武汉邮电科学研究院硕士学位论文

开题报告

专 业： 通信与信息系统

研究方向： 中文信息处理

论文题目： 基于文本的用户情感分析

及兴趣关注点的研究

研 究 生： 邹珞彬 学号： 20150065

导 师： 王峥

报告日期： 2016年11月25日

|  |
| --- |
| 1. 课题的来源、目的及意义 |
| 随着社交网络的快速发展，互联网逐渐倡导“以用户为中心,用户参与”的开放式构架理念。互联网用户由单纯的“读”网页,开始向“写”网页、“共同建设”互联网发展,并由被动地接收互联网信息向主动创造互联网信息迈进。人们可利用微博、论坛等平台发表、分享自己的观点以及抒发某种情绪，因此,互联网(如博客、论坛等)上产生了大量的用户参与的、对于诸如人物、事件、产品等有价值的评论信息。这些评论信息表达了人们的各种情感色彩和情感倾向性,如喜、怒、哀、乐和批评、赞扬等。基于此,潜在的用户就可以通过浏览这些主观色彩的评论来了解大众舆论对于某一事件或产品的看法。  由于越来越多的用户乐于在互联网上分享自己的观点或体验,这类评论信息迅速膨胀,仅靠人工的方法难以应对网上海量信息的收集和处理,因此迫切需要计算机帮助用户快速获取和整理这些相关评价信息。情感分析(sentiment analysis)技术应运而生(本文中提及的情感分析,都是指文本情感分析)。情感分析的研究通常需要大量包含真实情感的语料文本。而在微博、贴吧、论坛等内容中，用户往往可以自由的发表自己的看法，不必受传统媒体的限制。因此对于这部分内容，通过分析其中的情感倾向性，可以挖掘得到许多有益的信息，比如对于某部电影的评论，对最新热点的看法，或者对于特定商品的喜好或者厌恶。这其中就包含了巨大的价值，比如对于舆论的监控，对产品的反馈等等。研究如何有效处理这些信息就显得非常有意义。  近年来，微博在中国获得了长足的发展，中文微博情感倾向性分析研究在对国内市场的购物推荐和评价，商业行为的引导上将会有着广泛而重要的应用。由于微博拥有庞大的用户量、紧密的社交化网络、以及每天数以亿计的数据量，基于微博的研究已经是一个高关注度的研究领域。当前基于微博数据的研究，主要包括基于微博文本的文本分析、基于微博社交网的关系挖掘和对微博上信息的传播方式的研究等。其中对微博内容的挖掘，主要将用户的微博看成短文本然后使用文本挖掘技术进行内容分析，或是挖掘用户其自身的兴趣分布，或是挖掘出用户对于某个产品的情感倾向等等；基于微博社交网的关系挖掘主要研究用户与用户之间因关注行为而形成的社交网，对社交网上点之间的连接关系进行分析。对用户兴趣关注点及情绪变迁分析的研究是自然语言处理、文本挖掘与心理学交叉领域的重要研究课题，它能分析用户的兴趣及情感，分析其情绪变迁趋势，这对探究自然语言信息背后隐藏着的舆情及情感趋势分析具有重要的应用价值。  课题的主要目的：是运用文本挖掘、文本分类的方法基于微博文本信息对用户的情感变化趋势的分析以及对用户的兴趣进行识别。通过建立用户兴趣模型，将用户按其兴趣分类，并加入时间因素以便能反映用户在不同时间段关注的兴趣点与情感变化之间的关系。  课题的意义：基于文本的情感分析及用户兴趣关注点的研究本质上是一个文本挖掘中的主要问题。首先发现自己感兴趣的微博用户与信息是微博用户在平台上最重要的活动，微博平台则可以准确地基于用户的兴趣爱好和情绪变化进行相关账号的推荐；其次微博搜索系统可以根据用户的兴趣爱好来定制搜索结果，提升用户的搜索体验。同时对于微博平台，如果能掌握社交网络中各个用户群体的兴趣，就可以找到有效的广告发布渠道，增强广告的定向投放，增加微博平台的收益。 |
| 1. 本课题的国内外研究现状及发展趋势分析   （在文献调研的基础上完成此部分内容，并列出至少10篇相关文献。） |
| 自 2006 年 Twitter出现以来，微博等社交网络快速发展。目前，国内主要的社交网络代表有新浪微博、微信、腾讯QQ等，国外的主要有Facebook，Twitter等。微博自然语言处理研究已成为新的研究热点，情感分析是其中的关键课题之一。宋静静[1]在针对中文短文本进行情感倾向性研究中指出微博短文本具有不规范性、实时性、交互性、稀疏性的特点，因此微博文本具有不同与其他文本信息的处理方法。微博文本情感分析的主要目的就是从这类微博信息中识别主观信息，挖掘用户对产品、新闻、热点事件等评论信息所持有的观点和态度，分析判断其情感倾向。  李激[2]在研究情感分析问题时，认为情感分析是一个分类问题，进而可以通过相关性信息将其划分为不同的情感倾向类别。