文章编号：1003-0077(2015)04-0120-06

面向文本分类的特征词选取方法研究与改进

李国和[1,2,3]，岳 翔[1,2]，吴卫江[1,2,3]，洪云峰[3]，刘智渊[3]，程 远[3]

（1. 中国石油大学(北京) 地球物理与信息工程学院，北京 102249；
2. 中国石油大学(北京) 油气数据挖掘北京市重点实验室，北京 102249；
3. 石大兆信数字身份管理与物联网技术研究院，北京 100029）

摘 要：中文特征词的选取是中文信息预处理内容之一，对文档分类有重要影响。中文分词处理后，采用特征词构建的向量模型表示文档时，导致特征词的稀疏性和高维性，从而影响文档分类的性能和精度。在分析、总结多种经典文本特征选取方法基础上，以文档频为主，实现文档集中的特征词频及其分布为修正的特征词选取方法(DC)。采用宏F值和微F值为评价指标，通过实验对比证明，该方法的特征选取效果好于经典文本特征选取方法。

关键词：文本文档；特征词；特征选取；文本分类

中图分类号：TP391 **文献标识码**：A

Feature Word Selection for Document Classification

LI Guohe[1,2,3], YUE Xiang[1,2], WU Weijiang[1,2,3], HONG Yunfeng[3], LIU Zhiyuan[3], CHEN Yuan[3]

(1. College of Geophysics and Information Engineering, China University of Petroleum, Beijing 102249, China;
2. Beijing Key Lab of Data Mining for Petroleum Data, China University of Petroleum, Beijing 102249, China;
3. PanPass Institute of Digital Identification Management and Internet of Things, Beijing 100029, China)

Abstract: Feature words selection from texts is a significant step in Chinese text information pre-processing. After the segmentation of Chinese texts, a Vector Model constructed by feature words representing the Chinese text documents cannot avoid low accuracy of document classification (or document retrieval) due to the sparseness and high-dimension of feature words. On the basis of an analysis of several classical text feature selection methods, a new method of text feature selection (DC) is presented, which is based on a modified document frequency. Experiments prove the performance of DC, is better than that of typical other methods according to macro-F values and micro-F values.

Key words: Text document; Feature word; Feature selection; Text classification

1 引言

文本分类的目的是将未知类别的文本划归到具体的类别中，其在文档信息处理中主要具有信息过滤、内容查重、组织管理等功能，成为信息检索领域的重要应用之一[1]。由于文档的非结构化或半结构化特点，其中所隐含的信息难于直接进行比较，因此采用结构化的向量空间模型进行文本表示[2-3]。在该模型中，特征由文本中具有语义的词、短语等构成，统称特征词。这使得文档分类过程主要包括文本分词、特征词提取与优化、特征加权和分类器构建等阶段[1]。特征词的提取，除了可以去除停用词(如标点符号等)外，还可以完成文档的特征向量表示的结构化过程。由于采用统一特征向量形式表示所有文档，导致针对每一文档的特征向量具有高维性和稀疏性[4]。这不仅降低分类器的学习效率，而且影响甚至降低分类器的分类效果(包括精确率和召回率)。因此，通过特征选取方法，优化特征维数，选取确保分类效果不变或改善的特征词子集，成为文档分类的重要研究内容之一[5-6]。目前，特征词的选取方法主要采用统计学的方法[1,7]，但是没有考虑到特征词在文档中的分布特

收稿日期：2013-07-14 **定稿日期**：2013-11-12

基金项目：国家高新技术研究发展计划(2009AA062802)；国家自然科学基金(60473125)；中国石油(CNPC)石油科技中青年创新基金(05E7013)；国家重大专项子课题(G5800-08-ZS-WX)

