李子柒 YouTube 评论情感分析

小组成员：　　王靖淇 3220190879

孙天艺 3220190867

孙润庚 3220190866

石根 3220190861

目录

1 概述

近年来短视频异军突起,其移动化、社交化的特征使其很快成为最主流的内容形态之一。越来越多的内容创业者加入了短视频领域,该领域的发展也越来越成熟，并且短视频已经无意间扛起文化输出的重担。而视频的评论表明了观看者对视频情感倾向与喜爱程度，分析短视频的评论文本可以表现视频的质量、受欢迎程度以及文化输出的力度，为改进短视频文化、提高表现力提供了一个评价尺度。

情感分析（Sentiment analysis），又称倾向性分析，意见抽取（Opinion extraction），意见挖掘（Opinion mining），情感挖掘（Sentiment mining），主观分析（Subjectivity analysis）等，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。情感分析的研究与定义在计算机科学的蓬勃发展前，在心理学等领域已经是一个热门且成熟的研究方向了，人类在不同的情境下会有不同的情绪反应。通过一些特定的方法可以预测用户发表内容的情感。

目前情感分析主流的方法主要有基于情感词典和基于机器学习两种方法。基于情感词典法是最简单也是最符合直觉的方法[10,11,12]，通过判断特定的情感关键字是否出现在文本中从而给文本确定情感倾向性。基于情感词典法主要将情感词表和人工制定的相关规则结合，用已有的人工标注的情感词典去查找一个文本中包含正向情感词汇和负向情感词汇的个数，根据数量大小来判断情感极性。然而，基于情感词典的方法最主要的一个问题就是无法解决未登入词问题，尤其是很多包含情感倾向性的新兴网络词汇，导致最后结果的召回率偏低，而且需要手工制定情感字典，效率较低。

目前情感分析使用较多的还是基于机器学习方法，标注训练语料和测试语料，使用支持向量机(Support Vector Machine, SVM)、K 邻近(K-Nearest Neighbor, KNN)等分类器进行情感分类[2,3,8,9]。然而，基于机器学习的各项算法需要大量人工标注的数据，成本很高。

综上所述，考虑到两种方法缺陷，本文提出一种基于频繁模式树(Frequent Pattern Tree, FP-Tree)的无监督在线学习系统。FP-Tree 是一种关系分析、挖掘算法，其思想是构造一棵 FP-Tree，把数据集中的数据映射到树上，再根据这棵 FP-Tree 找出所有频繁项集。FP-Growth 算法是指，通过两次扫描事务数据

集，把每个事务所包含的频繁项目按其支持度降序压缩存储到 FP-Tree 中。在发现频繁模式的过程中，不需要再扫描事务数据集，而仅在 FP-Tree 中进行查找即可。通过递归调用 FP-Growth 的方法可直接产生频繁模式，因此在整个发现过程中也不需产生候选模式。由于只对数据集扫描两次，因此 FP-Growth 算法克服了 Apriori 算法中存在的问题，在执行效率上也明显好于 Apriori 算法。

具体来说，本项目将以美食短视频领域的领跑者李子柒及其团队为案例，通过爬虫技术获取其 YouTube 上的视频评论，利用 FP-Tree 算法构建频繁模式树，并构建在线系统对新进评论进行情感分析。

2 算法

2.1 频繁模式树

频繁模式树（FP-Tree）[13]是数据挖掘中用来挖掘频繁模式的一种算法模型。频繁模式（frequent pattern）是指频繁地出现在数据库中的模式，可以是项集，也可以是子序列或子结构。所谓项集，即项的集合。如果项集中包含 K 个项，则成为 K 项集。

挖掘关联规则的其中一种方法叫做频繁模式增长（Frequent-Pattern Growth, FP-growth），它可以挖掘全部频繁项集而无须多次扫描数据库，产生候选项集。分治策略是 FP-growth 的核心：首先是构造，将数据库压缩成一棵频繁模式树（FP 树）和保留项集关联信息的项头表，每个项通过一个结点链指向它在树中出现的位置。需要注意的是项头表需要按照支持度递减排序，在 FP-Tree 中高支持度的节点只能是低支持度节点的祖先节点。

在关联规则挖掘领域最经典的算法是 Apriori，其致命的缺点是需要多次扫描事务数据库。于是人们提出了各种裁剪（prune）数据集的方法以减少 I/O 开支， FP-Tree 算法就是其中非常高效的一种，FP-Tree 数据结构如图 1 所示。

