

doi:10.3772/j.issn.1000-0135.2011.04.002

停用词表对基于 SVM 的中文文本情感分类的影响¹⁾

夏火松¹ 陶敏^{1,2} 王一^{1,2} 魏翔¹

(1. 武汉科技学院经济管理学院, 武汉 430073; 2. 武汉科技学院商务智能与数据挖掘实验室, 武汉 430073)

摘要 运用非结构化信息挖掘,对网络评论情感进行分析是一个非常重要的方法。本文基于 Web 客户评论情感文本,在情感文本预处理过程中使用四种不同的停用词表,采用两种不同的特征选择方法,选用著名的 TF-IDF 权重计算方法,使用基于 RBF 核函数的支持向量机方法的分类器实现了对携程网上采集的 4000 个酒店客户评论情感文本的分类研究。通过实验,分析了不同特征选择方和停用词表的使用对客户评论文本情感分类的影响,提出了基于情感文本分类的有效的停用词表。

关键词 客户评论 情感分类 停用词表 特征选择 支持向量机

The Influence of Stop Word Removal on the Chinese Text Sentiment
Classification Based on SVM Technology

Xia Huosong¹, Tao Min^{1,2}, Wang Yi^{1,2} and Wei Xiang¹

(1. Department of economics and management, Wuhan University of Science and Engineering, Wuhan 430073;

2. Business Intelligence and Data Mining Lab WUSE, Wuhan 430073)

Abstract It is an important method to analyse Web reviews' sentiment categorization with unstructured information date mining. This paper based on the Web text reviews, using four different kinds of stop word removal way, two kinds of feature selection methods, the famous TF-IDF weighing assignment methods and the SVM (support vector machine) technology with the RBF kernel function categorize the 4,000 customer reviews text grasp on XIECHENG. With the results of the experiment, this paper analysis the influence of different kinds of feature selection methods and stop word removal on the Chinese text sentiment classification, represent the more effective stop word removal list.

Keywords customer review, sentiment classification, stop words removal, feature selection, support vector machine

1 引言

近年来 Web2.0 技术的广泛应用,使互联网的使用和发展出现了巨大的飞跃,大量用户创造的内容成为互联网上重要的信息来源。面对海量信息,人们已经不能简单地靠人工来处理所有的信息,需

要辅助工具来帮助人们更好地发现、过滤和管理这些资源。搜索引擎的出现大大提高了互联网上信息搜寻的速度,当人们需要了解一些未知信息时,就可以利用搜索引擎的在互联网上进行搜索^[1]。未来的搜索引擎将可能提供一个全新的自动分析功能——情感分类(sentiment classification),通过对客户评论的情感分类,将帮助我们了解更多的用户对

收稿日期: 2010 年 1 月 18 日
作者简介: 夏火松,男,1964 年生,博士,教授,研究生导师,主要研究方向:知识管理、离群数据挖掘、信息管理和电子商务、DSS、电子商务。陶敏,女,1987 年生,硕士研究生,研究方向:文本挖掘,知识管理,商务智能。E-mail: taomin111@gmail.com。王一,男,1989 年生,研究方向:金融市场,信息管理。魏翔,男,1987 年生,研究方向:数据挖掘,信息管理。
1) 基金项目:国家社会科学基金(07BTQ010)、湖北省课题(Z20091701,2008d062,2008244,2007097,HB092-21)、武汉市课题(200940833384-02,20041007072-08)和中国纺织工业协会(2007082)支持)。