处理情感分析问题通常有两种分类方式:谢星丽[3]等人在处理情感分析时直接把待分类文本分成中性、褒义和贬义三类情感倾向的文本，即一次三分类方式。而杨武[4]等人则采用两步二分类方式，第一次二分类实现主客观文本分类，即将待分类文本分为为主观性文本和客观性文本两类，第二次二分类则针对主观性文本，将其分成褒义和贬义两类文本。  在对情感进行分类的研究过程中，主要运用到各类分类模型以及特征选择算法。在模型方法方面，GARG[5]等人利用逐步回归、基于偏最小二乘回归的径向基函数、偏鲁棒 M-回归和主成分回归等展示了一些常见的统计模型方法，将其应用到多重共线性域中。LUO[6]等人提出一种基于传统的多信息特征选择的改进方法，通过对不同类别中词的不同表示来构建域特征词。PAIK[7]提出一个基于 TF-IDF 的权重计算框架，通过文档词频率归一化来决定对应词的重要性。WENG[8]实现了对微博网络结构的分析。刘志明[9]等人在针对中文微博情感分类的研究中，使用朴素贝叶斯和元模型对相同的微博文本数据进行了分类。对实验结果进行比较后发现，获得了比朴素贝叶斯和元模型更好的结果。此外，他们通过对相同的分类方法结合不同的权重计算算法进行了文本情感分类实验，结果显示当采用算法时，TF-IDF的性能更好。  在特征选择算法方面。目前特征选择方法主要包括基于频率的，基于信息熵的和基于统计量的方法。基于频率的方法主要是指TF-IDF方法。覃世安[10]等人在针对文本分类中TF-IDF方法的改进研究中指出改善此方法性能的主要途径是通过统计特征自身出现的频率和某个文档集合中出现的频率等因素，综合计算特征的重要性。张瑾[11]利用改进后的TF-IDF算法来提取情报中的关键词，所得到的研究结果更加准确。基于信息熵的方法包括“信息增益”，“互信息”和“期望交叉熵”。基于统计量的方法主要是指“卡方统计”（CHI），徐明[12]等人指出通过衡量特征和文本的开方统计值的高低来判断特征的重要性是提高卡方统计效果的重要手段。  总而言之，对于文本的情感分析经过了多年的研究和发展，呈现百花齐放的态势。和上述工作不同，本文是从微博用户的个性化兴趣关注点进行分析，通过一种基于微博文本的特征权重计算方法，将时间因素加入到此方法中，来计算不同时间段内特征词的权重，从而得到微博用户的关注信息，并根据此关注信息来分析微博用户在此时间段内情绪的变化情况，进而完成了用户个性化兴趣关注点及情绪的变迁的分析工作。  参考文献：  [1] 宋静静. 中文短文本情感倾向性分析研究[D]. 重庆理工大学，2013.  [2] 李激. 文本情感分析的发展综述[J]. 计算机工程与应用，2013，49（2）：136-137.  [3] 谢星丽，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报，2012，26（1）：74-84.  [4] 杨武，宋静静，唐继强. 中文微博情感分析中主客观句分类方法[J]. 重庆理工大学学报(自然科学). 2013，27（1）：51-56.  [5] GARG A,TAI K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity [J]. International Journal of Modelling, Identification and Control, 2013，18(4): 295-312.  [6] LUO Y, OUYANG N. Text similarity calculation based on domain feature word[A]. International Conference on Automatic Control and ArtificialIntelligence[C]. New York: IEEE. 2012:2049-2051.  [7] PAIK J H. A novel tf-idf weighting scheme for effective ranking[A]. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM. 2013:343-352.  [8] WENG Jianshu, LIM E P, JIANG Jing. Twitterrank: finding topic sensitive influential twitterers[A]. Proceedings of the Third ACM International Conference on Web Search and Data Mining[C]. New York: ACM. 2011:261-270.  [9] 刘志明，刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用，2012，48（1）：1-4.  [10] 覃世安，李法运. 文本分类中TF-IDF方法的改进研究[J]. 现代图书情报技术，2013，238（10）：27-30.  [11] 张瑾. 基于改进TF-IDF算法的情报关键词提取方法[J]. 情报杂志，2014，33（4）：153-155.  [12] 徐明，高翔，许志刚等. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用，2014，50（19）：113-117. |
| 三、课题中待解决的关键问题 |
| 本文主要针对微博用户利用微博动态及其评论，对获取到的文本进行情感分析。同时利用模型对用户兴趣关注点进行分类，在该过程中主要解决如下关键问题：  1、特征词的提取及权重计算。获取到文本数据后，需要对文本中的特征词进行提取从而确定用户的兴趣关注点，除了利用传统的文本分词和TF-IDF算法来计算特征词的权重，加入时间等因素从多个维度计算权重值将会更加准确地确定用户的关注点。  2、构建情感词典。语义知识库即词典的建立，给极性分析工作的开展以极大支持。利用词典能够直接判断相应词语的极性，同时也可以通过语义相似度、建立同义词典等方式对于未登录词进行极性的判断。基于情感词典的判别难度在于情感词典的构建。中文里有较多的一词多义现象，且在不同的语境下表达的意义可能相反。此外，微博作为一款互联网应用，各式各样的网络用语不断涌现，现缺乏相应的网络语言情感词库。  3、文本情感分析。情感分析的基础在于情感词典的构建，因此需要使用上述步骤中的情感词典作为基础情绪，同时利用对基础情绪的分类算法来分析微博用户的情绪及其变迁过程。 |
| 四、课题的研究内容 |
| 本研究选择了新浪微博的用户为主要研究群体，利用新浪微博平台的大量数据进行数据挖掘分析，构建基于算法的用户兴趣点模型，针对每个微博用户挖掘出相应的兴趣关键词，然后对微博用户进行情感分析得到其在一段时间内的情绪变化趋势。  具体的研究内容如下：  1、特征词的权重计算，利用新的特征值权重算法实现对微博用户微博内容和用户兴趣关注点的挖掘，然后从微博语料库中选取待定词汇，并根据它与标准情绪词汇在语料库中的互信息大小来计算特征词的权重，从而确定特征词找到用户兴趣关注点。  2、构建情感词典，本文将从已有的 4 种情感词库( 清华大学词典、台湾大学词典、大连理工大学词典、Word Net) 中构造关于( 高兴、愤怒、厌恶、恐惧、悲伤) 基本情绪的标准词库。  3、情感分析，基于情感词库，利用词汇匹配技术分析文本数据中情感倾向的方法目前还不成熟，仍存在许多缺陷，因此采用基于算法的情绪分类方法。所取得的实质研究成果，表明它确实是有效测量大众情绪的方法，所得的情绪指标能在一定程度上显著地反应用户的情感。 |
| 五、拟采取的研究方法、实验方案、技术路线 |
| 本课题的整体流程如下图所示，采用网络爬虫工具获取大量微博文本数据，并进行数据预处理（包括过滤链接博文、过滤广告博文、繁体与简体的转换以及博文去重等)将获取到的微博文本存储为结构化数据，以备后续分析使用。本文运用以传统的 TF-IDF 为基础的中文文本分词算法对文本数据进行处理，提取出微博博文的特征词并根据基于文本的特征词权重计算方法得到权重较高的特征词作为博文的主题词。通过这些主题词确定用户的兴趣关注点。同时对微博文本数据中的情感特征进行抽取，获取到微博用户的情绪类型并以已知的一些基本情绪类别作为参考，建立情感词典。利用建立好的情感词典对微博文本进行情感分类，发现微博用户的情绪特征及其变迁过程。最后结合微博用户的兴趣关注点与情绪变迁过程分析两者之间的关系。    图1.1 流程图（总）  1、中文分词  中文分词是研究本课题的重要基础。从中文自然语言句子中划分出有独立意义词的过程被称为分词。文本数据的预处理不仅需要正确的中文切分方法，还需要从切分的词语中利用词语的词性来正确地选取特征词、程度词和转义词。现有的分词工具有很多，大多是基于词典的切词工具，对于未登录词汇的识别能力还有待提高。在本课题中，我们需要针对涉及到的相应领域的词汇构建自己所需要的登陆词库。因此我们选择了利用Python语言开发的中文分词系统Jieba分词，从而在文本处理和情感分析的过程中保证最大程度上的减小错误率。  