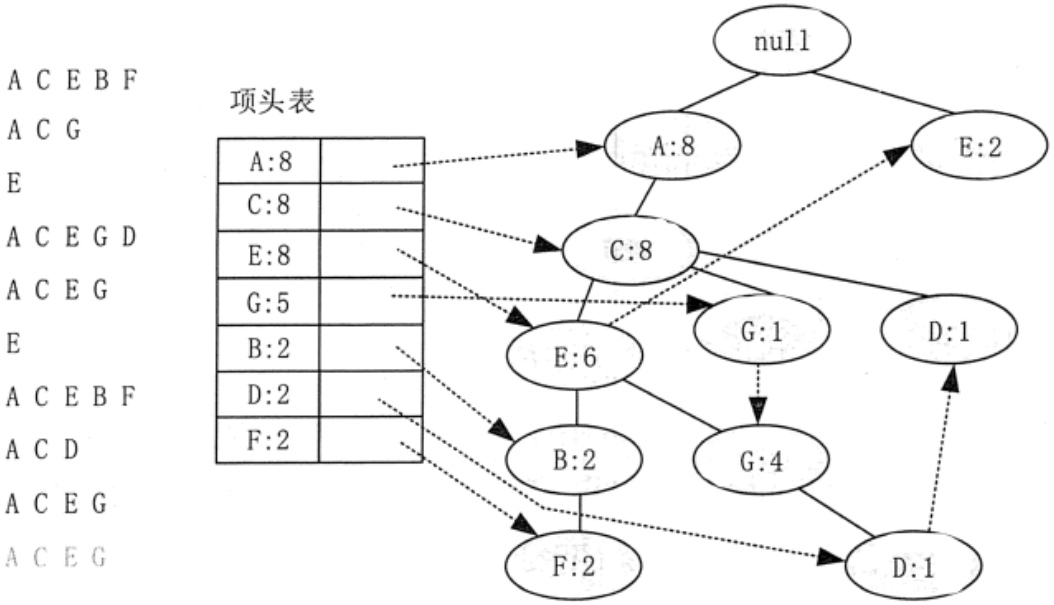

图 1 FP-Tree 数据结构

2.2 在线系统

在线学习是工业界比较常用的机器学习模型训练方法，例如[1]构建了一个对有关总统选举的评论信息进行情感倾向性分析的在线系统。它不是一种模型，而是一种模型的训练方法，在线学习能够根据线上反馈数据，实时快速地进行模型调整，使得模型及时反映线上的变化，提高线上预测的准确率。在线学习的流程如图 2 所示，包括：将模型的预测结果展现给用户，然后收集用户的反馈数据，再用来训练模型，形成闭环的系统。本项目首先输入收集的评论数据集，对数据集进行情感词典分词，根据词性剔除停用词和无关词，通过 FP 树构建算法对数据集进行处理，构建 FP 树并挖掘频繁项集，最终输出为频繁项集的词典。根据数据集构建积极与消极两个频繁项集，在线系统输入新评论，新评论分词并于两频繁项集进行匹配，获得分词的情感评分，最终合成输出评论的情感评分。

