文章编号：2095-2163(2019)01-0057-04 中图分类号：TP312 文献标志码：A

基于微博社交平台的舆情分析

盛成成，朱 勇，刘 涛
（南京工程学院 计算机工程学院，南京 211167）

摘 要： 社交网络当下已成为舆情传播的重要渠道之一，舆论在微博、贴吧等开放式社交网络传播尤为迅速。本文以微博社交平台为例，对网贷相关事件进行分析。主要利用网络爬虫进行数据采集、使用分词进行文本处理、使用 Word2Vec 结合机器学习技术进行文本分析并采用数据可视化呈现。研究网贷相关事件的热度及网友对事件的情感态度，分析该事件可能产生的舆论影响。

关键词： 舆情分析；网络爬虫；Word2Vec；机器学习

Public opinion analysis based on Weibo social network

SHENG Chengcheng, ZHU Yong, LIU Tao
(School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

【Abstract】 Social networking has become one of the important channels for public opinion communication nowadays. The public opinion spreads especially quickly on open social networks such as Weibo and Post Bar. This article takes the Weibo social platform as an example to analyze the online loan activities. It mainly uses Web crawlers for data collection, word segmentation for text processing, Word2Vec is combined with machine learning technology for text analysis and data visualization. Finally, the effect of online loan-related activities and the emotional attitude of netizens to the behavior are also studied, and the possible public opinion impact of the online loan activity is analyzed.

【Key words】 public opinion analysis; Web Crawler; Word2Vec; Machine Learning

0 引 言

近年互联网的不断发展，网络已经成为人们生活中不可或缺的一部分，而社交网络、移动互联网技术和智能手机的迅猛发展又为网民提供了快捷、迅速、有影响力的发声渠道。网民通过社交平台对于社会热点事件发表自己的观点情绪已成为主要现象，社交网络平台上网民的情感趋向产生了重大的社会影响。而随着大学生等低收入高消费群体的增长，以及大学生缺乏风险意识等原因，网络贷款引发的悲剧也屡见不鲜。对此，本文提出了一种基于社交网络的舆情分析方法，采用网络爬虫获取数据，使用 Word2Vec 建立词向量并采用机器学习中的半监督学习训练模型进行文本分类处理，最后采用数据可视化呈现网民的情感倾向、影响热度等，直观展现网贷行为的情况。

1 研究内容

本文进行社交网络网贷舆情分析的过程，主要包括数据挖掘、文本分类、情感分析、数据可视化 4 个步骤。数据挖掘采用网络爬虫实现，文本分类采用有监督的多层次二分类方法、基于标签传播算法的 MAD 吸附算法，情感分析利用文本分类中使用的有监督方法结合集成式学习提高准确度，最后利用 ECharts 进行数据可视化。技术路线如图 1 所示。

1.1 数据获取

数据获取主要包括 2 部分，分别是：信息爬取和分词。

(1)通过网络爬虫实现对微博的抓取。网络爬虫是一个自动提取网页的程序，为搜索引擎从万维网上下载网页，是搜索引擎的重要组成[1]。由于微博等社交网络平台采用反爬虫机制，对访问 IP 限制甚至直接封禁，为了自动化爬取足够数量的数据，在网络上众多代理网站上爬取大量免费代理 IP，并对其进行测试，高质量的存入数据库中，构建自己的代理池，通过代理池使用多个不同的 IP 访问微博 API，以此来消除访问 IP 限制对爬虫的干扰，绕过反爬虫机制进行数据获取。主要爬取的数据内容为：

① 短文本数据(描述性内容、评论等)。

基金项目： 南京工程学院大学生科技创新基金(TB201607004)。

作者简介： 盛成成(1996-)，男，本科生，主要研究方向：机器学习、数据分析；朱 勇(1998-)，男，本科生，主要研究方向：Web 开发、深度学习；刘 涛(1998-)，男，本科生，主要研究方向：机器学习。