性。针对这一不足，本文提出特征词分布的修正方法，完善特征词选取的功能。

2　相关研究工作

先简介一些相关基本概念，给出规范化的定义。

2.1　基本概念

$D=\{d_i|i=1,2,\cdots,n\}$为 n 个文档的文档集；$T=\{t_i|i=1,2,\cdots,k\}$为 k 个特征词的特征集；$C=\{c_i|i=1,2,\cdots,m\}$为文档的类别集；$W=\{W_i|i=1,2,\cdots,k\}$为所有特征值域的集合(即 ω：$D\times T\rightarrow W_i$，对于 $\forall d\in D,\omega(d,t_i)\in W_i$ 为文档 d 的特征词 t_i 的特征值，也称特征词加权的权值)；$D(t)\subseteq D$ 为含有特征词 $t\in T$ 的文档集，$D(c)\subseteq D$ 为属于类别 $c\in C$ 的文档集，$D(t,c)\subseteq D$ 为属于类别 $c\in C$ 并且含有特征词 $t\in T$ 的文档集，$DF(t)=|D(t)|$ 为含有特征词 $t\in T$ 的文档频(即文档集 D(t)中的文档数)，$DF(c)=|D(c)|$ 为属于类别 $c\in C$ 的文档频(即文档集 D(c)中的文档数)，$DF(t,c)=|D(t,c)|$ 为属于类别 c 并且含有特征词 $t\in T$的文档频(即文档集 D(c)中含有特征词 t 的文档数)，$TF(t,D')$为文档集 $D'\subseteq D$ 中出现特征词 $t\in T$ 的词频(即文档集 D'中出现特征词 t 的次数)。定义 c 类别概率、t 词频概率、t 词频与 c 类别关系的联合概率和条件概率，如下所示。

$$P(c)=\frac{DF(c)}{|D|} \tag{1}$$

$$P(t)=\frac{DF(t)}{|D|} \tag{2}$$

$$P(t\wedge c)=\frac{DF(t,c)}{|D|} \tag{3}$$

$$P(c|t)=\frac{P(t\wedge c)}{P(t)} \tag{4}$$

$$P(t|c)=\frac{P(t\wedge c)}{P(c)} \tag{5}$$

t、c 分别为非特征词 t 和非类别 c，也可定义有关概率。实际上，这些概率均为关于文档频的频率。

文本分类就是构造分类器函数 φ：$D\times T\rightarrow C$，即 $\varphi(d,\omega_1(d,t_1),\omega_2(d,t_2),\ldots,\omega_k(d,t_k))\in C$。特征词选取就是选取特征词子集 $SubT\subseteq T$，并使分类器函数 η：$D\times SubT\rightarrow C$ 满足 $\eta(d,SubT)=\varphi(d,T)$，即特征子集与原有特征集具有相同的分类能力。

2.2　经典文本特征选取方法

经典文本特征选取有信息增益 IG[8]、互信息熵 MI[8]、卡方 χ^2[8]、文档证据权 WET[9]、期望交叉熵 EC[9]、相对熵 H[5]等，还有与文本类别无关的文档频 DF[8]等，它们均为基于统计学的方法。各种特征词的选取方法是定义特征词 $t\in T$ 对文档所有类别 $c\in C$ 的分类能力评估函数。这些函数涉及如 2.1 所述的概率。通过特征词分类评估函数评价每个特征词 t 的分类能力，选取分类能力强的若干个特征词构成特征词子集 SubT。

这些评估函数(DF 方法除外)都体现出特征词 t 与文档类别 c 之间统计意义下的关联关系，如信息增益 IG：

$$\begin{aligned}IG(t)=&-\sum_{i=1}^{m}P(c_i)\lg P(c_i)+\\&P(t)\sum_{i=1}^{m}P(c_i|t)\lg P(c_i|t)+\\&P(\bar{t})\sum_{i=1}^{m}P(c_i|\bar{t})\lg P(c_i|\bar{t})\end{aligned} \tag{6}$$

其强调特征词 t 对文本分类的影响，即特征词 t 决策类别 c 的能力，而互信息 MI：

$$\begin{aligned}I(t,c_i)&=\lg\frac{P(t\wedge c_i)}{P(t)P(c_i)}=\lg\frac{P(c_i|t)}{P(c_i)}\\&=\lg\frac{P(t|c_i)}{P(t)}\end{aligned} \tag{7}$$