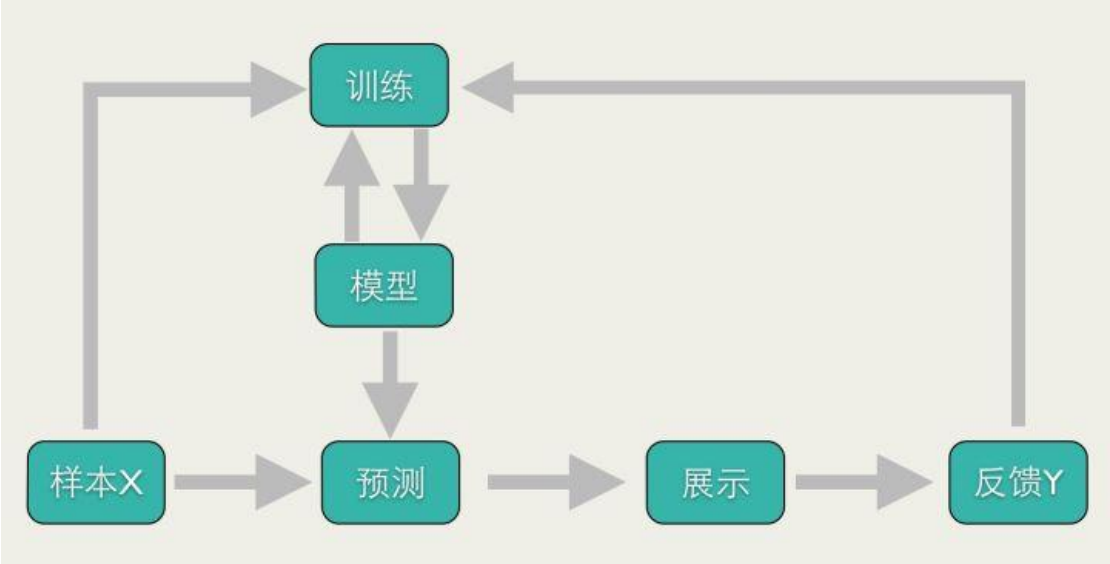

图 2 在线学习系统流程图

3 数据集

3.1 数据爬取

采用模拟浏览器行为的 selenium 库来进行数据爬取，数据爬取流程如图 3：

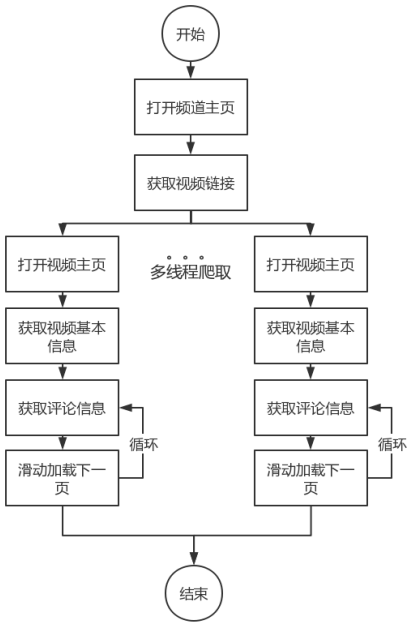

图 3 数据爬取流程图

我们爬取的数据主要包括视频的简要信息、视频详细信息、不完整的视频评论以及完整的视频评论。其中简要信息只包括视频题目、发表的时间、视频时长以及观看人数。视频的详细信息，除了简要信息之外，补充了每个视频的点赞数、点踩数以及评论数。李子柒一共是 100 个左右，李子柒的这 101 个视频的最早最晚时间差是 2 年 4 个月。不完整评论视频是指，每个视频爬取了 600-1000 条评论。李子柒 101 个视频一共爬取了 89748 条。完整评论视频是指爬取了这个视频的所有一级评论。李子柒 4 个视频一共是（48051）条，总的评论数据大概是 13.7w 条。

3.2 数据分析

图 4a 通过将评论按语言类型进行区分，同时将英语默认为一个国家。可以发现李子柒的中国观众仅占 16%，其观众国籍遍布世界各地，说明了李子柒在海外颇受欢迎。图 4b 统计 100 个视频的播放量,我们可以看到李子柒的每个视频播放量都很高，平均在 970w 左右，最高的有 4000w 左右的，最小的也有 200w 的播放量，由此可见李子柒的文化影响力是非常强的。图 4c、4d 分别统计了 100 个视频的点赞数/观看数，以及点踩数/观看数，也就是点赞率和点踩率，经过统计分析，点赞率平均为 1.8%，最高为 4.3%，最低为 0.8%；点踩率平均为 0.037%，最高为 0,065%，最低为 0.025%。通过对比可以看出，观众对于李子柒是喜爱远大于讨厌的。

