

一种面向机械领域文本分类器的设计

董丽丽, 魏胜辉

(西安建筑科技大学 信息与控制工程学院, 陕西 西安 710055)

摘 要: 提出了一种面向机械领域的文本分类器. 特征选择采用基于文档频率的特征提取法和灰色关联度计算相结合的方法, 有效降低了特征维数, 削弱了特征词之间的关联, 为采用贝叶斯分类创造了条件. 分类阶段引进了基于类别区分度的加权因子对朴素贝叶斯分类器进行优化. 实验证明, 该分类器能够有效地提高机械领域文本分类的召回率和正确率, 具有较好的使用效果.

关键词: 机械领域; 灰色关联分析; 贝叶斯分类器; 特征选择

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)04-0142-04

Design of Mechanical Information Text Classifier

DONG Li-li, WEI Sheng-hui

(School of Information & Control Engineering, Xi'an University of Arch & Tech, Xi'an 710055, China)

Abstract: A information text classifier of machinery-oriented is proposed in this paper. The method of document frequency and grey relation analysis are used to select feature, which reduce the characteristics of dimensionality, weaken relation between feature words and create the conditions for the Bayes. The Bayesian Classifier is ameliorated by using word's kinds-difference as weighted factor. The experimental results indicate that the classifier is able to improve recall and precision, and is useful in practice.

Key words: machinery-oriented; grey relation analysis; Bayesian classifier; feature selection

1 引言

随着互联网技术的迅速发展, 各行业信息化程度得以普遍提高, 机械行业相关的文本大规模积累, 如何对浩如烟海的文本进行分类以便于检索和管理, 是一个具有重要用途的研究课题. 目前常用的分类算法包括朴素的贝叶斯分类法、基于支持向量机的分类器、k-最近邻法、神经网络法、决策树分类法、模糊分类法等^[1]. 分类算法常应用的领域有反垃圾邮件处理, 医学图像分类, 网页分类等^[2-4], 针对机械领域文本分类还比较少见.

本文提出了一种面向机械领域的文本分类器, 可以准确有效的对机械类文本进行分类, 迅速整理大量基础资料, 为构建机械领域本体库, 实现机械产品的智能检索打下基础. 同时借助于分类结果企业

可以迅速检索、分析产品的性能、质量、规格、品种等相关属性, 可以增强企业的市场敏锐程度, 降低产品的开发成本, 提高产品性能和产品附加值.

2 机械文本分类器简述

机械领域文本分类包括两分类问题和多分类问题, 两分类问题是指将文本集合分为机械文本类和非机械文本类, 多分类问题是指待分类文本都是机械领域的文本, 将这些文本分成领域内的小类, 本文研究只讨论多分类问题.

机械领域文本分类流程如图 1 所示.

文本分类器分为训练部分和测试部分, 其中涉及到的关键技术有中文分词, 特征选择, 向量空间模型构建, 分类算法选取, 分类器性能评价等几个方面.

收稿日期: 2011-07-07; 修回日期: 2011-09-13

基金项目: 陕西省自然科学基金(2009JM8006); 陕西省教育厅专项科研项目(2010JK620)

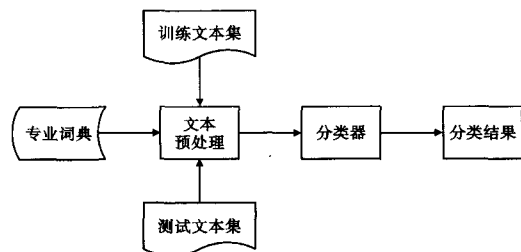


图1 机械领域文本分类流程图

根据机械领域专家的建议并结合实际项目需要,将机械领域文本分为以下几类:土方机械、混凝土机械、起重机械、筑养路机械、桩工机械、机动工业车辆、石方桥梁机械、地下及矿山机械、破碎设备、港口机械等十个分类。机械领域文本分类器可以将机械类文档迅速准确的分类到所属的类别中,便于用户查找。