收稿日期： 2018-11-08

② 发布消息的用户数据(昵称、年龄、地理位置、粉丝数等)。

③ 用户之间的联系数据(文本的转发关系、转发时间、用户粉丝的个人 id 等)。

④ 热度数据(发帖时间、发帖数、评论数、微博的转发数等)。

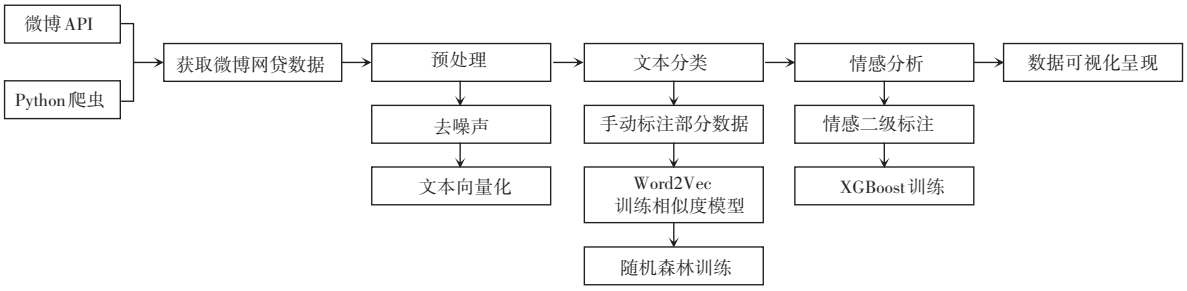

图 1 技术路线

Fig. 1 Technical roadmap

(2) 对抓取的舆情信息进行分词。分词按文本内容分为中文分词和英文分词。英文分词一般通过空格分开,易于实现。中文分词情况下词与词之间没有明显分隔表示,业界使用的主要方法有基于词表的分词方法、基于 n 元语法的分词方法、N-最短路径法等。本文采取了基于概率语言模型分词的 jieba 分词器中的全分词模式进行分词。jieba 自带了一个叫做 dict.txt 的词典, 里面有 2 万多条词, 包含了词条出现的次数和词性。在分词过程中,基于 Trie 树结构实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG),然后采用动态规划查找最大概率路径, 找出基于词频的最大切分组合。对于未登录词,采用基于汉字成词能力的 HMM 模型进行识别并重新计算最佳切分路径,输出分词结果。

1.2 文本分类

获取的数据当中存在多种类型的文本,包括客观新闻、网友的评论、参与者的发言等,不同文本反映不同的信息,所以需将其分类。采用多层次二分类方法,先将文本分成 2 类,再将分好的文本继续细化分类直至达到需求。

(1)首先对部分短文本进行手动标注来区分新闻报道、官方通告、网友发表的观点等。

(2)建立词向量,将文本进行数学化表达,作为训练模型的输入。这里使用了搜狐的互联网语料库进行 jieba 分词后,利用 Word2Vec 中的 CBOW 训练词嵌入(word embedding),将自然语言中的字词转为计算机可以理解的稠密向量,核心思路即“用词附近的词来表示该词”。在 Word2Vec 出现之前,自然语言处理经常把字词转为离散的单独的符号,也就是 One-Hot Encoder Word,每个词用长向量表示,向量维度是词表大小。向量中只有一个值为 1,其余都为 0。这种方法存在以下问题。一方面,单词编码是随机的,向量之间相互独立,看不出各个单词之间可能存在的关联关系。其次,向量维度的大小取决于语料库中字词的多少。如果将所有单词对应的向量合为一个矩阵的话,矩阵过于稀疏,会造成维度灾难(一个大的语料库维度超过几十万)。而 Word2Vec 将一个词所在的上下文中的词作为输入,而那个词本身作为输出。通过对一个大的语料库训练,得到一个从输入层到隐含层的权重模型。训练完成后,就得到了每个词到隐含层的每个维度的权重,就是每个词的向量表示(维度一般在 50~100)。对于句子“My major is computer science.”“my”与其它单词之间距离见表 1。其可视化表示如图 2 所示。

表 1 词向量示意

Tab. 1 Word vector

		距离
My	my	1.000 000 000 000 000 2
My	major	0.371 915 036 903 900 8
My	is	0.630 294 947 134 468 2
My	computer	0.491 617 662 280 982 64
My	science	0.392 278 897 587 195 86

(3)最后使用机器学习的有监督学习方法。有监督学习方法在训练过程中不仅输入训练数据,还输入分类的结果(数据具有的标签),模型经过训练后再输入未知特性的新数据也能计算出该数据导致各种结果的概率,输出一个最接近正确的结果。由于模型在训练的过程中不仅训练数据,而且训练结果(标签),因此训练的效果通常不错。将通过标注好的文本分词获得词列表并以此作为输入,采用集成方法随机森林进行训练。随机森林是 Bagging 方法的变体,Bagging 方法是对输入集进行有放回的随

机采样获得 m 个样本集利用不同的基学习器训练，然后将结果用投票法等策略组合。例如，若 3 个学习器结果为 1 1 0，那么最终结果就置为 1。对于质量高的文本，分类准确率在 85%以上。最后将数据输入训练好的模型中，获得全部文本的分类数据。分类流程如图 3 所示。

