doi: 10.3969/j.issn.1007-855x.2009.04.026

基于文本倾向性分类技术的图书评价模型

邓忠莹,严 馨,周历生,王卫东,常彦峰
(昆明理工大学 图书馆,云南 昆明 650224)

摘要: 介绍了文本分类技术和文本倾向性分类技术,并基于文本倾向性分类技术分析了图书评论中的信息,研究如何将机器学习方法应用在图书评论的倾向性分类中,提出了一种图书评价模型构建的解决方法.

关键词: 文本分类;文本倾向性分类;图书评价;机器学习方法
中图分类号: P208 **文献标识码**: A **文章编号**: 1007-855X(2009)04-0121-04

Book Evaluate Model Based on Text Tendency Classification

DENG Zhong-ying, YAN Xin, ZHOU Li-sheng, WANG Wei-dong, CHANG Yan-feng
(Library, Kunming University of Science and Technology, Kunming 650224, China)

Abstract: Text classification technology and text tendency classification technology are firstly introduced in this paper. Information of book review based on text tendency classification is then analyzed. Finally, in the research of the application of machine learning techniques to book review classification, the establishment of a book evaluation model is proposed.

Key words: text classification; text tendency classification; book evaluation; machine learning method

0 引 言

网络的发展使得各种信息急剧增加,如何有效地管理和使用这些信息,使其发挥自身价值已成为信息处理技术迫切需要解决的重要问题.而作为解决这些问题的关键技术之一,文本分类技术得到普遍关注,有了较好的研究.文本分类技术在数字图书馆、搜索引擎、电子商务、信息监控、信息过滤等方面都有很广泛的应用价值.

文本分类技术基本可以分为基于内容主题的分类和基于主观情感的倾向性分类两种.主题性分类是指依据文本所讨论的主题进行分类,将文本归类于若干个已知文本类别 其中之一.主题性分类技术在分类信息搜索、个性化搜索、搜索结果自动分类等领域已经进入了实用化阶段.文本倾向性分类技术依据文本对所讨论的主题所持有的观点、立场、感受或态度等相关信息进行分类,将文本分为正、反两种或正、反、中三种等褒贬倾向评价模式[1];它主要用于对主观性文本(Subjective Texts)进行分析和处理[2].

国内外在文本分类领域进行了大量研究,目前基于主题的中文文本分类已经出现许多成熟、高效的方法[3,4],分类准确率已达 80%～90%;国外在英文文本倾向性分类方面的研究已处于发展阶段,国内在中文文本倾向性分类方面的研究仍处于初级阶段.要想在短时间内获得人们对于诸如人物、事件、传媒、产品、政策等有价值的评价信息往往是十分困难的[2].分类技术可以应用到信息过滤、推荐系统、信息安全、网络舆情分析、自动文摘提取、信息挖掘、新政策法规的民众反映等领域,是一项具有较大实用价值的分类技术,是组织和管理数据的有力手段[5],成为文本分类领域的研究热点.

本文面向图书评论的倾向性分类问题,通过一定规模的语料信息研究,结合现有成熟的基于主题的文

收稿日期: 2008-03-24.
第一作者简介:邓忠莹(1978-),女,在读硕士研究生.主要研究方向:数据挖掘.E-mail:358486613@qq.com

本分类技术,研究倾向性分类技术的特点方法,使其更适合对文本褒贬倾向性分类的任务要求;根据评论情感态度的不同,将评论分为褒扬类、贬斥类2种,并给出了实现图书评价倾向性分类的模型.

1 倾向性分类技术

与主题性分类相比较,倾向性分类更为复杂[3],主观性文本中的字、词、短语、句式、修辞方式、标点符号等都可能体现文本所要表达的倾向性;另一方面,不同类别文章中会出现相同的词语[3].目前提出的文本倾向性分类技术方法主要分为3类[6]:基于倾向性词加权的分类方法、基于语义模式的分类方法以及基于机器学习的分类方法.

倾向性词加权的分类方法分析的对象是词或短语,以具有强烈倾向意义的词作为基准词来计算测试词与它之间的相似度距离,并对测试值进行词语褒贬倾向赋值,从而分析文本的倾向性.此方法不受文本主题限制,具有较强的通用性[6],但依赖语言修辞学研究,分类难度较大.

基于语义模式的分类方法是利用倾向性词汇表、倾向性模式库对句子的语义关系进行分析,在文本表示模型中加入语义信息[6],识别出文本中的倾向性.但目前语义分析模式体系不成熟,应用于倾向性分类中难度较大.

基于机器学习的分类方法将文本倾向性分类看作褒贬两类分类问题,使用传统的分类方法加以解决,具有代表性的方法有神经网络法、贝叶斯分类算法、K-近邻算法、SVM等方法.该方法的难点在于解决特征空间高维性和文本表示向量稀疏性的问题,采用合适的特征选取和特征降维方法,有利于提高分类效率和准确率. Pang等人使用机器学习法验证文本倾向性分类的效果,其实验结果表明,SVM的分类准确性最高,可达到80%[7],但分类精度小于文本主题性分类.

