基于词项关联的短文本分类研究

章 昉[1,2] 颜华驹[3] 刘明君[2] 赵中英[2]

[1](天津海量信息技术有限公司 天津 100029)
[2](中国科学院深圳先进技术研究院 深圳 518055)
[3](中山大学信息科学与技术学院 广州 510006)

摘 要 以短文本为主体的微博等社交媒体，因具备文本短、特征稀疏等特性，使得传统文本分类方法不能够高精度地对短文本进行分类。针对这一问题，文章提出了基于词项关联的短文本分类方法。首先对训练集进行强关联规则挖掘，将强关联规则加入到短文本的特征中，提高短文本特征密度，进而提高短文本分类精度。对比实验表明，该方法一定程度上减缓了短文本特征稀疏特点对分类结果的影响，提高了分类准确率、召回率和 F_1 值。

关键词 数据挖掘；短文本；分类；关联规则
中图分类号 TP 3 **文献标志码** A

The Research of Short Texts Classification Based on Association Rules of Lexical Items

ZHANG Fang[1,2] YAN Huaju[3] LIU Mingjun[2] ZHAO Zhongying[2]

[1](*Hylanda Information Technology Co., Ltd, Tianjin* 100029, *China*)
[2](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciencess, Shenzhen* 518055, *China*)
[3](*School of Information Science and Technology, Sun Yat-sen University, Guangzhou* 510006, *China*)

Abstract Due to its characteristics of shortness and sparseness, short text, as the main body of microblog and other social media, cannot be accurately classified by the traditional text classification methods. To solve this problem, a method of short text classification based on association rules of lexical items was proposed in this paper. Firstly, the training set based on the strong association rules was mined, and then the strong association rules was added to the features of short text so as to increase the feature density of short text, thereby to increase the accuracy of results of short text classification. Comparative experiments show that this method, to some extent, reduces the impact of sparseness of short text on the classification results, and it improves the classification accuracy, recall values and F_1 values.

Keywords data mining; short text; classification; association rules

1 引 言

近年来，随着互联网技术的日新月异，尤其是 Web2.0 技术的发展，Facebook、Twitter、MySpace、腾讯微博等社会化媒体不断出现，并日益成为人们制造信息、分享信息和传播信息的重要平台。相对于传统媒体，这些社会化媒体具有稳定性较高、传播较快和资源利用率高等优势，正逐渐取代传统媒体成为人们分享信息的重

收稿日期：2014-03-04 **修回日期**：2015-03-18
基金项目：深圳市知识创新计划基础研究项目(JCYJ20130401170306838)
作者简介：章昉(通讯作者)，硕士，研究方向为社会网络分析与挖掘，E-mail：zhangfang@hylanda.com；颜华驹，硕士研究生，研究方向为数据挖掘；刘明君，硕士研究生，研究方向为社会网络分析与挖掘；赵中英，博士，助理研究员，研究方向为社会网络分析与挖掘。

要平台。

随着微博的流行，中国互联网用户的参与度和活跃度呈现出爆炸式地增长，微博不仅成为了网民发布、共享、传播信息的平台，而且积累了大规模的网民行为数据。2012 年 5 月，新浪微博事业部副总经理芦义指出，新浪微博注册用户已超过 3 亿，其中有 60% 的活跃用户通过移动终端登录，用户平均每天发布超过 1 亿条微博内容。可见微博的数据量越来越大，因而对微博数据的挖掘具有可行性、创新性以及实用性，而对以上有关内容的研究亦已受到国内外学术界的广泛关注。

科学家们已经开始通过挖掘微博等社交网络数据来预测一系列和社会、经济、健康等相关的现象，如电影票房[1]、疾病传播[2]等。美国总统奥巴马在 2012 年美国总统竞选中的成功连任也离不开他身后庞大的社交网络数据挖掘团队[3]。

在我们的工作中，我们研究如何对如同微博的短文本进行精确的多类分类主要有以下三个原因：

(1) 微博等短文本具有篇幅短、特征少等特点，会给针对长文本的文本分类任务带来分类精度不高的困难。如何解决短文本的高精度分类是需要解决的实际问题。

(2) 丰富的短文本资源背后潜藏着巨大的商业潜能。研究人员可以对海量短文本数据进行挖掘，获取人们当前的兴趣热点，然后根据兴趣热点而制定相应的更准确的商业目标，比如根据用户的微博文本进行分类，获取微博用户的兴趣热点，从而为其定制个性化的广告推送，使得微博运营商、商家和用户三方都受益。

(3) 使用微博等短文本进行交互给人们的生活带来了方便，同时一定程度上也给社会的稳定带来了隐患，如垃圾短信、反动言论等非法信息也试图使用微博这样的短文本交互方式渗透到社会的各个角落。对短文本进行自动分类可以有效地对短文本进行监测和识别，并对其进行自动过滤，防止其贻害社会。