某种商品的态度倾向的分布,从而做出正确的购买决策。而且企业能够通过客户的评论信息作出产品的改进^[2]。

文本的情感分类研究是挖掘并利用包含在文本中大量的立场、观点、看法、情绪、好恶等隐含的主观信息,正确地分词和提取情感特征词汇能够提高分类的准确性。传统的文本分类通常是基于主题的分类,停用词表对分类结果有较大的影响^[3]。Silva 提出使用停用词表降低特征空间的维数,对提高文本分类器的准确率会产生积极的作用^[4]。对于停用词表对中文情感文本的分类的影响,国内学者王素格等采用 IG、MI、 X^2 三种特征选择方法、选用布尔型权重和词频型权重两种权重计算方法,分析了选用不同停用词表对情感文本分类的影响^[5]。对于基于较为有效的 TF-IDF 型权重计算方法条件下,考虑不同特征选择方法以及不同停用词表对情感文本分类结果的影响并没有学者做出相关报道。本文选择特征频度(TF)和 X^2 统计量两种不同的特征选择方法和较为有效的 TF-IDF 权重计算方法,基于支持向量机(SVM)的方法来构建分类器,考虑使用不同的停用词表情感文本,实现对情感文本的分类研究。

2 情感文本预处理关键技术

在对中文情感文本进行分类的过程中要对文本进行预处理,抽取代表文本特征的元数据(特征项),过程主要包括:中文分词、特征降维、文本的向量表示,经过预处理的数据才能利用于分类器,进行情感分类。我们总结其基本过程如图1所示。

由于基于主观的词汇的大量出现,中文情感文本的预处理同基于主题的文本分类的预处理要求不同,本文拟采用特征频度(TF)和 X^2 统计量(X^2)两种特征选择方法实现特征降维,文本的特征表示过程中,特征项的权值采用 TF-IDF 型权重计算方法。

2.1 特征选择方法

特征选择即是从特征集 $T = \{t_1, t_2, \dots, t_s\}$ 中选择一个真子集 $T' = \{t_1, t_2, \dots, t_{s'}\}$, 其中, s 为原始特征集的大小, s' 为选择后的特征集大小。在特征选择时一般是利用某种评价函数,独立地对每个原始特征项进行评分,然后按分值的高低将它们进行排序,从中选择若干个最高的特征项,以达到减少特征维数的目的。本文采用特征频度(TF)和 X^2 统计量(X^2)这两种常见的特征选择方法。分别计算候选特征的度量值,然后根据事先设定的阈值筛选出有效候选特征。下面为本文试验所选用的特征选择方法。

1) 文档频度(Document Frequency, DF)

DF 是训练集中含有特征项 t_k 的文本数在总文本数中出现的概率。其理论假设为稀有特征项或者对分类作用不大,或者是噪声,可以被删除。在机器学习时,统计样本文档中所有特征项的 DF 值。若某特征项在某类文档中的 DF 值越高即表明此特征项在该类文档中的代表性越强,反之则代表性越弱。在特征选择过程中,设定阈值,保留 DF 值高于此阈值的特征项作为候选特征。

2) X^2 统计

在统计学中, X^2 统计用于表征两个变量间的相关性,同时考虑了特征存在与不存在时的情况。在此是指特征项 t 对类 c 的独立缺乏程度,缺乏程度越高,则表示特征项 t 对于类 c 越重要。对于特征项 t , 其 X^2 统计值如公式(1)所示:

$$x^2(t, c) = \frac{[P(t, c)P(\bar{t}, \bar{c}) - P(\bar{t}, c)P(t, \bar{c})]^2}{P(c)P(t)P(\bar{c})P(\bar{t})} \quad (1)$$

面对多类问题时,对特征项的 X^2 统计量处理可以用取平均或者取最大值两种方法处理,在实验中,本文采用取平均值:

$$x_{avg}^2(t) = \sum_{i=1}^n p(C_i) x^2(t, C_i) \quad (2)$$

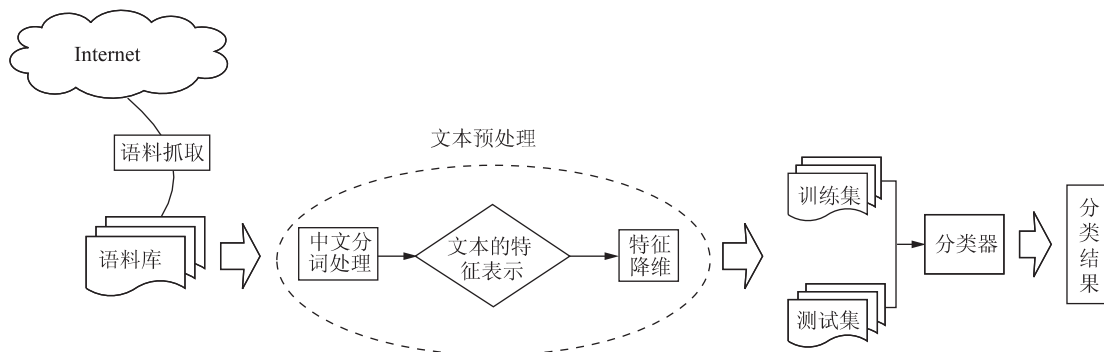


图1 中文文本分类的一般过程

2.2 特征权值的计算

特征项的权值综合反映了该特征项对标识文本内容的贡献度和文本之间的区分能力。常用的特征项权值计算有如下几种:布尔函数、平方根函数、对数函数、TF-IDF 函数。本文采用著名的权值函数 TF-IDF,其计算原理如公式(3)所示:

$$W_{ij} = freq_{ij} * \log(N/n_j) \tag{3}$$

其中, N 为所有文本的数目, n_j 为含有词条 t_j 的文本数目, $freq_{ij}$ 为 t_j 在 d_i 中出现的频率。在计算特征权值的时候,为了避免因文档长度引起的权值变化,应采用公式(4)对特征项权值评价函数进行归一化处理,将各特征项权值规范到 $[0,1]$ 区间中。

$$W_{ij} = \frac{\log(TF_{ij} + 1.0) * \log(N/n_j)}{\sqrt{\sum_k [\log(TF_{ik} + 1.0) * \log(N/n_k)]^2}} \tag{4}$$

其中, W_{ij} 为词 t_j 在文本 d_j 中的权值, TF_{ij} 为词 t_j 在文本 d_j 中的词频, $\log(N/n_j)$ 为文档频度,是训练集中含有特征项 t_j 的文本数在总文本数中出现的概率, N 为训练文本的总数, n_j 为包含 t_j 的文档数。

3 情感文本分类实验及评价指标

论文实验语料库来自于网上谭松波的酒店评论平衡语料^[6],语料库的总体统计特征如下:语料规模为 4000 篇,其中正面、反面语料各占 2000 篇,经分词和去标点处理后统计语料中包含 11 872 词,语料从携程网上自动采集,并经过整理而成。

3.1 停用词表的选择

在文本的向量表示过程中,要将文本数据表示成特征向量的形式。这其中的许多特征维对将要进行的分类学习未必全是重要的,而且高维的特征可能会大大增加其学习时间,也给分类过程带来了大量的噪声。去除停用词是在进行传统文本分类预处理过程中对词特征维去除噪声词的一个过程。中文文本的预处理过程对停用词的处理如图 2 所示。

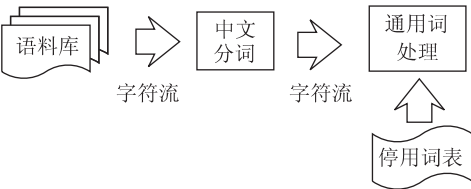


图 2 预处理过程中和停用词处理的流程

传统的停用词表是对主题没有描述能力和区别能力的词,是一些噪声词。而对情感分类来说,特征的选取就是要选择既要带有情感色彩又要有区分能力的词^[5]。在英文中,人们首先选择名词、动词、形容词、副词作为候选特征^[7~9],而在中文中,具有情感色彩的词除了名词、动词、形容词、副词外,还包括,区别词、叹词、拟声词、代词、成语、简称等具有情感色彩的词^[10]。为测试不同停用词表对情感文本分类的影响,我们构造了不同的停用词表,如表 1 所示。