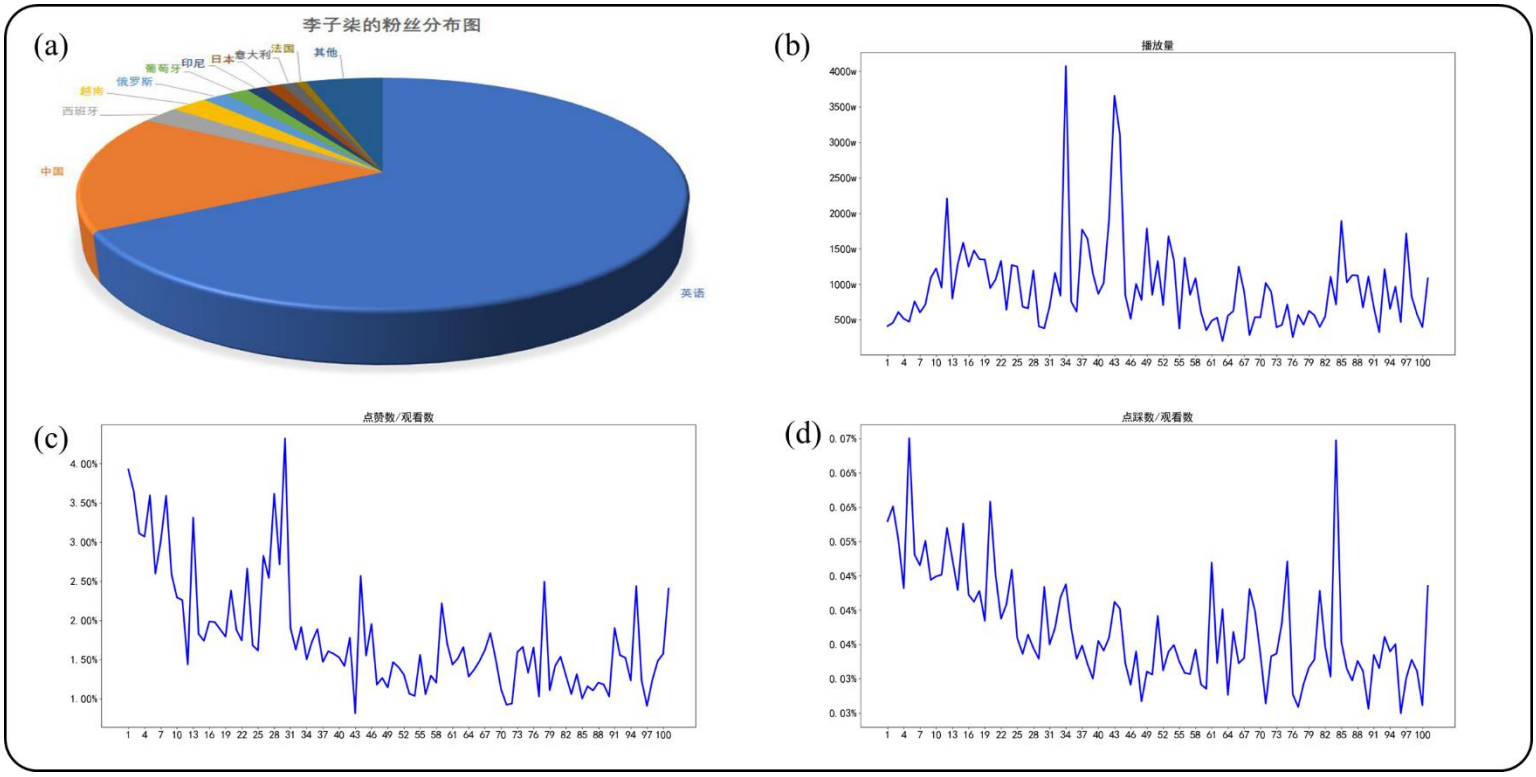

图 4 李子柒视频评论挖掘分析

4 系统展示

4.1 系统设置

本系统使用 python 语言实现，以网站形式呈现。客户端操作系统为 Windows10，浏览器为 chrome，系统服务器为 python 3.7，使用库资源包括 django 1.11，BeautifulSoup4 4.6.0，Jieba 0.39，Requests 2.18.4。

4.2 运行展示

运行 PowerShell (Windows)，进入项目根文件夹，输入 python manage.py runserver 8000 ，运行。在浏览器中输入：http://127.0.0.1:8000/sentiment/index/ ，即可看到系统主页面(图 5)。

随后输入短句评论，即可看到分词情况(图 6)，以及最终对该评论的情感评分(1-10),在 PowerShell 中也可看到对分词的评分状况。

图 5 在线分析系统主界面

图 6 分词与文本情感分析情况

5 结论与讨论

本次实验将关联规则挖掘与文本情感分析进行了融合。关联规则挖掘最初的目的是发现条目中频繁共现的项目，挖掘其中的规则。而文本情感分析是分析，处理，总结和推断带有情感色彩的主观文本的过程。互联网中的信息交流中心如博客，论坛，社交服务网络等，生成了大量用户参与的有价值的评论信息。 这些评论信息显示了了人们不同的情感色彩和倾向，例如喜悦，愤怒，悲伤，喜悦

