基于情感文本挖掘和分析的系统设计

余伟江[1] 李 恩[1] 章育林[1] 温志盛[1] 吴 健[1] 郑星锋[1]
([1]华南师范大学 物理与电信工程学院 广东 广州 510006)

摘 要:如何对网络中大量的文本数据进行挖掘和分析是大数据应用一个热点的问题,本文提供一种对文本数据进行挖掘和分析的新思路。以汽车口碑的文本数据为例,将采集的数据存入SQL SERVER 2008数据库,采用自然语言处理的方法处理数据,结合最大熵算法和支持向量机(Support Vector Machine, SVM)算法对数据进一步挖掘和分析。

关键词:文本分析;数据挖掘;汽车大数据;SVM

一、研究背景

情感文本挖掘和分析是自然语言处理中的一个研究领域[1]。如何有效地挖掘网络情感文本中的数据,是当今网络舆情分析所面临的关键问题。[2]本文借鉴现有的研究成果,提出一种基于最大熵算法结合SVM的文本情感分析新思路,设计出一个基于情感文本挖掘和分析的系统。

二、基于情感文本挖掘和分析的系统设计

(一)数据的采集

本系统使用基于WebCollector网络爬虫对汽车口碑进行爬取并将数据储存到SQL SERVER 2008数据库。

(二)数据的预处理

本系统创新地运用了HashSet类来存储不重复的对象[3];采用基于ANSJ的分词算法进行中文分词;使用基于哈工大停用词表的改进型停用词表进行停用词过滤操作。

(三)特征词的提取

针对"知网情感词典"和"台湾大学简体中文极性词典NTUSD"合并后的词典,我们通过人工添加新词的方法构建更合理的情感词典,提取评论的特征词。

(四)文本向量化

为了使计算机处理文本数据,我们需要将数据进行向量化。本文使用了著名的权值计算方法——词频-逆向文档频率 (term frequency - inverse document frequency , TF-IDF[4]) 实现汽车口碑的向量化。TF-IDF是一种统计方法,用以评估特征词对于汽车口碑中情感倾向的重要程度。

TFIDF的主要思想是:如果某个词或短语在一篇文章中出现的频率TF高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。特征词的重要性随著它在文件中出现的次数成正比增加,但同时会随著它在语料库中出现的频率成反比下降。

(五)情感分析

1基于最大熵算法的情感分类

这里我们使用的是Softmax回归模型,逻辑回归(Softmax)是最大熵对应类别为两类时的特殊情况[5]。在Softmax回归中,类型标记y可以取k个不同的值。于是,对于我们的训练集便有。首先计算Softmax回归概率值,其中是模型的参数。这一项对概率分布进行归一化,所有概率之和为1。然后添加一个权重衰减项来修改原代价函数,让参数值保持比较小的状态,这个衰减项会惩罚过大的参数值,得到新的代价函数,利用求偏导数,求最小化,从而实现一个可用的Softmax回归模型。

2基于SVM的情感细粒度分析

假设存在训练样本,可以被某个超平面没有差错地分开,其中,m为样本个数,为n维实数空间,是分类间隔。因此和两类最近的样本点距离最大的分类超平面称为最优超平面。在条件下对求解一下最大的函数值,为拉格朗日乘子,再根据公式求解最优分类函数,是偏移量,是是共轭表达。从而得到SVM分类器[6]。

三、结果分析

本文对网上7种车型的口碑进行爬取,利用最大熵算法的Softmax分类器进行情感倾向分类得到结果如下。

从图1可知购车者的汽车口碑的好坏评价比例, 用户对逸轩的认可度相对比较高,正向的口碑在7种热卖的汽车中最高,负向评论的数据最少。从上述的分类系统中,我们可以比较直观的得到哪一类汽车相对符合大部分人的需求并推荐给其他购车者,同时也可以将信息反馈给车商,帮助他们更好地改进汽车制作工艺。

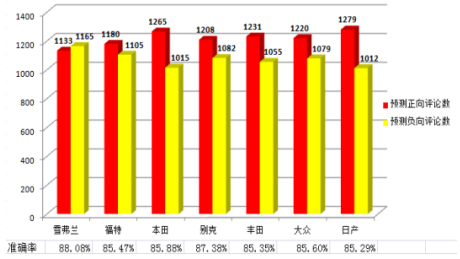

图1 Softmax模型的情感二分类效果

对一种汽车中的汽车属性进行细粒度分析,其可视化结果如图2所示。

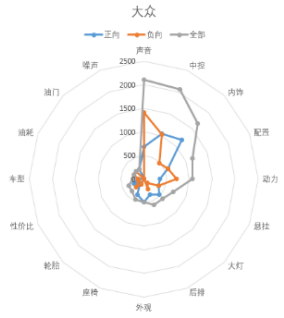

图2 对大众汽车的细粒度分析结果

细粒度分析可以人们对得到一类事物如汽车的各种属性的情感值,相对与综合情感倾向,有了更加细腻的倾向性,有利于更有方向的推荐。

四、总结

本系统将网络爬虫、文本数据预处理、特征词向量化结合最大熵算法和SVM,设计一个新的基于文本情感数据的分析系统,有良好的效果,希望可以对数据挖掘和分析领域有一定的参考价值。■

参考文献

[1] 涂慧明. 文本观点挖掘和情感分析的研究[J]. 电脑知识与技术,2016,05:235-237.

[2] 冯时. 面向网络舆情分析的观点挖掘关键技术研究[D].东北大学,2011.

[3] 王小华,卢小康. 基于N-Gram的文本去重方法研究[J]. 杭州电子科技大学学报,2010,02:61-64.

[4] 张建娥. 基于TFIDF和词语关联度的中文关键词提取方法[J]. 情报科学,2012,10:1542-1544+1555.

[5] 李学相. 改进的最大熵权值算法在文本分类中的应用 [J]. 计算机科学,2012,06:210-212.

[6] 王文华,朱艳辉,徐叶强,杜锐,鲁琳,邓程.基于SVM的产品评论属性特征的情感倾向分析[J].湖南工业大学学报,2012,26(5).

作者简介:余伟江(1994年),男,汉族,广东汕头人,华南师范大学物理与电信工程学院,2013级本科生,通信工程专业。

(接上页)

代农业科技;2008年06期.

[3] 吴向伟.转变农业经济发展方式的内涵与途径[J].经济纵横.2008(2).

[4] 孙剑,李崇光.论农产品营销渠道创新与对策[J].商业时代,2003(14).

[5] 张耀华.都匀市农村土地承包经营权流转现状与对策研究[D].贵州民族大学,2013.

作者简介:王立(1995)女,苗族,贵州都匀人,安徽财经大学经济学院,2014级本科生,经济学专业。