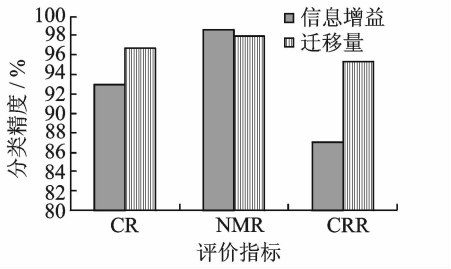



图 2 垃圾邮件各指标的平均值

Fig. 2 Average value of spam filtering

**表 2 垃圾邮件过滤精度**

**Tab. 2 Accuracy of spam filtering** %

| 组别 | 方法 | CR | NMR | CRR |
|---|---|---|---|---|
| 第 1 组 | CHI 统计 | 94.0 | 98.7 | 89.3 |
| | 迁移量 | 97.6 | 98.6 | 96.7 |
| 第 2 组 | CHI 统计 | 94.7 | 99.3 | 90.0 |
| | 迁移量 | 97.5 | 99.3 | 95.7 |
| 第 3 组 | CHI 统计 | 90.2 | 99 | 81.3 |
| | 迁移量 | 96.3 | 98.3 | 94.3 |
| 第 4 组 | CHI 统计 | 92.7 | 97.7 | 87.7 |
| | 迁移量 | 95.3 | 95.7 | 95.0 |

**表 3 垃圾邮件过滤精度**

**Tab. 3 Accuracy of spam filtering** %

| 组别 | 方法 | CR | NMR | CRR |
|---|---|---|---|---|
| 第 1 组 | 文档频率 | 98.50 | 99.33 | 97.67 |
| | 迁移量 | 97.83 | 99.67 | 96.00 |
| 第 2 组 | 文档频率 | 96.50 | 96.00 | 97.00 |
| | 迁移量 | 95.33 | 99.00 | 91.67 |
| 第 3 组 | 文档频率 | 97.83 | 98.67 | 97.00 |
| | 迁移量 | 95.17 | 99.67 | 90.67 |
| 第 4 组 | 文档频率 | 95.00 | 93.67 | 96.33 |
| | 迁移量 | 96.50 | 99.33 | 93.67 |

### 3.2 中文网页分类

实验采用由复旦大学整理的中文网页语料库[13]，从中选取出环境、计算机、交通、教育、太空、体育、农业、艺术和政治 10 大类共 2 600 篇文本。实验结果如表 4 所示。实验共分为 4 组，每组随机抽取 1 000篇作为训练集，另随机抽取 400 篇作为测试集，其中每类文本在训练集与测试集中所占比例相同。实验通过查准率、召回率和 $F_1$ 评价算法的好坏，评价指标分别定义为

$$\text{查准率}=\frac{\text{检索到相关文档数}}{\text{检索到文档总数}} \tag{11}$$

$$\text{召回率}=\frac{\text{检索到相关文档数}}{\text{所有相关文档总数}} \tag{12}$$

$$F_1=\frac{2\times\text{查准率}\times\text{召回率}}{\text{查准率}+\text{召回率}} \tag{13}$$

通过观察表 4 可以发现，在对训练集数据进行迁移学习后，网页分类的 3 个指标均明显得到提升。

在体现整体分类精度的 $F_1$ 中,经迁移学习后,4 组数据集提高了 3.9%～11%;在查准率上,增幅最大的是第 1 组,为 13.7%,增幅最小的是第 4 组,为 4.8%;在召回率上,第 1 组经迁移学习后达到了 100%,其余 3 组的召回率也普遍较高。图 3 为 3 个评价指标在 4 组数据集中平均值的对比。从图 3 可以看出,在对训练集模型进行迁移后,3 个评价指标 $F_1$、查准率和召回率均明显提高,其中查准率提升了 8.9%,平均召回率值达到了 99%以上。