$$MI(t)=\max_{i=1}^{m}P(c_i)I(t,c_i)$$

或

$$MI(t)=\sum_{i=1}^{m}P(c_i)I(t,c_i) \tag{8}$$

其强调特征词 t 与类别 c 之间的相互关联性，即特征词 t 决策类别 c 并且类别 c 也决策特征词 t 的能力。又如相对熵 RE 是文档类别 c 概率分布 $\{P(c_i)|i=1,2,\ldots,m\}$，与 $t\wedge c$ 概率分布 $\{P(t\wedge c_i)|i=1,2,\ldots,m\}$的相似性比较，表达形式如式(9)所示。

$$\begin{aligned}RE(t)=\sum_{i=1}^{m}\Bigg[&P(c_i)\lg\frac{P(c_i)}{(P(c_i)+P(t\wedge c_i))/2}+\\&(1-P(c_i))\lg\frac{1-P(c_i)}{1-(P(c_i)+P(t\wedge c_i))/2}\Bigg]\end{aligned} \tag{9}$$

其也是反映文档类别 c 与特征词 t 的关联关系。还有其他评估函数，都全部或部分涉及到 $P(c|t)$、$P(\bar{c}|\bar{t})$、$P(\bar{c}|t)$、$P(c|\bar{t})$、$P(c)$、$P(t)$ 概率，而且可以看出来：$P(c|t)=\frac{P(t\wedge c)}{P(t)}$ 和 $P(\bar{c}|\bar{t})=\frac{1-(P(c)+P(t)-P(t\wedge c))}{1-P(t)}$ 具有一致性(即一个增大，另一个不减小)，$P(\bar{c}|t)=\frac{P(t)-P(t\wedge c)}{P(t)}$ 和 $P(c|\bar{t})=\frac{P(c)-P(t\wedge c)}{1-P(t)}$ 也具有一致性，而 $P(c|t)$ 和 $P(\bar{c}|t)$ 具有互斥性(即一个增大，另一个减小)，$P(\bar{c}|t)$ 和 $P(c|\bar{t})$ 也具有互斥性。

这些概率及其一致性和互斥性的内涵为：$P(c|t)$表示特征词t的出现对类别c的肯定，$P(\bar{c}|\bar{t})$表示特征词t不出现对其他类别c的否定，$P(\bar{c}|t)$表示特征词t的出现对类别c的否定，$P(c|\bar{t})$表示特征词t不出现对类别c的肯定。在给定的文档集中，$P(c)$、$P(t)$是常数，所有评估函数为关于$P(c|t)$、$P(\bar{c}|\bar{t})$的非线性递增函数，关于$P(\bar{c}|t)$、$P(c|\bar{t})$的非线性递减函数，而$P(c)$、$P(t)$作为归一化的系数。不同评估函数只是这些概率组合和系数各有差异。总之，$P(c|t)$、$P(\bar{c}|\bar{t})$表示特征词t对类别c的决策程度，它与评估函数的递增关系意味着特征词t对分类的正面贡献。$P(\bar{c}|t)$、$P(c|\bar{t})$表示特征词t不能决策类别c的程度，它与评估函数的递减关系意味着特征词t对分类的负面贡献。

2.3 存在问题

经典特征选取方法只是涉及到文档频DF，并不涉及到词频TF。实际上，词频TF对文档分类具有很大的影响。一般情况下，文档d∈D中某一特征词t的词频TF(t,{d})越高，文档内涵越明确，文档的类别就越清晰，所以该特征词对文档的分类能力也越强。鉴于此，文献[10]提出了基于TFIDF的文档特征选取方法，其核心思想是用属于c_i类且含有特征词t的文档数来刻画特征词t对c_i类文档的分类能力，即

$$\begin{aligned}\mathrm{IDF}(t,c_i) &= \log\left(\frac{\mathrm{DF}(t,c_i)}{\mathrm{DF}(c_i)}\times|D|\right)\\ &= \log(P(t\mid c_i)\times|D|), \quad (10)\end{aligned}$$

还结合词频TF，定义特征词t对c_i类文档的分类能力

$$\mathrm{DIS}(t,c_i)=\frac{\mathrm{TF}(t,D(c_i))\times\mathrm{IDF}(t,c_i)}{\mathrm{TF}(t,D(\bar{c}_i))\times\mathrm{IDF}(t,\bar{c}_i)},c_i\in C \quad (11)$$