在本文中，我们提出了一种基于词项关联的短文本分类方法，其中第 2 部分介绍了现有的针对短文本分类的研究工作；第 3 部分概述了腾讯微博及其数据特征，给出了人工标注的类别及结果；第 4 部分给出了短文本分类的实现过程及其关键技术；第 5 部分给出了实验过程及结果分析；文章最后对本研究进行了总结并且提出了进一步的工作计划。

2 相关研究

文本分类是指文本分类器对待分类文本特征进行分析，进而将文本归类到预先设定的类别里的过程。很多研究学者对中文短文本分类进行了研究，但仍旧处于起步阶段。

Yan 等[4]提出了一种针对短文本分类的动态分类方法，用一个树状分类器来减轻短文本的稀疏特性和不平衡特性对分类结果产生的影响；在训练阶段，提出了动态适应策略。实验结果指出，与传统的分类器相比，其方法在针对短文本的分类中可以获得较高的分类准确率和召回率，但分类器的稳定性尚未得到较好的验证。

胡吉祥等[5]提出了针对短文本聚类的重复串方法，通过使用有意义的重复串抽取技术代替文本分词，使得由分词产生的词条数大幅减少、降低了特征空间的维度，进而达到缓解短文本固有的高维度问题和高系数问题。而其实验结果指出，性能参数 F-measure 比传统聚类方法提高了将近 40%，说明所提出方法有效可行。然而实现重复串抽取方法的复杂度很高，增加了短文本分类问题的难度。

滕少华等[6]提出了使用条件随机域(CRFs)解决短文本分类问题。首先将文本转换为一个待标注的序列，再使用训练集得到的 CRFs 模型对该序列进行标注。实验结果表明，相对于支持向量机(Support Vector Machine，SVM)，使用 CRFs

对短文本分类能够得到更高的正确率。然而实现 CRFs 方法的复杂度较高，增加了短文本分类问题的难度。

丁亚辉等[7]提出了基于领域词语本体的短文本分类方法。首先抽取领域高频词作为特征词，借助知网从语义方面将特征词扩展为概念和义元，然后通过计算不同概念所包含相同义元的信息量来衡量词的相似度，从而进行分类。实验表明，该方法在一定程度上弥补了短文本特征不足的缺点，且提高了准确率和召回率。

由此可以看出，以上研究成果中均存在一定的问题需要克服。基于此，本文提出了一种基于关联规则的短文本分类研究。本研究中，我们基于训练微博集挖掘高质量的关联规则，对微博短文本进行特征拓展，从而减轻了短文本的高特征稀疏问题对分类结果产生的影响，提升了短文本分类的性能。最后通过实验验证该方法的有效性。

3　数据准备及人工标注

3.1　腾讯微博

腾讯微博是一个国内微博网站，于 2010 年 4 月由腾讯控股有限公司推出。在国内，腾讯微博已是十分地受欢迎，有超过 5 亿的用户。和美国的推特(Twitter)一样，每个腾讯微博用户有一组听众(followers)，所以腾讯微博可以被视为一个社交网络。用户可以和其听众分享带有照片、视频以及 140 字以内的文字微博，而这些微博包含了关于用户的一些个人信息。用户发出的微博显示在用户的主页上，之后其听众便可以阅读、评论或者转发该条微博并显示在其个人主页上。除此之外，用户之间还可以直接相互发送私信。转播微博使得腾讯微博内的照片、视频、文本和链接等信息可以快速传播。由于腾讯微博庞大的用户群体，越来越多的公司和组织使用腾讯微博来推销产品或者传播信息。在我国，挖掘腾讯微博数据已经成为一个热门的、创新的方法来预测一些未来的社会现象或者判断潜在的消费和用户群体。

3.2　数据库特征

实验中所使用的数据通过腾讯微博搜索 API 从腾讯微博网站上下载而获得。2013 年 10 月 15 日至 10 月 20 日，通过 API 给出的接口对北京市、上海市、广州市和深圳市共 736 万多条腾讯微博进行下载收集。在上述微博集中随机选出 15000 条微博作为本实验的实验微博集，并将这 15000 条微博等分成三份，用于交叉验证本实验的有效性。

3.3　标记准则

经过市场调查，我们将微博文本分为 12 类，如表 1 所示。

13 个标记员负责对收集到的实验微博集进行标记，将实验微博集内的每条微博标记为上述 12 类中的一类。对于转发微博，如果评论部分可以判断该微博的类别，则直接判断；如评论部分不能直接判断该微博的类别，则结合原微博进行判断。根据鸽笼原理，每条微博都会有得票最多的类别，以此为该微博的最终类别。分类结果如表 2 所示。