2 图书评价倾向性分类

文本分类需要将文本表示成计算机可以处理的模式,目前常用的四个基于统计的模型是:传统的的布尔模型、扩展的布尔模型、向量空间模型以及概率模型;其中向量空间模型在文本分类领域中有较广泛的应用.

向量空间模型(Vector Space Model,简称VSM)是由哈佛大学的Gerard Salton提出的,Gerard Salton是现代信息检索的奠基人,开发了著名的SMART系统. VSM的基本思想是用项的向量空间来表示文档信息,根据向量空间的相似度比较进行文档分类.基本方法是将文档表示成如$(w_1, w_2, w_3, \cdots)$的形式,其中w_i是出现在文档中的各个特征词的权重.目前通常使用TF-IDF公式来计算权重,最常用公式如下:

$$W(t,d) = \frac{tf(t,d)\log(N/n_t + 0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\log(N/n_t + 0.01)]^2}} \quad (1)$$

其中,$W(t,d)$为词在文本d中的权重,而$tf(t,d)$为词t在文本d中的词频,N为训练文本的总数,n_t为训练文本集中出现词的文本数,分母为归一化因子[8].

研究结果表明支持向量机(SVM)方法分类准确率和精度最高,SVM是基于统计学的分类方法,它是对结构化风险最小化原则的近似,其理论基础是统计学习理论[9].本文模型采用SVM方法对文本进行倾向性分类.

2.1 模型框架

图1所示是针对某个文本的分类全过程,包括预处理、特征抽取、特别语句的处理和分类.

2.2 否定词处理方法

否定词的出现使其所匹配的特征词褒贬义性质变反,例如:这本书的内容不是很新颖;"新颖"在褒贬义词典中属于褒义词,其前面加上了"不是"这样的否定词,使整个句子性质变反转变成了贬义性质.论文对否定词的处理分为以下2个步骤:

1) 首先在褒贬义词典中提取所有含有否定词的褒贬义词语,将其作为特征项的一部分D_j;

2) 对于含上述词之外的含其它否定词的句子查找其所匹配的特征词,并对此特征词进行取反处理.

2.3 特征词选取算法

形容词、名词、副词、动词、成语和习惯语是最常见的具有语义倾向性的词语,文本中这类词语的语义倾向性决定分类结果,文献10研究表明,当特征词仅为形容词和名词时,依然有很高的准确率. 因此模型选择形容词和名词作为特征词来研究分类.

特征抽取是倾向性分类的重要步骤,由于向量空间模型的特征维数比较高,利用矩阵变换,将训练集矩阵 $T = (D_{ij})_{mn}$ 利用矩阵变换,转换成 $T' = (D_{ij})_{mp(p<<n)}$ 去除那些冗余的、对分类贡献不大的项,降低文本特征空间维数,得到更为独立的特征空间,提高分类精度达到更好的分类效果. 它有主成分分析 (PCA)、潜在语义索引 (LSI)、非负矩阵分解 (NMF) 3种常用方法. 在此采用非负矩阵分解方法,它具有收敛速度快、存储空间小的特点,适用于处理大规模文本.

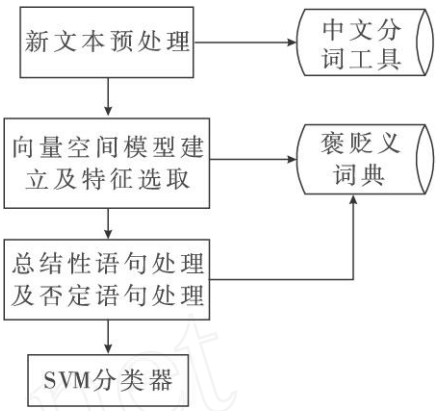

图 1　倾向性分类模型结构框图

Fig. 1　Frame chart for model of tendency classification

本算法主要步骤如下:

1) 打开已做过分词及词性标注处理的文本 T_i;

2) 在文本中查找词性标注符 adj和 n,将其对应的词语提取出来;构成特征项集合 t_i;

3) 对照褒贬义词典,将不在词典中的特征项剔除,形成新的特征项集合 t_i'

4) 根据公式 (1) 计算权重、建立矩阵向量空间模型;

5) 采用非负矩阵分解方法进行特征抽取,特征项集合变换为 D_i;

6) 将否定特征词加进来,$D_i + D_j$构成模型的完整特征项;

7) 根据公式 (1) 对 D_j部分计算权值,并将其标注在矩阵向量中;

8) 对于文档中包含否定词,且此否定词不在褒贬义词典中的情况,依照 2.2节步骤进行处理.