对$\mathrm{DIS}(t,c_i)$进行从大到小排序，得到特征词t对所有类别的分类能力从大到小序列$\mathrm{DIS}_1(t,c'_1)$，$\mathrm{DIS}_2(t,c'_2),\ldots,\mathrm{DIS}_{|C|}(t,c'_{|C|})$，最后特征词t的分类能力定义为$\mathrm{DIS}_1(t,c'_1)-\mathrm{DIS}_2(t,c'_2)$。这一方法尽管考虑到不同文档类别间的词频分布，但没有考虑到同类文档中词频分布的不均匀性对分类效果的影响。因此，针对这一问题，本文提出基于文档词频分布修正的特征词选取方法(DC)。

3 基于文档词频分布修正的特征词选取方法

基于文档词频分布修正的特征词选取方法(DC)以文档频DF为主，兼顾文档集中词频的分布对文档分类的影响，涉及以下基本算子：

(1) $P(c_i\mid t)\lg\frac{P(c_i\mid t)}{P(c_i)}=P(c_i\mid t)I(t,c_i)$表示特征词t的出现对文档类别$c_i$概率分布的正面贡献，即特征词t对文档类别$c_i$的分类能力。

(2) $\overline{\mathrm{TF}}(t,D(c_i))=\frac{\mathrm{TF}(t,\mathrm{D}(c_i))}{|\mathrm{D}(c_i)|}$为$c_i$类文档集中每个文档含有特征词t的平均词频，表示特征词t在c_i类文档中的普遍性。该值越大，越能代表该类文档，分类c_i类能力越强。

(3) $\sum_{d\in D(c_i)}|\mathrm{TF}(t,\{d\})-\overline{\mathrm{TF}}(t,D(c_i))|$为$c_i$类文档集内特征词t词频离散程度，表示$c_i$类文档集内特征词t的均匀性。如果该值越小，意味着特征词t分布越均匀，越能代表该类文档，分类c_i类能力越强。

(4) $\frac{\mathrm{TF}(t,D(c_i))}{\mathrm{TF}(t,D(t))}$为$c_i$类文档集对特征词t的占有率，表示特征词t与$c_i$类文档的关联程度。该值越大，特征词t对$c_i$类文档分类能力越强。

综合上述四个基本算子，特征词t的文档分类能力评估函数定义如下：

$$\begin{aligned}\mathrm{DocClassify}(t) &= P(t)\sum_{i=1}^{|c|}\left[\frac{\mathrm{TF}(t,\mathrm{D}(c_i))}{\mathrm{TF}(t,D(t))}\times\right.\\ &\frac{1}{1+\sum_{d\in D(c_i)}|\mathrm{TF}(t,\{d\})-\overline{\mathrm{TF}}(t,D(c_i))|}\times\\ &\left.P(c_i\mid t)\lg\frac{P(c_i\mid t)}{P(c_i)}\right]=\sum_{i=1}^{|c|}\left[\frac{\mathrm{TF}(t,\mathrm{D}(c_i))}{\mathrm{TF}(t,D(t))}\times\right.\\ &\frac{1}{1+\sum_{d\in D(c_i)}|\mathrm{TF}(t,\{d\})-\overline{\mathrm{TF}}(t,D(c_i))|}\times\\ &\left.P(c_i\wedge t)I(t,c_i)\right] \quad (12)\end{aligned}$$

从式(10)可以看出，$P(t)P(c_i\mid t)\lg\frac{P(c_i\mid t)}{P(c_i)}=P(c_i\wedge t)I(t,c_i)$为特征词t与类别$c_i$之间的相互决策能力，也就是特征词t的分类能力，而$\frac{\mathrm{TF}(t,\mathrm{D}(c_i))}{\mathrm{TF}(t,D(t))}\times\frac{1}{1+\sum_{d\in D(c_i)}|\mathrm{TF}(t,\{d\})-\overline{\mathrm{TF}}(t,D(c_i))|}$表明了特征词t在$c_i$类文档集中词频较大，而且分布比较均匀(即类中每个文档的词频大小相当)，而在其他类词频较小，而且分布不均匀(即类中每个文档的词频不相当)，则表示此特征词的分类能力较强，因此该方法充分体现了类内文档相似性高、类间