表 2　实验微博集人工标记结果

Table 2　The result of artificial labels of Tencent Weibo sets

类别	子微博集 1	子微博集 2	子微博集 3
体育	4	18	2
健康	62	93	27
教育	2	74	10
旅游	8	31	1
科技	69	89	5
汽车	6	8	2
游戏	2002	1930	1966
美容美发美体	397	417	461
美食	67	68	35
服装鞋靴包	484	385	438
娱乐文化	235	125	95
其他	1664	1762	1958

表 1 微博文本分类

Table 1 Tencent Weibo text classification

类别	类别特征	微博示例
体育	体育赛事、体育报刊、体育明星等	一班和二班的篮球赛！
健康	健康常识、药物、身体健康状况等	【史上最全经期健康饮食法】女人面色红润、精神饱满比任何化妆品都给力。而要想始终精气神十足，就得注意经期保健。做好经期护理，特别是月经前后的饮食，也潜藏着大学问哦！
教育	新东方、新航道等培训机构，个人的学习状况、学习意向，出国留学	#2013 澳际秋季国际教育展“——西安站#名校提前看，贝尔法斯特女王大学[1]于 1845 年由维多利亚女王建校，其教研历史悠久，学术成就昭著。是英国历史最悠久的十所大学之一。它坐落于英国北爱尔兰首府贝尔法斯特市的南部，本市的空中交通方便，每天有 35 趟航班来往于伦敦空中飞行时@西安电子科技大学
旅游	景点、游乐场、出国游、自由行、酒店	11 月 30 日出发海南三亚！
科技	手机、电脑、数码产品、网络等	德阳移动三人消费 108 元以上，就可以送你 6 M 宽带，均可共享 108 话费，四人消费 138 以上就可以送 8 M 宽带，均可共享 138 话费五人消费 168 以上，就送 20 M，均可共享 168 话费，有需要的电话联系
汽车	汽车、汽车杂志等	【奔驰 S600 加长版继任迈巴赫广州车展亮相】想买车的人…可以参考一下 (分享自@Qzone)
游戏	手机游戏、网页游戏、网络游戏等	【炫耀一下！】种菜那件小事，贵在坚持，功到自然成。轻松升到 47 级了，更广阔的土地、更新奇的作物等着我开垦和收获~
美容美发美体	护肤品、化妆品、美甲、纤体、洗护用品等	长斑的人伤不起啊！再好的遮瑕膏也遮不住姐的一脸斑！最近用的老中医祛斑面膜，让姐激动的想哭~~一周就开始淡斑坚持用后越来越淡，直到现在脸上再也找不到斑痕的影子！闺蜜都以为我做了手术呢~~必须推荐给需要的妹纸们。
美食	食品、吃货、食谱等	黄记煌焖锅真好吃！
服装鞋靴包	服装、鞋靴、包、网购等	™夏季新款坡跟罗马凉鞋女鞋子拼色韩版高跟鞋厚底松糕鞋露趾鞋™
娱乐文化	娱乐圈、演唱会、话剧、展览等	#不二神探# 文章，这二货榛的不错，看戏劲头.....
其他	个人状态、个人情感、社会看法、生活看法等	今天真是太折磨人了，有种想死的感觉。

4 基于词项关联的短文本分类方法

本研究将使用传统分类器支持向量机对微博短文本进行分类。为了减轻短文本长度短、特征稀疏特征对分类结果产生的影响，我们挖掘关联规则对短文本特征进行扩充，从而提高传统分类器对短文本分类的效果。本文的微博短文本分类流程如图 1 所示。

首先，对微博文本进行去除特殊符号、分词和去除停用词的预处理，并去除微博中转发标识、表情标识和提及标志后的内容。然后对文本

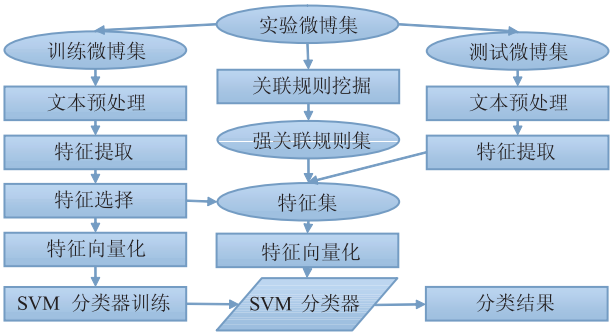


图 1 基于关联规则的微博短文本分类过程

Fig. 1 The process of short text classification based on association rules

特征进行特征选择。这样做主要有以下三个原因：

(1) 提高训练和测试过程的效率；

(2) 去除噪音；

(3) 提高分类精度。

我们计算训练微博集中经过预处理后的特征提出来的每一个词项的 CHI 卡方检验值，对词项集合按照 CHI 卡方检验值进行由大及小排序，选出最高的前 3000 个词项作为 SVM 分类器的特征，并使用 tf-idf 值对每条微博进行特征向量化处理。