2、基于微博文本的权重计算方法  微博用户和它对应的博文可用四元组Q形式化表示，如公式（1）所示:  Q = ( User-id，keywords，weight，timestamp) （1）  式中: User-id 代表微博用户；keywords 表示在某一时间段从博文中提取出的特征词；weight 代表对应特征词权重；timestamp 表示相应的博文发布时间。  首先，对于四元组 Q 中特征词 keywords 参数的确定，使用基于篇章分析、中文分词、同义词和未登录词处理的算法，完成对博文主题词的抽取，此方法可弥补单独使用统计方法的缺陷以及避免语义分析中的低频词；其次，对于四元组 Q 中特征词权重 weight 参数的确定，是以传统的 TF-IDF 为基础的，但传统的 TF-IDF 方法只反映静态文档集中特征词的权重，不能有效地表示特征词在不同时间段内的分布信息。因此，本文首先利用基于点互信息(PMI) 的方法提取出特征词，然后提出与时间参数 timestamp 相结合的特征词权重计算方法。在此过程中，本文从已有的 4 种情感词库( 清华大学词典、台湾大学词典、大连理工大学词典、Word Net) 中构造关于( 高兴、愤怒、厌恶、恐惧、悲伤) 基本情绪的标准词库，然后从微博语料库中选取待定词汇，并根据它与标准情绪词汇在语料库中的互信息大小来确定特征词，以下为互信息计算公式，见式（2）:  PMI（W,S）=log （2）  式中: W表示在语料库中属于 k( 1≤k≤6) 类情绪下的词; S表示标准词库中第 k 类情绪下的第 j 个标准词。最后本文从语料库中选择与标准词汇的互信息最大的词作为特征词。为了反映特征词与关注点之间的关系，本课题加入了时间因素，特征词权重会在不同时期会有所改变，一些新的特征词将代替旧的特征词出现在特征词集合中。   1. 基于SVR算法的微博用户情绪分析   虽然目前在情绪认知分析领域还没有形成一个统一的基础情绪类别标准，但有些基本的情绪类别通常是被认可的。本文使用基本情绪( 高兴、愤怒、厌恶、恐惧、悲伤) 作为基础情绪来分析微博用户的情绪及其变迁过程。另外，除利用自然语言处理工具从文本信息中挖掘用户的基础情绪外，对表情信息的分析也至关重要。因此，建立一个微博表情库，形成一个情绪类型与表情图片的对应关系库来对表情符号进行分析。首先，进行数据预处理（包括过滤链接博文、繁体与简体的转换以及博文去重等）；其次，提取出微博文本特征，比如在微博文本中往往包含大量的表情信息，而这些表情最能直接反映用户当时的情绪。本文将抽取出的表情符号与基础情绪联系起来，形成一个表情库。对于修饰词、否定词、关联词、标点符号等，这些特征在影响情绪产生过程中也起到一定的作用。  最后，利用式（3）所示的卡方测试方法( 式中 Ni表示观察频数，n 是总频数，pi是期望频率) ，  = （3）  将文本中的高频词以及类别相关度大的情绪特征词挑选出来，作为辅助的情绪特征，并选择最优的特征数量形成 VSM 向量维度，以此为基础来调节SVR 参数和分类使用的阈值，实现对微博文本的最佳情绪分类效果。    图1.2 情绪处理过程   1. 性能评价指标   情感分析中常用的评价指标有准确率（Precision），召回率（Recall），F-测度值（F-measure）。  准确率是衡量情感分类结果的准确性，实验分类后，实验结果中判定为该类的正确数目占属于判定为该类数目的比值，其公式如式（4）所示：  （4）  召回率是衡量情感分类模型的完备性，实验分类后，实验结果中判定为该类的正确数目占本应判定为该类别数目的比值，其公式如式（5）所示：  （5）  准确率和召回率从两个不同的方面反映了分类的效果，但准确率和召回率是互相影响的，理想情况下系统希望同时提高准确率和召回率，做到两者都高。但是一般情况下准确率高、召回率就低，召回率高、准确率就偏低。由于准确率和召回率有时会出现矛盾的情况，为了综合考虑两者，F值（F-measure）就被选作为平衡两者的评价指标，其公式如式（6）所示：  （6）  式中，是一个综合度量指标，p是准确率，r是召回率，是一个可调参数，代表准确率p或召回率r在评价函数中的影响力度。小于1时，表示召回率r对于的影响力度较大；大于1时，准确率p对于的影响力度较大。通常认为准确率和召回率的重要性相当，将的值选定为1，这时的评价指标变为：  （7）  此时F-测度值也称为F1值。当F1值较高时，说明实验的结果较好。  为验证本文算法的有效性，在抓取的微博数据集上进行测试,并与传统的特征值选择算法进行比较，通过对比上述三项性能指标并不断改进算法使得本文算法的效果更加准确。 |