文档差异性大的特点。

4　实验结果对比

4.1　实验基础

实验数据采用复旦大学计算机学院提供的文档集，其类别数|C|＝20，文档数|D|＝19 637。采用ICTCLAS 分词系统进行分词，得到特征词数|T|约为 13 万。采用 TFIDF 对所有文档进行加权[7]：

$$\omega(d,t) = \mathrm{TF}(t,\{d\}) \times \log\left(\frac{|D|}{|D(t)|}\right) \quad (11)$$

表示文档 d 的特征词 t 的权值。

分类器选用 KNN[11]，并取 K 值为 15。对文档集 D 中的所有文档进行统一加权后，采用 5-交叉验证实验，即所有文档随机均分成五组，一组为测试集，其他四组为训练集，共进行五次实验，最后评价指标的平均值作为特征词选取的依据。

4.2　效果评价标准

文档分类的评价标准有精确率、召回率、F 值。设 x(c)为测试文档的测试结果与真实类别均为 c 类的文档数；y(c)为测试文档的测试结果为 c 类的文档数；z(c)为测试文档类别为 c 类的文档数，文档分类的评价标准定义：精确率 $pre(c) = \frac{x(c)}{y(c)}$，召回率 $rec(c) = \frac{x(c)}{z(c)}$，F 值 $\mathrm{F}(c) = \frac{2 \times pre(c) \times rec(c)}{pre(c) + rec(c)}$，宏精确率 $macro\_pre = \frac{\sum_{\forall c \in C} pre(c)}{|C|}$，宏召回率 $macro\_rec = \frac{\sum_{\forall c \in C} rec(c)}{|C|}$，宏 F 值 $macro\_\mathrm{F} = \frac{\sum_{\forall c \in C} \mathrm{F}(c)}{|C|}$，微精确率 $micro\_pre = \frac{\sum_{\forall c \in C} x(c)}{\sum_{\forall c \in C} y(c)}$，微召回率 $micro\_rec = \frac{\sum_{\forall c \in C} x(c)}{\sum_{\forall c \in C} z(c)}$，微 F 值 $micro\_\mathrm{F} = \frac{2 \times micro\_pre \times micro\_rec}{micro\_pre + micro\_rec}$。可以看出，宏 F 值和微 F 值综合了召回率和精确率，因此宏 F 值和微 F 值对特征词选取进行评价。

4.3　特征词分类能力有效性实验

根据式(10)对每一特征词的分类能力进行评估，并根据评估值从大到小对所有特征词进行排序。为了证明此特征选取方法的有效性，从有序的特征集中分别“从前到后”(即正向选取)、“从后到前”(即反向选取)和“随机”(即随机选取)选取特征词 n 个，构成三个 n 维特征向量，分别进行文档分类效果实验。特征向量维数 n 的范围为 100 到 4 000。每隔 100 个特征词做一次 5-交叉实验。实验结果如图 1 和图 2 所示。从实验结果可看出：①正向选取的特征词集分类效果好于反向选取特征词集的分类效果，而随机选取特征词集的分类效果介于正向选取和反向选取的分类效果之间。②随着特征词数的增加，正向选取特征词的文档分类效果逐渐变好，而反向选取特征词的分类效果基本不变。说明反向选取的特征词分类能力特别弱。③特征词数目大于 4 000 以后，正向选取特征集的分类效果基本保持不变，随机选取特征集和反向选取特征集的分类效果在缓慢逐渐增大，但最大值也难于接近正向选取特征集的分类效果。其他特征词分类能力的评价实验结果与宏 F 值和微 F 值测试结果具有相似的变化趋势。说明有序特征集中靠前的特征词分类能力比较强，特征选取方法 DC 能够对特征词分类能力进行有效评估，成为特征词选取的依据。