以下将给出本方法中的两类关键技术：支持向量机和关联规则。

4.1　支持向量机

支持向量机[8-11]属于一般化线性分类器，是一种监督式学习的方法，被广泛地应用于统计分类以及回归分析。

4.1.1 二类线性可分条件下的支持向量机

如图 2 所示，二类线性可分问题存在大量可能的线性分界面。对于 SVM 而言，它的准则是寻找一个离数据点最远的决策面。从决策面到最近数据点的距离决定了分类器的间隔。这种构建方法也意味着 SVM 的决策函数完全由部分数据子集决定，并且这些子集定义了分界面的位置。这些子集的点被称为支持向量。在分类构建过程中，SVM

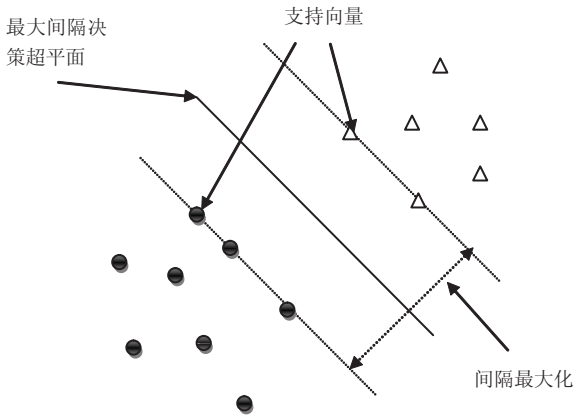


图 2　分类器间隔两端的 5 个点是支持向量

Fig. 2　The support vectors are the 5 points right up against the margin of the classifier

强调在分类决策面上下有一个大的分类间隔。

4.1.2 软间隔分类

对于在文本分类中很普遍的高维空间问题来说，有时数据是线性可分的。但是一般情况下这都不成立，而且即使线性可分成立，我们也可能优先考虑那些能够将大部分数据分开而忽略一些奇异噪音文档的解决方案。

如果训练数据集 D 线性可分，常规的做法是允许决策间隔间犯一些错误(有些离群点或者噪音点在间隔内部或者在决策面的错误一方)。于是，我们要根据每个错分例子满足间隔的程度定义其惩罚代价(Penalty)。为了实现这一目的，引入松弛变量 ξ_i，一个非零的 ξ_i 表示允许 x_i 在未满足间隔需求下的惩罚量或代价因子。如图 3 所示：

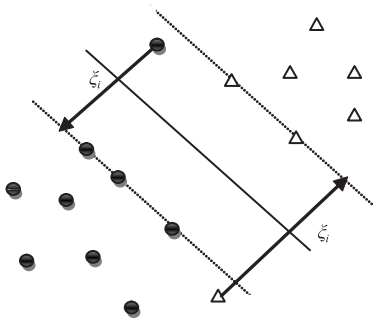


图 3　引入松弛变量的大间隔分类

Fig. 3　Large margin classification with slack variables

4.1.3 非线性支持向量机

如果数据集不允许线性分类器分类时怎么办？图 4 中上面的数据集显然可以被线性分类器直接分开，而中间的数据集却显然不可能被线性分类器直接分开。我们需要做的就是将他们间隔开。一个解决这个问题的方法是将数据映射到一个高维空间并在此空间上使用线性分类器将数据分开。例如，图 3 中最下面的图表明，如果采用二次函数将原始数据映射到二维空间，那么在新空间中就可以很容易将数据分开。也就是说，尽可能保留与数据相关性有关的特征维，将原始的特征空间映射到某个更高维的线性可分的特征空间中去。这样，最终的分类器仍然具有很好的泛化能力。

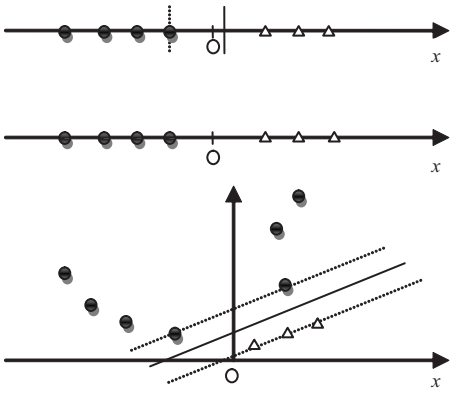


图 4 将非线性可分的数据映射到高维空间中使它们可分类

Fig. 4 Projecting data that is not linearly separable into a higher dimensional space can make it linearly separable