3 文本预处理阶段的关键技术

3.1 分词和停用词处理

领域专业文本包含较多的专业词汇,仅使用通用词典进行分词准确率不高,专业领域词典的制定应确保其权威性与完整性。本文采用通用词典和专业词典结合的方式,分词时使用逆向最大匹配分词法。本研究中所采用的机械专业词典主要来源于机械专业词典、机械设计制造研究人员手工录入,还有部分词汇来源于搜狗实验室经过选择、去重等处理后得到。专业词典在收录词时采用长词优先的方法。如:“铲斗液压缸”作为一个词收录,同时也将铲斗,液压缸也收录在词典中。

机械专业词典部分词汇摘录如下:梅花式抓斗、动力绞盘、抓斗、可换斗齿、主卷扬筒、驱动轮、钢索滑轮、推压机构、钢丝绳悬吊、发动机、基坑等。

停用词处理利用停用词集剔除掉语义虚泛的虚词,如:的、我们、可以、是等。

3.2 类特征向量构建

文档频率(DF)是指出现某个特征项的文档的频率,从训练语料中统计出包含某个特征词的文档的频数^[5]。在文本分类中通常高频特征项和低频特征项对于类别的区分能力较低,但是在机械类文本处理中,某些专业词库中的低频词汇同样具备较高的区分能力,所以统计每个分类中候选特征词的频率,设定最大阈值和最小阈值,满足以下条件时做剔除处理:

(1)候选特征词的频率超出最大阈值且不在专

业词典之中。

(2)候选特征词的频率小于最小阈值且不在专业词典之中。

通过以上处理,剔除掉大量信息量较少的词汇,将出现频次适当并且重要的词组成候选特征词集 $W = (w_1, w_2, \dots, w_n)$ 。

用词频统计法标注每篇文档,可以得到每篇文档的文本向量,每篇文档可以表示为 $d_i = (w_{i1}, w_{i2}, \dots, w_{in})$, w_{in} 表示第 n 个特征词在第 i 篇的文档中出现的次数。用同样的方法标引所有文档,可得到该类的文本向量表示模型^[6]:

$$D = \begin{matrix} & w_1 & w_2 & \cdots & w_n \\ \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{matrix} & \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{pmatrix} \end{matrix}$$

将文本向量特征矩阵转置可得到文本特征词向量矩阵

$$D^T = \begin{matrix} & d_1 & d_2 & \cdots & d_m \\ \begin{matrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{matrix} & \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{m1} \\ w_{21} & w_{22} & \cdots & w_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{mn} \end{pmatrix} \end{matrix}$$

文本特征词向量矩阵可表示某一特征词在文档中的分布情况。

3.3 利用灰色关联进行二次特征选择

通过词频统计设定阈值的方法进行第一次特征选择后得候选特征词之间可能存在某些联系,不利于采用贝叶斯分类器进行分类。在此引入灰色关联分析^[7]做进一步处理。

对于两个系统或系统中两个因素之间,随时间或不同对象而变化的关联性大小的程度,称为关联度。

把每一个候选特征词在文本中的分布情况作为参考因素,计算候选特征词集中包含在专业词典中的词与不包含在专业词典中的词之间的关联度。若两个因素的变化具有一致性,则两个的关联程度就高,反之,则低。

具体计算方法如下:

将候选特征集调整为 $W = (w_1, w_2, \dots, w_l, w_{l+1}, \dots, w_n)$ 其中 $w_1, w_2, \dots, w_l$ 为包含在专业词典中的候选特征词, $w_{l+1}, \dots, w_n$ 为其他候选特征词,组成矩阵:

$$\begin{matrix} & w_1 & w_2 & \cdots & w_l \\ \begin{matrix} w_{l+1} \\ w_{l+2} \\ \vdots \\ w_n \end{matrix} & \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1l} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2l} \\ \vdots & \vdots & \cdots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nl} \end{pmatrix} \end{matrix}$$

ϵ_{ij} 表示 w_i 与 w_j 的灰色关联度, 计算公式

$$\epsilon_{ij} = \frac{1 + |s_i| + |s_j|}{1 + |s_i| + |s_j| + |s_j - s_i|} \quad (1)$$

当 ϵ_{ij} 的值较高时, 说明 w_i 与 w_j 关联度较高, 设定阈值, 剔除掉与专业词典中的候选词关联度较高的词汇. 例如铲斗(专业词典中词汇)与开挖(通用词典中词汇)的灰色关联度为 0.62, 则说明文档中“铲斗”和“开挖”同时出现在挖掘机类文档中概率较高, 两词的相关性较强, 而且铲斗更能标注土方机械类文档, 因此“开挖”相对于“铲斗”为冗余信息, 候选特征词集中保留“铲斗”, 剔除掉“开挖”. 照此思路对候选特征词进行处理, 得到每类的类特征向量.