|  |
| --- |
| 六、论文的创新点或实用性 |
| 本文的创新点有：  1、本文是从微博用户的个性化兴趣关注点进行分析，通过一种基于微博文本的特征权重计算方法，首次提出将时间因素加入到特征权重的计算方法中，以计算不同时间段内特征词的权重，从而得到微博用户的关注信息，并根据此关注信息来分析微博用户在此时间段内情绪的变化情况，进而完成用户个性化兴趣关注点及情绪的变迁分析工作。  2、本文利用基于SVR的微博用户情绪变迁方法首创了基于情感分析的微博用户个人兴趣关注点动态变迁算法。与传统的 TF-IDF 方法对比，本方法得到的特征词权重将会有所提高，加入时间因素可以更加直观有效地展现微博用户在某个时期的兴趣关注点，而基于传统的 TF-IDF 得到特征词权重不能较好地反映该词在微博中随时间变化而产生的影响。  本文的实用性：  微博作为一种新型且受大众喜爱的社交媒体，用户量十分庞大，微博中的文本数据量也十分巨大，如果可以将微博的文本的内容进行情感和兴趣的分类，可以帮助了解文本发布者的当下情感和一段时间内的兴趣偏好。对于商家的个性化销售有很好的帮助，具有很强的商业价值，同时，个性化的服务对与用户来说，可以很大程度上提升微博用户的体验。 |
| 七、研究计划进度和预期成果 |
| 计划进度：  1、2016年10月-2016年11月，确定论文题目，进行相关理论研究；  2、2016年12月-2017年2月，学习机器学习的算法，以及切词工具的应用；  3、2017年3月-2017年4月，对微博文本爬虫，收集训练集；  4、2017年5月-2017年6月，对训练集进行中文分词处理，构建词库；  5、2017年7月-2017年8月，应用机器学习的方法进行文本分类；  6、2017年9月-2017年11月，完成论文撰写、修改定稿。  预期成果：  划分微博文本的情感和兴趣关注的种类，构建较为完整的参考词库，利用中文文本分词和机器学习算法对训练集处理分类，达到能将该类型的文本进行情感和兴趣关注方面的分类，并能提高分类结果准确率的效果。 |

|  |
| --- |
| 八、导师意见  （请对项目的意义、具体内容、创新点和取得预期成果的可能性等进行评价） |
| 本论文以微博文本为主要研究对象，通过基于微博文本的特征权重计算方法，同时考虑时间因素，提取一段时间内微博用户的个性化兴趣关注点，并根据此关注信息来分析其情绪变化趋势。论文选题合理，目标清楚，工作量适中，同意开题。  导师签字： |
| 九、研究生部审核意见 |
| 审核老师签字： |