4.2 关联规则

大多数的微博都有着长度短和特征稀疏的短文本。如果潜在的相关联的特征能够被挖掘并加入到原文本中，使得短文本文本长度变长、特征更多，那么短文本的分类效果也会得到提升。而在数据挖掘领域中，关联规则挖掘[12-15]是一种流行的并被仔细研究过的在大型数据库中挖掘变量间联系的方法。鉴于以上理由，我们使用关联规则来提高对微博数据进行分类的效果。

Agrawal 等[12]将关联规则定义为，描述在一个交易中物品之间同时出现的规律的知识模式，更确切地说，关联规则是通过量化的数字描述物品 X 的出现对物品 Y 的出现有多大的影响。在我们的研究中，对关联规则如下定义：将 $I=\{i_1,i_2,\cdots,i_n\}$定义为 n 个文本特征的集合，$D=\{t_1,t_2,\cdots,t_m\}$数据库中的 m 个微博文本。在一个给定的数据库 D 中，一个关联规则如同 $A\Rightarrow B$，$A, B\in I$ 并且 $A\cap B=\varnothing$ 的形式。其中 A 和 B 分别叫做这个规则的先行词和导出词。判断一个关联规则是否为一个强关联规则的关键是计算这个规则的支持度和置信度，因而挖掘关联规则是获取强关联规则的关键。

Apriori 算法[13]可以被用来挖掘关联规则和频繁模式，因为 Apriori 算法需要找到所有候选项集并且在此过程中反复对数据库进行扫描，所以 Apriori 算法不是一个高效的算法。然而在我们的研究中，只需要找到有两个项的候选项集而不考虑多于两个项的候选项集，因此 Apriori 算法成为一种有效的并且能在我们研究中应用的算法。

支持度和置信度都达到最小阈值的频繁模式被看做是可以用来拓展微博短文本进而提高微博短文本分类精度的强关联规则。假设在我们的数据库中，“吃饭”⇒“睡觉”是一个强关联规则，那么词项“睡觉”会作为特征被添加到含有词项“吃饭”的微博文本词项集合中。

5 实验与分析

5.1 评价指标

评价文本分类器的常用指标主要包括分类准确率(Precision，简记为 P)、召回率(Recall，简记为 R)、F_1 测量值(简记为 F_1)、微平均(Micro)和宏平均(Macro)。下面将对这些常用指标进行简要描述。

5.1.1 准确率、召回率、F_1 测量值

某个文本分类器的分类结果如表 3 所示。其中，真正例(tp)表示实际属于该类且被分类器分到该类的文本数目；伪正例(fp)表示实际不属于该类但被分类器分到该类的文本数目；伪反例(fn)表示实际属于该类但未被分类器分到该类的文本数目；真反例(tn)表示实际不属于该类且未被分类器分到该类的文本数目。

表 3 某文本分类器的分类结果

Table 3 Result of a classifier

	文本属于该类	文本不属于该类
标记为该类	真正例（tp）	伪正例（fp）
未标记为该类	伪反例（fn）	真反例（tn）

准确率是指被分类器分到该类的文本中实际为该类的文本所占比例，用 P 表示：

$$P=tp/(tp+fp)$$

召回率是指实际属于该类的文本被分类器分为该类的文本所占比例，用 R 表示：

$$R=tp/(tp+fn)$$

通常我们希望文本分类器达到一定准确率的同时也希望能够同时达到一定的召回率，融合了准确率和召回率的指标是 F 值，指准确率和召回率的调和平均值：

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}，其中\beta^2=\frac{1-\alpha}{\alpha}$$

默认情况下，平衡 F 值中准确率和召回率的比重相同，即 $\alpha=0.5$ ，或记为 $\beta=1$，则公式简化为：

$$F_1=\frac{2PR}{P+R}$$

5.1.2 微平均和宏平均

当对多类分类器进行评价时，我们需要对所有类别的准确率和召回率综合评价，此时用到的评价方法便是微平均和宏平均。

微平均将所有类别的分类结果综合起来计算出一个总的准确率和召回率，计算微平均时需要计算 tp_{all}、fp_{all} 和 fn_{all}。其中，tp_{all} 表示在所有测试集文档中被正确分类的文档数目；fp_{all} 表示在所有测试集文档中被错误分类的文档数目；fn_{all} 表示在所有测试集文档中应正确分类却没有正确分类的文档数目。微平均法的计算公式如下：

$$micro-p=\frac{tp_{all}}{tp_{all}+fp_{all}}$$

$$micro-r=\frac{tp_{all}}{tp_{all}+fn_{all}}$$

宏平均则是在类别中求平均值，计算公式如下：

$$macro-p=(\sum_{i=1}^{k}P_i)/\ k，\ k\ 为类别数$$

$$macro-r=(\sum_{i=1}^{k}R_i)/\ k，\ k\ 为类别数$$

微平均和宏平均的计算结果可能会相差很大，微平均对每篇文档的判定结果等同对待，而宏平均对每个类别等同对待。微平均的计算中，大类起支配作用，需要度量小类的分类结果，则需要计算宏平均指标。