**表 4 中文网页分类精度**

**Tab. 4 Accuracy of Chinese web page classification** %

| 组别 | 方法 | $F_1$ | 查准率 | 召回率 |
|---|---|---|---|---|
| 第 1 组 | CHI 统计 | 74.5 | 60.9 | 95.5 |
| | 迁移量 | 85.5 | 74.6 | 100.0 |
| 第 2 组 | CHI 统计 | 75.4 | 62.6 | 94.8 |
| | 迁移量 | 83.5 | 72.4 | 98.6 |
| 第 3 组 | CHI 统计 | 80.8 | 69.7 | 96.0 |
| | 迁移量 | 86.7 | 77.1 | 99.0 |
| 第 4 组 | CHI 统计 | 81.6 | 70.6 | 96.5 |
| | 迁移量 | 85.5 | 75.4 | 98.7 |

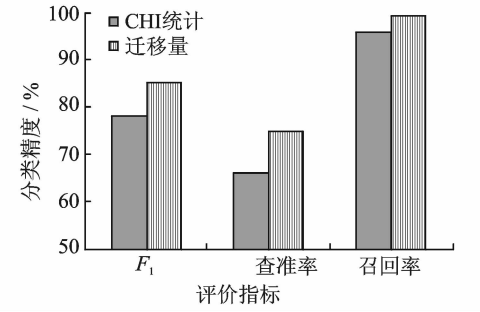



图 3 中文网页分类各指标的平均值

Fig. 3 Average value of Chinese web page classification

### 3.3 英文网页分类

实验采用 20-newsgroup 英文网页语料库[15],从中选取出 alt. atheism, rec. autos, rec. sport, baseball, sci. crypt, sci. electronics, misc. forsale, comp. graphics, rec. sport. hockey, sci. med, rec. motorycles, sci. space 10 大类共 2 600 篇网页。实验共分为 4 组,每组随机抽取 1 000 篇作为训练集,另随机抽取 400 篇作为测试集,其中每类文本在训练集与测试集中所占比例相同,通过查准率、召回率和 $F_1$ 评价算法的好坏。实验结果如表 5 所示。表 5 表明,通过迁移量算法对训练集空间向量模型进行修正后,网页分类的 3 个指标均明显得到提升:在指标 $F_1$ 中,迁移量算法在 4 组数据集上提高了 6.4%～13.1%;在关键指标查准率上迁移量算法在 4 组数据集上有较大提升,增幅最大的是第 1 组,为 16.82%,增幅最小的是第 4 组,为8.69%;在召回率上,迁移量算法也有 2.39%～3.1%的提升。图 4 为 3 个评价指标在 4 组数据集中平均值的对比。从图 4 可以看出,通过迁移量算法对训练集空间向量模型进行修正后,3 个评价指标 $F_1$、查准率和召回率均明显提高,关键指标查准率平均提升了 11.44%,平均召回率达到 98%以上。

**表 5 英文网页分类精度**

**Tab. 5 Accuracy of English web page classification** %

| 组别 | 方法 | $F_1$ | 查准率 | 召回率 |
|---|---|---|---|---|
| 第 1 组 | CHI 统计 | 72.46 | 57.99 | 96.56 |
| | 迁移量 | 85.47 | 74.81 | 99.66 |
| 第 2 组 | CHI 统计 | 75.47 | 62.66 | 94.86 |
| | 迁移量 | 82.79 | 73.59 | 97.29 |
| 第 3 组 | CHI 统计 | 79.88 | 68.39 | 96.00 |
| | 迁移量 | 86.84 | 77.72 | 98.39 |
| 第 4 组 | CHI 统计 | 79.39 | 67.00 | 95.97 |
| | 迁移量 | 85.79 | 75.69 | 99.00 |

