基于朴素贝叶斯分类算法的微博文本的情感分析研究

□王彬菁　成都师范学院

【摘要】 随着移动互联技术的发展，微博作为一种新媒体形式日益成为国内主流的移动社交媒体平台，微博使用的人群数量庞大，微博平台包含的内容丰富，网络社交的功能突出。微博包含海量的信息数据且数据种类多样，即有文档文本数据，也有图片、表情符号、视频动画等非结构化的数据。因此，对各政府部门和企业单位的网络舆情监管提出了艰巨的挑战，有关中文微博文本的情感分析的研究也成为近几年数据挖掘领域的关注方向之一，情感分析研究主要围绕着信息的抽取和情感倾向的判定，均离不开对微博文本的分词工作。本文提出一种基于朴素贝叶斯分类算法的分词归类方法，对比 PMI 互信息和特征频度 TF 方法的优劣，为微博文本中的词汇进行归类，分为喜、怒、悲、惊、乐五大类，从而分析文本的情感倾向。

【关键词】 微博文本　朴素贝叶斯算法　情感分析　数据挖掘

引言

近些年，随着移动互联技术的迅猛发展和日益成熟，移动互联技术已然进入社会大众的生活，并且逐渐改变着我们的消费方式、沟通交往方式；其中，微博作为一种成熟的新媒体形式已经成为国内最大的移动社交媒体平台。根据中国互联网络信息中心（CNNIC）最新发布的第 41 次《中国互联网发展情况统计报告》显示，截止 2017 年 12 月底，中国网民规模已经达到 7.72 亿，这其中手机用户的占比为 97.5%，手机成为网民上网的主要终端设备 [1]。这些网民获得信息的方式又主要通过微博，微信，各类手机 APP, 移动社会化的传播格局逐步形成，微博作为承载信息发布，互动交流功能的社交媒体平台已经被社会大众所熟知和使用。据《2017 年微博用户发展报告》显示，截止 2017 年 9 月，微博月活跃人数共计 3.97 亿，日活跃 1.65 亿，[1] 用户的使用习惯趋向移动化，微博讨论方式碎片化，强调高社交粘性的互动方式，这些特性吸引着年轻群体，他们在微博上表达带有个人喜好的观点和看法，对网络舆论的传播具有重大影响。所以，微博应该成为各级政府机构和企业关注的舆论阵地，积极引导正面舆论，及时监控不良的舆论导向。

微博文本主要使用文本形式传播信息，其中也包含其他非结构化的数据，比如种类繁多的网络表情符号、各式各样的图片、视频、音频。这些都为文本词汇信息的提取增加了难度。微博平台提供的 API 可以方便微博语料的获取。另外，谷歌公司开发的 Word2vec 也可以将微博文本快速转化为计算机可以识别的数据，作为一种机器学习方法，他可以在深度学习算法应用以前对语料进行预处理，将语料自动加载到模型中，通过设定相关参数，模型算法会将其训练成对应的词向量，通常使用在文本词性分析、聚类和查找同义词等方面，为微博文本的情感分析提供了便捷的处理手段。[2] 通过查阅文献可知，关于微博文本的情感分析的研究已经成为近几年数据挖掘领域的主要研究方向。目前，情感分析研究主要围绕着信息的抽取和情感倾向的判定，完成这两项工作必须对微博文本中的数据信息进行预处理，包括分词处理；网络表情符号识；词汇的情感分类汇聚以及情感判定等，为了更好判定微博文本的情感倾向，文章提出一种基于朴素贝叶斯分类算法的分词归类方法对微博文本进行情感分析研究。

一、数据挖掘技术

数据挖掘技术（Data Mining technology, 简称 DMT），它是从海量数据中挖掘出有意义模式的技术，数据挖掘技术又称为知识发现（KDD），[3] 该技术如在采矿中挖掘有价值矿藏的过程，伴随着大数据时代的到来，数据的收集与存储急速增长，数据库中蕴藏着丰富巨大的知识供人类发现和学习使用。有关数据挖掘的方法和工具成为大数据分析研究方向之一，传统的数据挖掘方法有分类和聚类，分类即在海量数

RT:ex:4801:100010001；分部设置为 RT:im:4801:100010001，RT:ex:4801:100010000。根据 MPLS BGP VPN 的技术原理我们可以知道，PE 通过 RT 的 im 值决定是否接受路由，由图可知，只有公司总部的 im 值和分部的 ex 值是一致的，才能接收分部的路由信息，而公司分部之间 im 值和 ex 值不一致，故公司分部之间无法接受彼此的路由；而公司分部的 im 值和总部的 ex 值是一致的，所以可以接受总部的路由，从而就实现了总部和分部路由共享，而分部之间路由无法共享的要求，这也就是我们所说的星型网络，达到了用户分部只能与总部通信，而不能与分部进行通信的要求。

通过以上的实施方案我们更深刻的体会到了 MPLS BGP VPN 网络安全性、私密性、灵活性、可扩展性的优秀表现，随着公司 VPN 私有网络的迅猛发展，MPLS BGP VPN 技术将是构建 VPN 的发展方向，会越来越受到客户和运营商的青睐。