表 1 不同停用词表的特征

| 停用词表 | 词表特征 |
|------|--|
| 1 | 不含名词(n)、形容词(a)的停用词表,即将不含名词(n)、形容词(a)作为候选特征 |
| 2 | 不含名词(n)、动词(v)、形容词(a)、副词(d)的停用词表,即将名词(n)、动词(v)、形容词(a)、副词(d)作为候选特征 |
| 3 | 不含名词(n)、动词(v)、形容词(a)、副词(d)、区别词(b)、代词(r)、叹词(e)、方位词(f)的停用词表,即将名词(n)、动词(v)、形容词(a)、副词(d)、区别词(b)、代词(r)、叹词(e)、方位词(f)作为候选特征 |
| 4 | 包含非语素字(x)、标点(w)和时间(t)的停用词表,即选取除非语素字(x)、标点(w)和时间(t)外的所有词作为候选特征 |

考虑到李荣陆^[11]的基于主题的停用词表,为考察不同词性的词在情感文本分类中的特殊作用,我们选择以上四种停用词表。

3.2 文本分类器的选择

支持向量机算法应用到文本分类时取得了较好的效果。Joachims 将 SVM 用于文本分类,实验在 Reuters 和 Ohsumed 两个标准语料库上进行,实验得出,在与贝叶斯、Rocchio、K 近邻算法和决策树这四种分类方法的比较中,SVM 方法不仅取得了更好地分类效果,还表现出了更强的鲁棒性和处理高维数据的优良特性^[12]。Dumais 和 Yang 等很多学者也相继对此进行了研究,并得出了相类似的结果。基于 SVM 方法应用于高维文本数据分类时良好的表现,本论文选择使用 SVM 方法实现分类器的构建。

实验选用基于支持向量机(SVM)的方法构建的分类器,所有的SVM均使用RBF核函数,通过交叉验证(cross validation)的方法找到最佳 c 、 g 值,构造最优SVM分类器模型,使用台湾国立大学林智仁教授的LIBSVM-2.89进行分类实验^[13],采用50%作为训练集,50%作为测试集,训练集和测试集不重复。

4 情感分类试验评价及结果分析

4.1 评价指标

评价指标是在测试过程中所使用的一些用来评价分类性能的量化指标,通常采用的性能评价指标有查全率(Recall,简记为 r)、查准率(Precision,简记为 p),其定义如下:

$$\text{反面查全率 RN} = \frac{a_1}{c_1} \text{反面查准率 PN} = \frac{a_1}{b_1}$$
$$\text{正面查全率 RP} = \frac{a_2}{c_2} \text{正面查准率 PP} = \frac{a_2}{b_2}$$
$$\text{正反查全率 } F_1 = \frac{a_1 + a_2}{c_1 + c_2} \text{正反查准率}$$
$$F_2 = \frac{a_1 + a_2}{b_1 + b_2}$$

其中, a_1 表示系统判断为反面的文档与实际应为反面的文档的文集的数量, a_2 表示系统判断为正面的文档与实际应为正面的文档的交集的数量; c_1 表示实际应为反面的文档数量, c_2 表示实际应为正面的文档数量; b_1 表示系统判断为反面的文档数量, b_2 表示系统判断为正面的文档的数量;由于本文判断的是正反两类问题,因此 $c_1 + c_2 = b_1 + b_2$,即 $F_1 = F_2$,这样混合两种类别的查全率和查准率应为相等。

4.2 实验结果及分析

在实验中,中文的分词及词性标注采用中科院的ICTCLAS系统,文本的向量表示模型采用向量空间模型,采用TF-IDF值作为特征项权重,文本预处理在MATLAB7.0.3平台下编程实现,选取的特征维数为4000维,实验结果如表2所示。

从表2和表3中可以看出:

(1) 特征选择方法方面,使用 X^2 统计量的特征选择方法的分类精度的差异梯度较使用TF频度的特征选择方法的分类精度的差异梯度要小,这说明, X^2 统计量作为特征选择方法具有较好的稳定性,这个结果基本与其他实验者的认识是一致的^[14]。