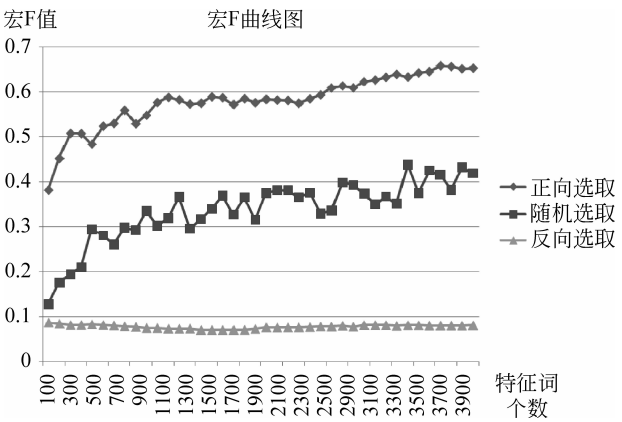

图 1　宏 F 反映特征集分类能力

4.4　文本特征选取方法对比实验

对文档集进行 TFIDF 加权后，分别采用词频修正 DC、文档频 DF、信息增益 IG、互信息熵 MI、统计量 χ2(CHI)、文档证据权 WET、期望交叉熵 ECE 和 DIS 以及相对熵 RE 进行文档分类效果对比实验，实验结果如图 3 和图 4 所示。

由图 3 所示，所有特征选取方法的微 F 值随着特征词的增多都能达到一个比较稳定的效果，其中

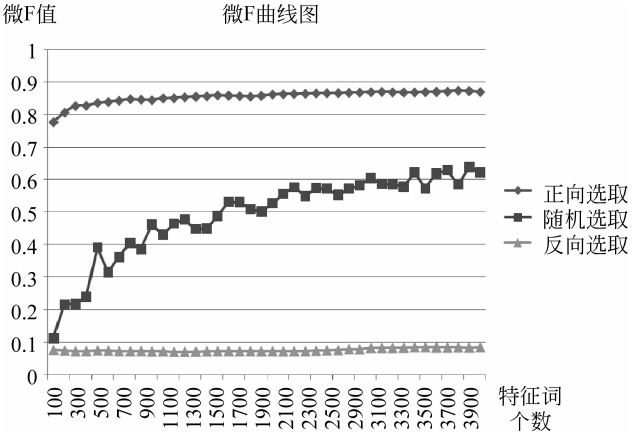

图 2 微 F 反映特征集分类能力

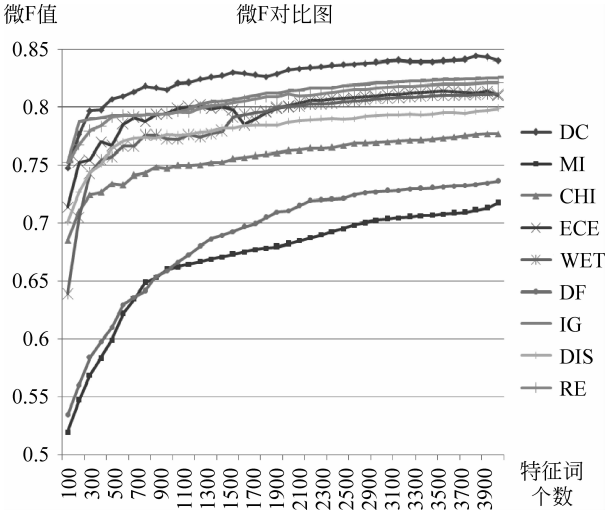

图 3 不同特征词选取方法比较(微 F)

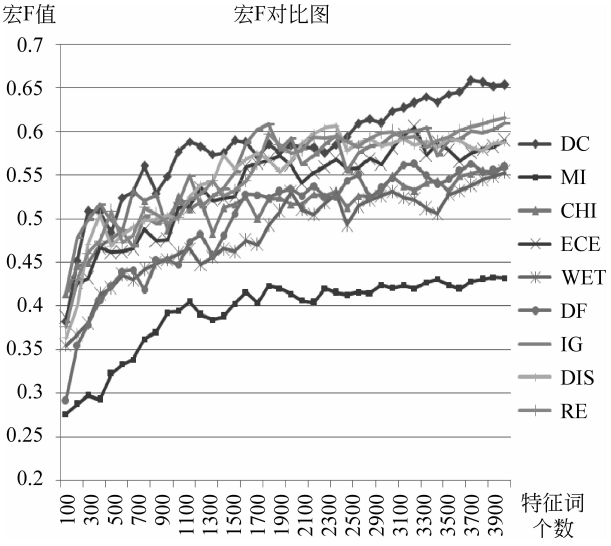

图 4 不同特征词选取方法比较(宏 F)