和批评以及赞美。

两者融合的基础是文本的情感倾向是根据文本中的单词而定的，而单词的情感倾向可以基于词与词的关联关系来确定。关联分析是数据挖掘的核心技术，通过关联分析挖掘技术，获得已有文本中词间潜在关系。对于新的文本，首先将其分词，然后将分词与已有词进行关联分析匹配，根据已有词的情感倾向对匹配词进行情感评分，从而获得整个文本的情感极性。

本项目也由此构成。我们首先利用了爬虫技术获取了李子柒在 YouTube 上近期的视频评论，构成了初始的文本。随后基础情感词典的方法，将评论分成了积极评论与消极评论两部分，并对其进行分词处理，利用上述提到的 FP-Tree 关联算法，挖掘积极词与消极词之间的关联关系，获得积极频繁项集与消极频繁项集。当需要预测一个新的文本评论的情感倾向时，只需要将其分词，然后与频繁项集进行匹配，获得分词的情感倾向，进而得到整个新文本的情感极性。于此同时，我们还构建了一个在线学习系统，以此可以边预测边学习，更好地丰富训练样本。

本项目构建有以下优点：

1. 充分利用数据。互联网中存在着大量文本，但是对其进行人工情感标注十分困难，数据量少效果又会欠佳。而本系统进行情感分析，可以直接获得文本情感极性，无需人工标注，分析迅速，适合大量数据情况。
2. 实时分析训练。本项目构建了一个在线系统，可以边预测边训练，提供情感态势感知。同时提供时序分析，分析一段时间内的情感倾向变化。

近年以来，也出现了许多基于深度学习的情感分析方法[4]。例如，基于卷积神经网络的文本情感分析方法[7]。也有利用注意力机制与 Transformer 来进行短文本情感的方法[5,6]。基于深度学习的模型准确率较高，但其也有训练时间长，硬件成本高昂，上线难度大等缺点。

参考文献

[1] Wang H, Can D, Kazemzadeh A, et al. A system for real-time Twitter sentiment analysis of 2012 U.S. Presidential election cycle[C]// ACL 2012 System Demonstrations. 2012:115-120.

[2]戚天梅,过弋,王吉祥,王志宏,成舟.基于机器学习的外汇新闻情感分析[J].计算机工程与设计,2020,41(06):1742-1748.

[3]许诺,王德广,赵煜,王宇.基于机器学习的舆情分析系统[J].微型电脑应用,2020,36(05):53-56+63.

[4]李丽华,胡小龙.基于深度学习的文本情感分析[J].湖北大学学报(自然科学版),2020,42(02):142-149.

[5]李福鹏,付东翔.基于 Transformer 编码器的金融文本情感分析方法[J/OL].电子科技,2020(09):1-6[2020-07-02].http://kns.cnki.net/kcms/detail/61.1291.TN.20191017.1338.074.html.

[6]吴小华,陈莉,魏甜甜,范婷婷.基于 Self-Attention 和 Bi-LSTM 的中文短文本情感分析[J].中文信息学报,2019,33(06):100-107.

[7]刘书齐,王以松,陈攀峰.基于 CNN-ATTBiLSTM 的文本情感分析[J].贵州大学学报(自然科学版),2019,36(02):85-89.

[8]范文慧. 基于机器学习的网络舆情文本情感分类方法研究[D].电子科技大学,2019.

[9]韩开旭. 基于支持向量机的文本情感分析研究[D].东北石油大学,2014.

[10]郑毅. 基于情感词典的中文微博情感分析研究[D].中山大学,2014.

[11]李钰. 微博情感词典的构建及其在微博情感分析中的应用研究[D].郑州大学,2014.

[12]周咏梅,杨佳能,阳爱民.面向文本情感分析的中文情感词典构建方法[J].山东大学学报(工学版),2013,43(06):27-33.

[13]宋余庆,朱玉全,孙志挥,陈耿.基于 FP-Tree 的最大频繁项目集挖掘及更新算法[J].软件学报,2003(09):1586-1592.