5.2　实验结果与对比分析

由于短文本的特征稀疏特性使得直接使用 SVM 分类器进行分类而达不到较好的分类效果，我们使用关联规则对微博短文进行词项拓展。基于实验微博集，我们挖掘到了一些支持度高于 0.002、置信度高于 0.6 的强关联规则，表 4 展示了其中八个强关联规则。

表 4　强关联规则示例及其支持度和置信度

Table 4　The samples of strong association rules with support and confidence

前向词	导出词	支持度	置信度
难受	感冒	0.0022	0.6111
托福	出国	0.0031	0.7419
游戏	玩	0.1991	0.6661
玩	游戏	0.1991	0.6581
痘痘	祛痘	0.0073	0.7365
祛痘	痘痘	0.0073	0.7842
吃货	好吃	0.0023	0.6140
折扣	返利	0.0059	0.6067

为了和我们的研究进行对比，我们首先进行了三次实验，每次实验分别以子微博集 1、2、3 为训练集，另外两个子微博集为测试集，每次实验中先使用 SVM 分类器直接分类，而后加入关联规则后再进行对比，实验结果如图 5 所示。图 5 针对单个类别进行评价，$D=1,2,3$ 分别表示子微博集 1、2、3；P、R、F_1 值分别为文类评价指标准确率、召回率和 F_1。表 5 对分类器的整体性能进行评价，使用微平均和宏平均方法对分类器使用关联规则前后进行性能比较。

从图 5 可以看出，实验一、实验二和实验三在使用关联规则后，各类的分类准确率和召回率大部分都呈现上升的趋势。其中升高十个百分点

以上的用粗体标出，而用斜标出的是指使用关联规则后评价标准呈下降趋势，并且集中在微博条数不多的类别中，比如体育、健康、教育等类别，分类性能下降的原因如下：

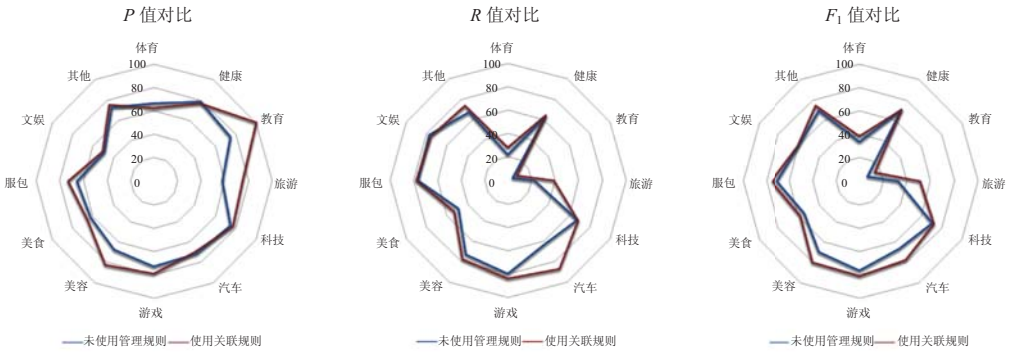
（1）训练集和测试集类别微博数目差异较大。如子微博集 1 里教育类的微博只有 2 条，而在子微博集 2 和子微博集 3 里，教育类微博分别有 74 条和 10 条，分别作为训练集时，导致文类效果相对差。

（2）加入关联词后引入了噪音使得分类结果错误。

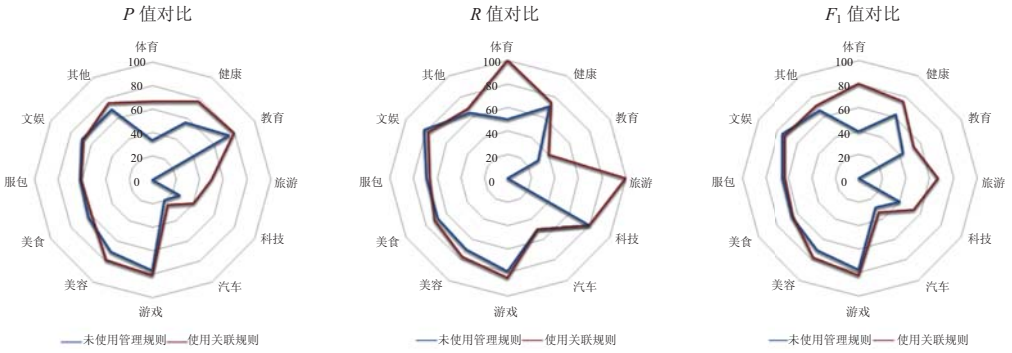
从表 5 可以得到以下结论：

（1）以子微博集 1 为训练集时，分类效果相对最差；以子微博集 3 为训练集时，分类效果相对最好。主要是由于子微博集 1 内的文本类别分布最不均匀，而子微博集 3 内的文本类别分布相对最均匀。