DC 方法优于其他方法,最快达到的稳定值和最大值。由图 4 可知,上述所有方法的宏 F 值随着特征词的增多都呈现明显的上升趋势,但是 DC 方法上升趋势更明显,且达到的极值最大。当特征词大于 4 000 个以后,微 F 值和宏 F 值就没有明显增大趋势,文档分类效果基本达到极值。其他特征词分类效果评价的变化趋势与微 F 值和宏 F 值实验结果变化趋势相似。

5 结束语

通过分析现有多种基于统计学的文本特征词选取方法,其实质上是利用某个特征词在一个文档中是否出现对文档类别的概率分布的影响来刻画特征词对文档分类能力,只是应用到文档频的信息。而这些特征选取方法缺乏考虑文档词频和文档词频分布对文档分类的影响。针对这个缺陷,以文档频为主,结合文档词频和文档词频分布为修正,重新定义文本特征词分类能力的评估函数。在此基础上完成特征词的分类能力排序,形成基于词频分布修正的文本特征词选取方法 DC。目前,文本特征词选取方法主要是构造特征词分类能力评估函数,实现特征词分类能力排序。以特征词分类能力为启发信息,采用分类能力强的特征词组合(即特征词分类能力的强强联合),逐一测试每一组合的分类效果(即精确率、召回率、F 值),人工选取认可的特征词子集集,但还缺少特征词自动选取方法。另一方面,特征词选取后采用其他方法(主要是 TFIDF 方法)对特征词加权,即特征词选取和特征词加权是分离的。因此,下一步工作是实现自动特征选取方法和特征加权后再进行特征词选取的研究。

参考文献

[1] 苗夺谦,卫志华. 中文文本信息处理的原理与应用[M]. 北京:清华大学出版社,2007

[2] 刘铭. 大规模文档聚类中若干关键问题的研究[D]. 哈尔滨工业大学博士学位论文. 2010.

[3] 熊忠阳,张鹏招,张玉芳. 基于 χ2 统计的文本分类特征选择方法的研究[J],计算机应用,2008,28(2):513-514

[4] 熊云波. 文本信息处理的若干关键技术研究[D]. 复旦大学博士学位论文. 2006.

[5] 王 辉,张成锁,卓呈祥. 一种改进的相对熵特征选择方法[J]. 计算机工程,2011,37(10):167-169.

[6] 柴玉梅,王宇. 基于 TFIDF 的文本特征选择方法[J]. 微计算机信息,2006,22(8-3):24-26

[7] 苏丹. 一种基于最少出现文档频的文本特征提取方法

[J]. 计算机工程与应用，2012，48(10)：164-166＋178.

[8] Bong Ch，K. Narayanan. An empirical study of feature selection for text categorization based on term weightage[C]//Proceedings of the International Conference on Web Intelligence，2004：599-602.

[9] 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报，2004，18(1)：26-32.

[10] Saltong，Clementty. On the construction ofeffective vocabularies for information retrieval[C]//Proceedings of the 1973 Meet-ing on Programming Languages and Information Retrieva. l NewYork：ACM，1973：11.

[11] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2011.

[12] 陈键. 面向文本分类的特征词选取方法研究[D]. 合肥工业大学硕士学位论文. 2009.

[13] 余俊英. 文本分类中特征选择方法的研究[D]. 江西师范大学硕士学位论文. 2007.

[14] 周茜，赵明生等. 中文文本分类中的特征选择研究[J]. 中文信息学报，2003，18 (3)：17- 23.

[15] 单松巍，冯是聪，李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用，2003，39(22)：146-148

[16] Yang Yiming，Pedersen J O. A comparative study on feature selection in text categorization[C]//Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco，CA，USA：ICML97 Morgan Kaufmann Publishers Inc，1997.

李国和(1965—)，博士，教授，博士生导师，主要研究领域为智能信息处理，知识发现，数据可视化等。

E-mail：ligh@cup. edu. cn

吴卫江(1971—)，在职博士研究生，副教授，主要研究领域为智能信息处理，知识发现，数据可视化等。

E-mail：allan1226@163. com

岳翔(1988—)，硕士研究生，主要研究领域为智能信息处理，知识发现等。

E-mail：yuexiang19881@126. com