表2 不同停用词表以及不同特征选择方法所得的情感分类结果比较

| 停用词表 | 评价指标 | TF | | X^2 统计量 | |
|------|-------|--------|--------|-----------|--------|
| | | 查全率 | 查准率 | 查全率 | 查准率 |
| 1 | 正面 | 70.80% | 76.54% | 74.40% | 78.32% |
| | 反面 | 78.30% | 72.84% | 79.40% | 75.62% |
| | F_1 | 74.55% | 76.90% | | |
| 2 | 正面 | 75.70% | 80.62% | 76.80% | 81.61% |
| | 反面 | 81.80% | 77.10% | 82.70% | 78.09% |
| | F_1 | 78.75% | 79.75% | | |
| 3 | 正面 | 82.70% | 78.76% | 82.90% | 77.88% |
| | 反面 | 77.70% | 81.79% | 77.80% | 81.98% |
| | F_1 | 80.20% | 80.35% | | |
| 4 | 正面 | 78.00% | 83.96% | 82.30% | 80.69% |
| | 反面 | 85.10% | 79.46% | 80.30% | 81.94% |
| | F_1 | 81.55% | 81.30% | | |

表 3 基于停用词表 4 的不同特征选择方法
所得到的前 30 个词特征

| 特征选择方法 | 前 30 个特征 |
|--------|---|
| TF | 的/u、是/v、不/d、酒店/n、很/d、我/r、房间/n、了/u、有/v、住/v、了/y、还/d、也/d、在/p、就/d、好/a、服务/v、都/d、到/v、我们/r、房/n、说/v、去/v、要/v、不错/a、大/a、太/d、人/n、差/a、可以/v |
| X^2 | 是/v、不/d、房间/n、酒店/n、很/d、住/v、不错/a、还/d、差/a、有/v、也/d、好/a、服务/v、我/r、就/d、没有/v、说/v、要/v、都/d、携程/n、太/d、才/d、方便/a、人/n、到/v、再/d、比较/d、脏/a、能/v、房/n |

使用 X^2 统计量的特征选择方法在特征项的选择上显得更为有效,这从表 3 所列举的特征项可以看出。表 3 所列举的 30 个词特征为实验选取的 4000 个词特征中的前 30 个。在这些词特征中,包含多个共有词,但共有词所处的位置并不同。在相异词中,不具备明显情感区分度的词,如“的”、“我们”等在 X^2 统计量方法下已经从前 30 特征项中退出,事实上,特征词“我们”已经下降到 186 位;而具有更明显的情感区分度的词,如“脏”、“方便”等则出现在前 30 特征项中。同时,共有词中具有显著情感区分度“不错”、“差”、“好”等也从较靠后的位置上升到较靠前的位置。这些变化同样发生在实验所选取的 4000 维特征中,更多的具有情感分类区分度的特征词项在 X^2 统计量的特征选择中较 TF 方法下的位置有所上升,同时不具备明显情感分类特征的特征词项排位有所下降,这更好保证了实验所取的前 4000 维特征项能更好的表征情感文本,使情感文本分类取得较好效果。这也保证了,使用 X^2 统计量选择的特征比通过 TF 选择的特征具有更高的情感分类区分度。

(2)停用词表的选择方面,使用停用词表 1 到 4 所得到的最终分类精度呈上升状态,即使用停用词表 1 进行的情感文本分类的分类精度最低,停用词表 2、表 3 呈上升状态,使用停用词表 4 所得到的情感文本分类的精度最高。我们发现,在情感文本中,停用词表的使用情况对分类精度的影响不同于主题文本分类,停用词表所保留的候选特征越多,对应的情感文本分类精度越高。本实验中,停用词表 4 表现了最优的有效性。