经过灰色关联度的处理, 一方面选择出更能代表分类特征的特征词, 降低了维度, 另一方面, 剔除掉部分关联度较高的词汇, 确保了词与词之间的独立性. 由此得到的特征词集汇聚了该类文档的特征, 为贝叶斯分类创造了条件.

4 基于改进的朴素贝叶斯分类器

4.1 朴素贝叶斯分类器

朴素贝叶斯分类器的基本思想是利用特征项和类别的联合概率来估计给定文档的类别概率. 假设文本是基于词的一元模型, 即文本中当前词的出现依赖于文本类别, 而不依赖于其他词及文本的长度.

对于文本分类来说, 特征词 w_i 与 w_j 之间相互独立, 那么在给定目标类别的情况下, 可以采用特征词对文本进行向量表示, 则联合的特征词 $w_1, w_2, \dots, w_i$ 的概率等于各个特征词的概率乘积, 则条件概率可转化为式(2)所示:

$$\begin{aligned} P(d | C_i) &= P((w_1, w_2, \dots, w_m) | C_i) \\ &= \prod_{j=1}^m P(w_j | C_i) \end{aligned} \quad (2)$$

式中, $P(w_j | C_i)$ 表示分类器预测特征词 w_j 在类别 C_i 的训练文本中出现的概率. 得到式(3):

$$P(C_i | d) = P(C_i) \prod_{j=1}^m P(w_j | C_i) \quad (3)$$

文档 d 采用 TF 向量表示法, 如式(4)所示:

$$P(w_j | C_i) = \frac{1 + \text{TF}(w_j | C_i)}{|V| + \sum_{k=1}^{|V|} \text{TF}(w_k, C_i)} \quad (4)$$

式中, $|V|$ 表示训练文本特征词集中的所有特征词总数; $\text{TF}(w_j | C_i)$ 表示 C_i 分类的所有训练样本中特征词 w_j 出现的总次数, $\sum_{k=1}^{|V|} \text{TF}(w_k, C_i)$ 表示 C_i 类别中所有特征词出现的总次数.

朴素贝叶斯是基于概率统计的学习算法, 其主要特点是运行速度快且易于实现, 但是算法还存在以下的不足:

首先, 贝叶斯分类的前提是假设词与词之间是独立的, 而实际情况中词与词之间会存在某些联系, 这是导致分类效果不佳的一大原因.

其次, 贝叶斯分类算法假定所有的特征词对类别的区分程度是相同的. 而实际情况是专业词库中的特征词的区分效果要优于其他的特征词.

由于以上两点不足, 采用朴素贝叶斯分类器进行机械领域文本分类的效果不佳, 召回率和正确率都偏低(实验结果见表 1).

4.2 贝叶斯算法加权因子的改进

类别区别度用来表示某一个词语对于文本分类的贡献程度, 即词语的领域区别程度. 本文研究的是机械领域文本的二次分类, 这种分类的特点是文本都属于机械领域, 文本的相似化程度较高, 采用朴素贝叶斯算法进行分类的效果不佳. 造成这一结果的原因是朴素贝叶斯算法没有在特征词集中判断出类区别度较高的特征词并且给予它较高的权重. 例如“机械”, 在各个分类的文本分类中出现频次较高, 但是单根据这一词很难判定属于哪一类, 因此它的类区别度较低. 而文档中包含“铲斗”时, 很大程度上属于土方机械类, 因此它的类区别度较高, 对于区分度高的特征词应增大它的计算权重, 突出其在分类中所起的作用.

我们用 λ 表示类区分度, 定义为式(5):

$$\lambda_j = \frac{\text{TF}(w_j | C_i)}{\text{TF}(w_j | C') + 1} \quad (5)$$

式中, $\text{TF}(w_j | C_i)$ 表示 C_i 分类的所有样本文档中特征词 w_j 出现的总次数, $\text{TF}(w_j | C')$ 表示其他分类(除去 C_i 类)中特征词 w_j 出现的总次数.