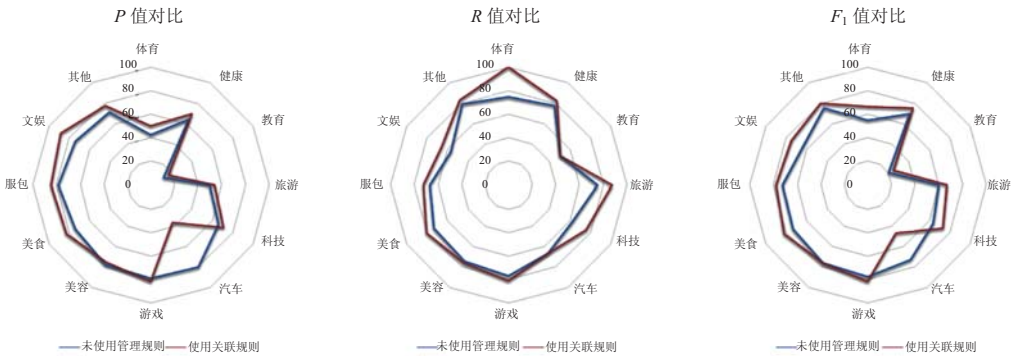
（2）三次实验中，使用关联规则后，微平均



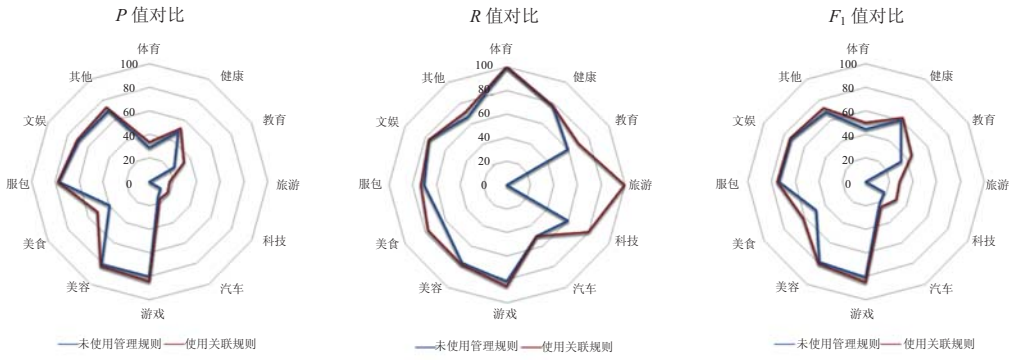
（a）实验一 D＝2



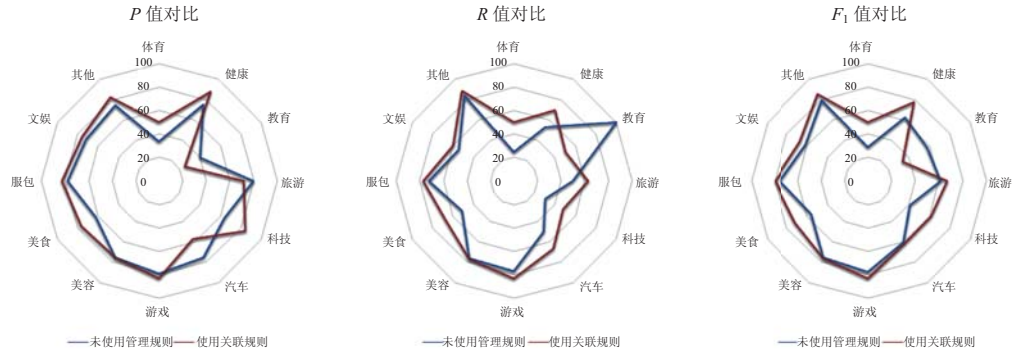
（b）实验一 D＝3



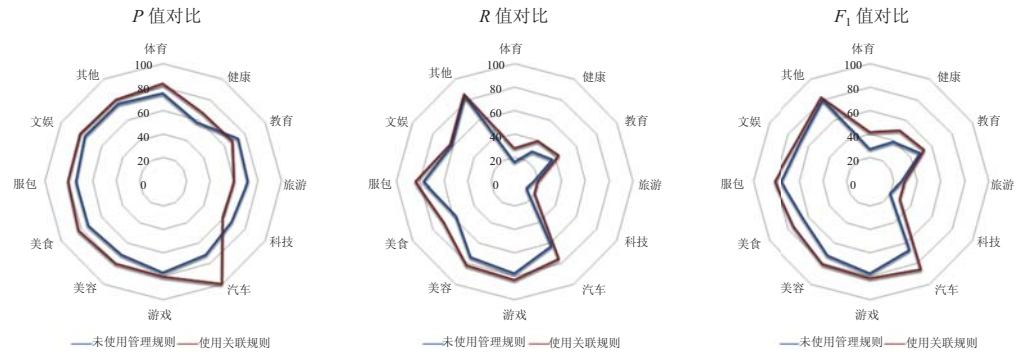
（c）实验二 D＝1



(d) 实验二　D＝3



(e) 实验三　D＝1



(f) 实验三　D＝2

图 5　使用关联规则前后分类效果比较-1

Fig. 5　Summary of evaluation results-1

表 5　使用关联规则前后分类效果比较-2

Table 5　Summary of evaluation results-2

	未使用关联规则					使用关联规则				
	Micro-p	Micro-r	Macro-p	Macro-r	Macro-f	Micro-p	Micro-r	Macro-p	Macro-r	Macro-f
实验一	71.41	71.41	59.81	57.39	54.85	75.78	75.78	68.68	67.73	64.57
实验二	74.69	74.69	54.75	67.34	60.11	79.78	79.78	58.30	77.49	63.14
实验三	76.04	76.04	69.60	55.30	58.36	80.82	80.82	73.90	62.03	65.60

准确率 Micro-p 平均增加 4.75 个百分点，宏平均准确率 Macro-p 平均增加 5.54 个百分点，宏平均召回率 Macro-r 平均增加 9.07 个百分点。这些评价标准的提高表明，使用关联规则后分类器的综合效果有较为明显的提高。

从实验可以看出使用关联规则后能够有效提高微博等短文本的分类精度，然而提升幅度有限。

6 结论与展望

本文提出了基于词项关联的短文本分类方法。该方法通过挖掘强关联规则，拓展微博短文本长度，增加微博短文本特征数，减轻短文本特征稀疏性对分类结果产生的影响，从而提高传统分类器对微博短文本分类的有效性。在真实的微博数据上进行的实验结果表明，短文本分类的准确率、召回率和 F_1 值都有一定程度的提高。然而，仅仅使用词项关联对短文本分类，还不能得到非常理想、有效的结果，我们将在后续的研究工作中不断探索和完善，如：建立主题词库，对每个分类中添加词项，微博短文本分类时判断是否含哪些词项，从而判断该短文和哪些类相关；或者对微博短文本建立上下文关系，微博中经常会有转发微博，判断转发微博和原微博之间的情感、逻辑关系，通过原微博来判断转发微博的类别。

参 考 文 献

[1] Sadilek A, Kautz HA, Silenzio V. Predicting disease transmission from geo-tagged micro-blog data [C] // Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012: 11.

[2] Asur S, Huberman BA. Predicting the future with social media [C] // 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, 1: 492-499.