参 考 文 献

[1] 华为 MPLS BGP VPN 原理
[2] 华为 HCDP 系列教程

邓丽云，女，（1979.7–），汉族，广东遂溪，本科，现就职于中国电信股份有限公司海南分公司，研究方向：IP 网络安全、优化、维护工作

据中提取关键的数据，形成对数据信息的提取与分类；近几年对数据分类应用于文本情感词提取的方法的研究也很多，袁婷婷，杨文忠[2]等人在《一种基于性格的微博情感分析模型 PLSTM》中提出基于性格将文本观点词进行分类，人类性格模型可以为五大类，相似性格的人发表观点大致相同，建立 PLSTM 模型。李小龙[4]在《基于统计的分词系统字典模型研究》中提出基于统计的文本分词方法，通过 SVM 机器学习方法构建数据模型，从而提取情感分类器进行分词。刘刚[5]提出基于 Jieba 分词工具，Word2vec，网络爬虫工具等进行文本分词。另外，还有 k-means、层次聚类算法针对 Web 文档的分类。

Yu[6]等人对微博的主观文本分类时，使用了朴素贝叶斯算法对相似句子进行分类。数据挖掘算法为微博本文情感分析提供了丰富的实现方法和途径。

二、朴素贝叶斯算法

针对微博文本的情感分析研究主要围绕着信息的抽取和情感倾向的判定两大方面，均离不开对微博文本的分词工作，本文采用基于朴素贝叶斯分类算法对微博文本中的词汇的情感极性进行分类，并对比 PMI 互信息的优劣，为微博文本中的词汇进行归类，分为喜、怒、悲、惊、乐五大类，从而分析文本的情感倾向。

2.1 朴素贝叶斯算法

朴素贝叶斯算法（Naive Bayesian，NB）和决策树算法 (Decision Tree) 作为最常使用的两种分类方方法，经常用来解决数据分类的问题。[7]然而前者与后者相比，拥有更加稳定且高效的分类效果；在构建朴素贝叶斯分类器的过程中所需要的相关参数较少，同时对微博文本中缺损数据也不敏感；算法容易理解，分类效果较好。

朴素贝叶斯算法的执行公式：

$$P(B/A)=P(A/B)P(B)/P(A)$$

算法在执行的过程中首先通过预先描述的数据集对待分词文本进行建模，随后为样本事先预设类别，NB 算法是通过样本 1 的先验概率，计算出样本 1 的后验概率，选择最大的后验概率的类别作为样本 1 的最终归属分类类型。

2.2 PMI 算法

PMI 又称点互信息（Pointwise Mutual Information）该算法可以简化为一个样本数据中包含另一个样本数据的信息量【】，该算法定义成：

$$PMI(X;Y)=\log p(x,y)/p(x)p(y)$$

PMI 满足对称性；其中 P(x,y) 是样本数据（x,y）的联合分布，而 p(x) 和 P(Y) 是样本数据的边缘分布，点互信息的值可为正数或负数。

三、微博文本情感词分类实验结果

3.1 实验环境

文章通过实验验证朴素贝叶斯分类方法和 PMI 互信息分类方法，实验工具采用的 window7 平台和 VS2013；实验数据来自知网情感词典中的数据集，其中知网情感词典中待分类词有 4175 个，其中积极情感词 1240 个，负面情感词 2935 个。

3.2 分类实验结果对比分析

情感类别	待分类词分类个数
喜	779
怒	1666
悲	636
乐	461
惊	633

表 3-1 基于 NB 算法的 HowNet 分类结果

情感类别	待分类词分类个数
喜	779
怒	1666
悲	636
乐	461
惊	633

表 3-2 基于 NB 算法的 HowNet 分类结果

两类算法都是根据相似性进行数据分类，对情感词的分类效果一样，正确率都是相同，但朴素贝叶斯算法比 PMI 点互信息算法时间复杂度更低，且算法简单容易理解，所以本文采用朴素贝叶斯算法对微博文本中的情感词进行分类。

四、结束语

微博文本的情感分析作为数据挖掘的研究热门领域，随着微博使用人数的增长，和网路舆论监管力度不断加深，本文提出了一种针对微博文本情感分析中的情感判定的研究方法 – 基于朴素贝叶斯分类算法的情感分词方法，并通过实验对比分析了该方法与传统 PMI 点互信息分类方法的优越性，实验选取知网中的数据集作为实验数据，并将两种分类算法的效果通过表格的形式展示，研究后发现朴素贝叶斯算法具有执行速度快，时间代价小的优势。最终选着该算法完成对微博文本情感词的分类工作。未来考虑在复杂数据形式的微博文本中如何准确有效地提取情感词，并且尝试扩展情感词典中的网络符号，网络流行语以及特殊符号等代表情感的数据。

参 考 文 献

[1]http://www.xinhuanet.com/newmedia
[2] 袁婷婷 , 杨文忠 , 仲丽君 , 张志豪 , 向进勇 . 一种基于性格的微博情感分析模型 PLSTM[J/OL]. 计算机应用研究 :1-6[2019-01-04].
[3] 王彬菁 , 李明东 . 基于云计算的数据处理及数据挖掘方法 [J]. 软件导刊 ,2015,14(03):148-149.
[4] 李小龙 . 基于统计的分词系统字典模型研究 [J]. 湖北工业大学学报 ,2010,25(05):71-73+79.
[5] 刘钢 . 基于文本情感分析的企业舆情监测方法研究 [D]. 大连海事大学 ,2018.
[6] Yu Hong. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences [J]. Pediatrics, 2003, 116(3):58-59.
[7] 张俊飞 . 基于改进朴素贝叶斯算法实现评教评语情感分析 [J]. 现代计算机 ,2018(32):3-6.

王彬菁 1990 年 6 月 1 日 女 汉族 四川省成都市 成都师范学院 助教 硕士 管理信息系统