2.4 图书评价倾向性分类算法

模型构建过程中考虑中文语言习惯,总结性句子可以代表评论人的整体意见,如"总 "'一般来说 "等等,这些具有总结性意义的词语后面的句子可以直接反应出评论的褒贬倾向,因此直接分析其后面的各个特征项的倾向性可以代表整条评论的倾向性,可以有效缩短分类时间.

将评价语料中的 1/3部分作为训练集,提供给机器学习; 2/3部分作为分类集. 训练阶段算法的主要流程如下:

1)新文本预处理 (包括数据清洗、分词、词性标注等);

2)依据文中 3.3节提出的算法进行特征选择及特征抽取;

3)采用 SVM方法进行机器学习.

分类阶段算法的主要流程如下:

1)新文本预处理 (同上);

2)判断评论中是否存在总结性词语,直接提取总结性词语后面的名词及形容词作为特征输入;

3)对于不存在总结性词语的评论文本进行上述训练阶段的 (2)步骤;

4)采用 SVM方法判断各文档所属类别;

5)输出分类结果.

3 总 结

论文研究了文本分类技术和文本倾向性分类技术,将"含总结性词语的文本进行特别处理 "引入到模型中,将褒贬义词典中的部分含否定意义词并入特征项中,采用向量空间模型方法将分类问题转换成计算机可以处理的数学模式,采用非负矩阵分解方法对特征空间进行了降维处理,采用 VSM算法进行文本倾

向性分类,给出了一种图书评价系统的构建模型,并详细阐述了模型建立的方法和流程.

参考文献:

[1] 李智超,马少平.针对搜索引擎的媒体倾向性研究[J].江西师范大学学报,2008,32(2):127-131.
[2] 姚天昉,娄德成.汉语语句主题语义倾向性分析方法的研究[J].中文信息学报,2007,21(5):73-79.
[3] 李艳玲,戴冠中,朱烨行.基于类别空间模型的文本倾向性分类方法[J].计算机应用,2007,27(9):2 194-2 196.
[4] 李艳玲,戴冠中,覃森.快速的文本倾向性分类方法[J].电子科技大学学报,2007,36(6):1232-1236.
[5] 唐慧丰,谭松波,程学旗.基于监督学习的中文倾向性分类技术比较研究[J].中文信息学报,2007,21(6):88-94.
[6] 马海兵,刘永丹,王兰成,等.三种文档语义倾向性识别方法的分析与比较[J].现代图书情报技术,2007(4):43-47.
[7] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up Sentiment Classification Using Machine Learning Techniques[C]. In Proceedings of EMNLP 2002: 79-86.
[8] 罗欣,夏得麟,晏蒲柳.基于词频差异的特征选取及改进的TF-IDF公式[J].计算机应用,2005,25(9):2031-2033.
[9] Vapnik V. The Nature of Statistical Learning Theory[M]. New York; Sprfinger-Verlag, 1995.
[10] 徐军,丁宇新,王晓龙.使用机器学习方法进行新闻的倾向性自动分类[J].中文信息学报,2007,21(6):96-100.

(上接第116页)

4 结束语

1)笔者利用物元分析矩阵和理想点法,对多指标决策问题进行了研究,建立了基于物元矩阵的优化决策模型.该模型具有计算过程简单,使用起来方便等特点.特别在权重取值上采用拉格朗日函数来确定指标权重值.不但克服了权重的主观性,而且决策结果客观、公正.

2)由于笔者既考虑了信息的透明度原则,又避免了理想点法和物元分析模型取值中的一些主观性,所以该模型在实际应用上有一定的实用价值.

3)应用实例说明,该模型理论简捷、可操作性好,克服了决策中人为因素影响大的缺点,解决了不相容指标权重值分配难的问题,为多指标决策问题提供了一条新的途径.

参考文献:

[1] 钱钢.三种基于理想点的不确定多属性决策优化模型[J].系统工程与电子技术,2003,25(5):1-3.
[2] 孙晓东.基于灰色关联度和理想解法的决策方法研究[J].中国管理科学,2005,13(4):63-67.
[3] 刘家学,黄德成.无信息多指标决策的层次——关联优化模型[J].系统工程与电子技术,2000,22(12):7-10.
[4] OLSON D L. Comparison of Weights in TOPSIS Models[J]. Mathematical and Computer Modeling, 2004, 40: 82-85.
[5] 蔡文.可拓学概述[J].系统工程理论与应用,1994.
[6] 蔡文.物元模型极其应用[M].北京:科学技术文献出版社,2001.
[7] CHEN C T. Extensions of the TOPSIS for Group Decision Making Under Environment[J]. Fuzzy Sets and Systems, 2000, 114: 1-9.
[8] 罗党,刘思峰.灰色关联决策方法研究[J].中国管理科学,2005,13(1):101-106.
[9] 张吉军.权重为区间数的多指标决策问题的逼近理想点法[J].系统工程与电子技术,2002,24(11):76-78.
[10] 胡启洲,石琴,张卫华,等.城市公交线网优化的理想决策法[J].交通运输工程学报,2005,5(1):82-85.