[3] Tumasjan A, Sprenger TO, Sandner PG, et al. Predicting elections with twitter: What 140 characters reveal about political sentiment [C] // Proceedings of the Fourth International Conference on Weblogs and Social Media, 2010: 178-185.

[4] Yan R, Cao XB, Li K. Dynamic assembly classification algorithm for short text [J]. Acta Electronica Sinica, 2009, 37(5): 1019-1024.

[5] 胡吉祥, 许洪波, 刘悦, 等. 基于重复串的短文本聚类研究 [C] // 2005 全国第八届计算语言学联合学术会议论文集, 2005: 355-361.

[6] 腾少华. 基于 CRFs 的中文分词和短文本分类技术 [D]. 北京: 清华大学, 2009.

[7] 宁亚辉, 樊兴华, 吴渝. 基于领域词语本体的短文本分类 [J]. 计算机科学, 2009, 36(3): 142-145.

[8] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[9] Lin CJ. A practical guide to support vector machines classification [D]. Taipei: Taiwan University, 2006.

[10] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2008.

[11] Meyer D, Leisch F, Hornik K. The support vector machine under test [J]. Neurocomputing, 2003, 55(1): 169-186.

[12] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases [C] // Proceedings of the 1993 ACM SIGMOD International Conference on Manangement of Data, 1993: 207-216.

[13] Agrawal R, Srikant R. Fast algorithms for mining association rules in lager databases [C] // Proceedings of the 20th International Conference on Very Large Data Bases, 1994: 487-499.

[14] Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining--a general survey and comparison [J]. ACM SIGKDD Explorations Newsletter, 2000, 2(1): 58-64.

[15] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques [M]. Morgan Kaufmann, 2005.