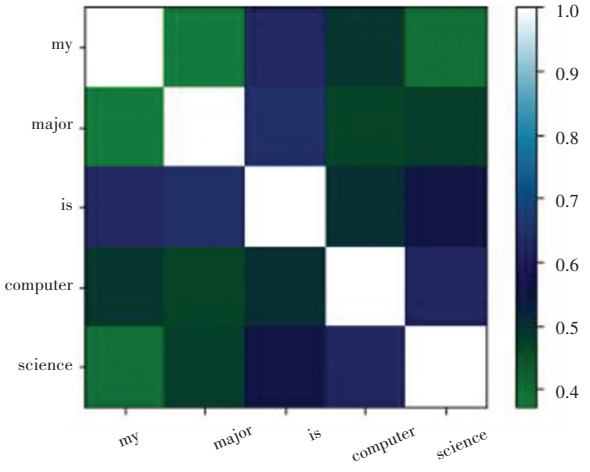

图 2　词向量可视化表示

Fig. 2　Word vector visualization

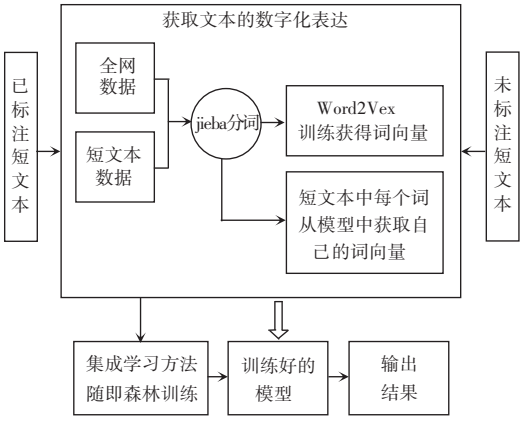

图 3　文本分类流程

Fig. 3　Text classification process

1.3　MADDL 吸附算法

MADDL（Modified Adsorption for Dependent Labels）算法是 MAD（Modified Adsorption）在考虑不同标签具有不同程度的依赖性的情况下的改进版本。MAD 算法是一种 GBSSL 算法（Graph-based Semi-Supervised Learning），即基于图的半监督算法。这是一种 LP（Label Propagation）标签传播算法。

（1）半监督学习

半监督学习（Semi-supervised learning）发挥作用的场合是：数据集中仅有部分数据存在 label。通常情况下数据集中大部分数据没有 label，即整个数据集只有少许几个 label。半监督学习算法会充分地利用 unlabeled 数据来捕捉整个数据集的潜在分布。其基于 3 大假设：

① Smoothness 平滑假设：相似的数据具有相同的 label。

② Cluster 聚类假设：处于同一聚类下的数据具有相同 label。

③ Manifold 流形假设：处于同一流形结构下的数据具有相同 label。

（2）LP 标签传播算法

标签传播算法（label propagation）的核心思想即相似的数据应该具有相同的 label。

构建节点相关系数的算法为：

$$w_{ij} = \exp\left(-\frac{1}{\sigma^2}\sum_{d=1}^{m}(x_{id} - x_{jd})^2\right)$$

（3）MAD 算法

吸附算法是一个用于传感器学习的通用算法框架，在这个框架中，学习者通常会得到一小套标记的示例和一组非常大的未标记示例。目标是标记所有未标记的示例，并可能在标签噪声的假设下，也重新标记已标记的示例。

与许多相关算法一样，“吸附”假定学习问题是以图形形式给出的，其中示例或实例表示为节点或顶点，并且边缘代码在示例之间相似。某些节点与预先指定的标签相关联，该标签在无噪声情况下是正确的，或者可能受到标签噪声的影响。其它信息可以以标签权重的形式提供。吸附通过边缘将标签信息从标记的示例传播到整个顶点集。标签使用每个标签的非负分数表示，某些标签的高分值表示高关联。

1.4　情感分析

在情感分析的技术实现上，先对短文本标注情感极性—消极或积极。利用 Word2Vec 计算词向量，然后使用支持向量机，贝叶斯等进行训练，准确度在 70%左右。了解集成学习后，利用了随机森林、XGBoost 模型，XGBoost 是一种集成学习 boost 方法，模型输出百分比，即对正负文本的相似度，本文直接采用为情感二极性。由于模型训练所耗费时间较多，无法进行增量的实时训练等问题，所以使用 SnownNLP 进行新的情感分析训练，以达到实时训练、实时预测的目的。

2　结束语

在对以微博为代表的社交网络“网贷”相关字

（下转第 64 页）