由式(5)可以看出, 各类中出现概率相近的特征词, 区分度较小, 得到的权值很小. 各类中出现概率差别较大的特征词, 区分度较大, 得到的权值也较大.

将区分度 λ 作为加权因子引入贝叶斯公式, 可得式(6):

$$v_{NB} = \arg \max_j \lambda_j P(C_i) \prod_j P(w_j | C_i) \quad (6)$$

其中

$$\lambda_j = \frac{\text{TF}(w_j | C_i)}{\text{TF}(w_j | C') + 1},$$

$$P(w_j | C_i) = \frac{1 + \text{TF}(w_j | C_i)}{|V| + \sum_{k=1}^{|V|} \text{TF}(w_k, C_i)}$$

5 实验结果与分析

5.1 测试评价标准

采用召回率、正确率、F1 测试值来评估改进的贝叶斯分类器, a 表示分类器将输入文本正确的分类到某个类别的个数; b 表示分类器将输入文本错误的分类到某个类别的个数; c 表示分类器将输入文本错误的排除在某个类别之外的个数; d 表示分类器将输入文本正确的排除在某个类别之外的个数。

该分类器的召回率、正确率、F1 测试值分别采用式(7)~(9)计算:

$$\text{召回率} \quad r = \frac{a}{a+c} \times 100\% \quad (7)$$

$$\text{正确率} \quad p = \frac{a}{a+b} \times 100\% \quad (8)$$

$$\text{F1 测试值} \quad F1 = \frac{2 \times p \times r}{p+r} \quad (9)$$

5.2 测试结果

本实验采用的硬件环境为 Intel(R) Core(TM) 2Duo CPU T6600, 内存为 2 GB, 用 C# 语言在 VS2008 的开发环境上将面向机械领域的文本分类器进行了编程实现。

由于机械文本分类实验没有统一的数据集可供参考, 所以分类的语料库来自中国工程机械商贸网, 中国机械专家网, 中国机械网等专业资源系统网站。该语料共有 1 500 篇, 其中包含土方机械、混凝土机械、起重机械、筑养路机械、桩工机械、机动工业车辆、石方桥梁机械、地下及矿山机械、破碎设备、港口机械等十类文章各 150 篇。每类训练样本 100 篇, 待

测样本 50 篇。实验结果表明改进后的方法文本分类的召回率、正确率、F1 值均有大幅度提高。

6 结束语

本文设计并用 C# 语言实现了一种面向机械领域的文本分类器。经实验验证, 相对于朴素贝叶斯分类器, 该分类器具有更好的召回率和正确率。

由于分类器对文本训练集的要求很高, 选择高质量的训练文本填充训练集可以进一步提高分类器的精确度, 这是下一步工作的重点。

参考文献:

- [1] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2000.
- [2] 蒙海涛. 一种基于神经网络和主动反馈的反垃圾邮件技术的研究[J]. 微电子学与计算机, 2011, 28(6): 135-137.
- [3] 华翔, 孙蕾. 基于 SVM 的医学图像分类器的设计[J]. 微电子学与计算机, 2011, 28(6): 171-175.
- [4] 李晓黎, 刘继敏, 史忠植. 基于支持向量机与无监督聚类相结合的中文网页分类器[J]. 计算机学报, 2001, 24(1): 62-68.
- [5] Yang Yiming, Pederson J O. A comparative study on feature selection in text categorization [C]// Proceeding of the 14th International Conference on Machine Learning, 1997. Nashville: Morgan Kaufmann, 1997: 412-420.
- [6] 刘海峰, 刘守生, 姚泽清. 一种基于类别的混合型文本特征降维[J]. 微电子学与计算机, 2010, 27(10): 13-17.
- [7] Deng J. The theory and method of socioeconomic grey systems[J]. Social Sciences in China, 1984(6): 47-60.

作者简介:

董丽丽 女, (1960—), 教授, 硕士生导师. 研究方向为分布式系统与计算机网络应用、数据挖掘。

魏胜辉 男, (1987—), 硕士研究生. 研究方向为数据挖掘。