5 结束语

本文基于 TF-IDF 特征权值计算方法,通过在大规模的酒店客户评论文本上的实验,采用两种不同的特征选择方法(X^2 统计量和特征频度(TF)以及不同的停用词表的使用);选取较多候选特征以及选取传统文本分类中的几类候选特征词性,实现不同特征选择方法和停用词表对情感文本分类的影响研究。实验得出,使用停用词表 4,即保留除标点、非语素词和时间外的所有词性的候选特征时情感分类效果较好。这是由情感文本分类不同于传统文本分类的特点决定的,在情感文本中区别词(f)、叹词(e)、拟声词(o)、代词(r)等词同其他词的组合往往具有较强情感倾向。 X^2 统计量的稳定性以及良好的区分情感特征的性能在本实验中也得到了验证。

虽然这篇论文在其他学者研究的基础上实现的分类的数值效果略有提高,但仍有一定的局限性,我们这此进行分析并提出未来的研究方向:①对语料集分类标记进行修正,以期得到更精确的实验结果。由于语料库分类标记的准确度直接影响着最终的分类结果,邀请领域专家,建立一个统一的标准进行类别的人工修正,会对分类结果产生积极的影响。②考虑使用基于语义的分词方法、实验更多的特征选择方法进行进一步的情感分类实验。初步小样本实验基于语义的分词方法在分词结果上更为准确,有益于提高分类的数值结果。同时选择更多的特征选择方法进行实验,从中选择最有效的特征选择方法,构建更高效的情感文本分类模型。③对情感倾向程度的类别区分。建立一个统一的标准,实现对情感倾向分类类别的扩充,拟扩充为强正向、次强正向、中立、次强负向、强负向五个情感倾向等级。在文本实验基础上,实现情感文本多类情感分类。

参 考 文 献

[1] 叶强,张紫琼,罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 信息系统学报, 2007,1(1):79-91.

[2] 余传明. 从产品评论中挖掘观点:原理与算法分析[J]. 信息系统,2009,32(7):124-128.

[3] 顾益军,樊孝忠,王建华,等. 中文停用词表的自动选取[J]. 北京理工大学学报,2005,25(4):337-340.

[4] Silva C, Ribeiro B. The importance of stop word removal on recall values in text categorization [J]. Neural

- Networks, 2003,3:20-24.
- [5] 王素格,魏英杰. 停用词表对中文文本情感分类的影响[J],情报学报,2008,27(2):175-179.
- [6] 谭松波. 中文情感挖掘语料-ChnSentiCorp[OL]. [2009-11-10]. [http://www. searchforum. org. cn/tansongbo/corpus-senti. htm](http://www.searchforum.org.cn/tansongbo/corpus-senti.htm) .
- [7] Pang B, Lee L, Vaithyanathan S. Thumbs up Setiment classification using machine learning techniques[C]. The Conference on Empirical Methods in Natural Language Processing,2002:79-86.
- [8] Turney P D, Littman M L. Measuring praise and criticism: inference of semantic orientation form association [J]. ACM Transaction on Information Systems,2003,21(4):315-346.
- [9] Vasileios H, Kathleen R M. Predicting the semantic orientation of adjectives[C]. ACL/EACL,1997:174-181.
- [10] 王治敏,朱学峰,俞士汶. 基于现代汉语语法信息词典的词语情感评价研究[J]. Computational Linguistics and Chinese Language,2005,10(4):581-592.
- [11] 李荣陆. 文本分类若干关键技术研究[D]. 上海:复旦大学,2005.
- [12] Joachims T. Text categorization with support vector machines; Learning with many relevant features. LS-8 Report 23, Computer Science Department, University of Dortmund,1998.
- [13] Chang C C, Lin C J. LIBSVM:a library for support vector machines[OL]. [2009-11-01]. [http://www. csie. ntu. edu. tw/~ cjlin/libsvm](http://www.csie.ntu.edu.tw/~cjlin/libsvm).
- [14] 尹世群. Web 文本分类关键技术研究[D]. 重庆:西南大学,2008.

(责任编辑 王建平)