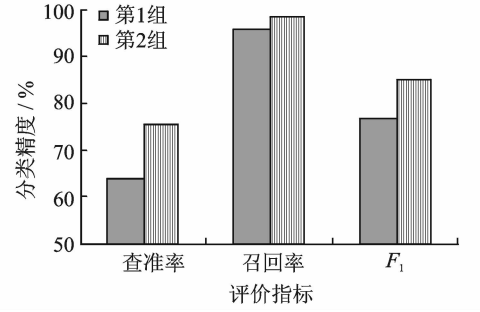



图 4 英文网页分类各指标的平均值

Fig. 4 Average value of English web page classification

## 4 结束语

本文首先提出文本分类中存在特征迁移问题,并通过实验进行了验证。其次提出一种基于迁移量的迁移学习解决方法,对从训练集中构造出的模型进行修正,以减少训练集与测试集模型中存在的差异,最终提高文本的分类精度。最后在中文垃圾邮件过滤、中文网页分类和英文网页分类中对提出的方法进行验证。文本分类中的迁移学习问题还有很大的研究空间,例如部分词语所含的信息量会随着时间的发展而变化。这样在文本分类中,不仅仅是特征词的特征分布发生变化,而且组成特征空间的特征词典需要不断地更新。此外对于引言中提到的邮件过滤问题,其中存在特征词类迁移的情况,不能用本文提出的方法进行有效解决。在未来的工作中,还需要进一步的学习和探讨。

**参考文献:**


[1] Ministry of Education of the People's Republic of China. The national language committee. 2011 Chinese language life report [EB/OL]. http://www.moe.gov.cn/, 2012-05-30.

[2] China Internet Association. The third anti-spam survey report of China in 2005[EB /OL]. http://www.isc.org.cn/, 2005-12-09.

[3] China Internet Association. The first anti-spam survey report of China in 2013[EB /OL]. http://www.isc.org.cn/, 2013-07-01.

[4] Sinno J P, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[5] Dai W, Yang Q, Xue G R. Boosting for transfer learning[C]// 24th International Conference on Machine learning. New York: [s. n.], 2007: 193-200.

[6] Wu P, Dietterich T G. Improving SVM accuracy by training on auxiliary data sources[C]// 21th International Conference on Machine learning. Banff: [s. n.], 2004: 1-8.

[7] Arnold A, Nallapati R, Cohen W W. A comparative study of methods for transductive transfer learning [C]// 7th IEEE International Conference on Data Mining Workshops. Washington: IEEE Computer Society, 2007: 77-82.

[8] Lee S I, Chatalbashev V, Vickrey D, et al. Learning a meta-level prior for feature relevance from multiple related tasks[C]// 24th International Conference on Machine Learning. Corvallis: [s. n.], 2007: 489-496.

[9] Yang Y M. An evaluation of statistical approaches to text categorization [J]. Information Retrieval, 1999, 1(1): 76-88.

[10] 邸鹏,段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理, 2014, 29(1): 71-75.
Di Peng, Duan Liguo. New naive Bayes text classification algorithm[J]. Journal of Data Acquisition and Processing, 2014, 29(1): 71-75.

[11] Drucker H, Wu Donghui, Vapnik V N. Support vector machines for spam categorization[J]. Neural Networks, 1999, 10(5): 1048-1054.

[12] CCERT. Spam dataset[EB/OL]. http://pan.baidu.com/s/1gdtRBiV, 2013-06-11.

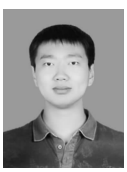
[13] Su J S, Zhang B F, Xu X. Advance in machine learning based text categorization[J]. Journal of Software, 2006, 17(9): 1848-1859.

[14] Saad O, Darwish A, Faraj R. A survey of machine learning techniques for spam filtering [J]. International Journal of Computer Science and Network Security, 2012, 12(2): 66-73.

[15] Fudan University Chinese Corpus. Chinese web page corpora[DB/OL]. http://pan.baidu.com/s/1o61KQ74, 2015-06-11.



**作者简介:**



**赵世琛**(1989-),男,硕士研究生,研究方向:机器学习。

**王文剑**(1968-),女,博士,教授,博士生导师,研究方向:机器学习、计算智能等,E-mail: wjwang@sxu.edu.cn。