本科生毕业论文

题目： 微博实时热点分析系统的设计与实现

Title Design and Implementation of Weibo Real-time Hotspot Analysis System

姓　　名：牛嘉诚

学　　号：1300016631

院　　系：信息管理系

专　　业：信息管理与信息系统

指导教师：王继民 教授

2017 年 6 月

北京大学本科毕业论文导师评阅表

<table>
<tr><td colspan="2">学　　号</td><td>1300016631</td><td>学生姓名</td><td>牛嘉诚</td><td>论文成绩</td><td></td></tr>
<tr><td colspan="2">学院（系）</td><td colspan="3">信息管理系</td><td>专　　业</td><td>信息管理与信息系统</td></tr>
<tr><td colspan="2">导师姓名</td><td>王继民</td><td>导师单位</td><td>信息管理系</td><td>职　　称</td><td>教授</td></tr>
<tr><td rowspan="2">论文题目</td><td>中文</td><td colspan="5">微博实时热点分析系统的设计与实现</td></tr>
<tr><td>英文</td><td colspan="5">Design and Implementation of Weibo Real-time Hotspot Analysis System</td></tr>
<tr><td>导师评语</td><td colspan="6">导师签名：</td></tr>
</table>

版权声明

微博实时热点分析系统的设计与实现

摘　要

随着互联网的飞速发展，新浪微博作为新生代网络媒体，成为了海量信息的集散地。在此基础上，新浪微博基于用户的搜索数据建立了实时热搜榜。该榜单能够较为客观地反映实时热点，成为用户获取热点资讯的重要渠道。通过对微博实时热点进行分析，可以为用户提供更为优质的热点资讯服务，达到热点追踪、舆情监测的目的。

本文设计并实现的就是这样一个微博实时热点分析系统。本文调研了国内外的微博分析系统、实时热点分析方法和短文本分类方法，设计出了适用于微博实时热点的分类算法，实现了对微博热点的自动分类。然后，本文设计了系统的功能、框架和数据库结构，将系统划分为五大模块：数据获取模块，数据存储模块，机器学习模块，数据分析模块和前端展示模块。最后，本文用 bootstrap 前端框架和 web.py 后端框架实现了整个系统。

关键词：微博实时热点 数据分析 文本分类 贝叶斯模型 TF-IDF 算法

Design and Implementation of Weibo Real-time Hotspot Analysis System

Niu Jiacheng (Information Management and Information System)

Supervised by Wang Jimin

Abstract

With the rapid development of the Internet, Sina Weibo as a new generation of network media, has become a distribution center of mass information. On this basis, Sina Weibo set up a real-time hot search list based on users' searching data. The list reflects the real-time hotspot objectively and becomes an important channel for users to get hotspot information. By analyzing the real-time hotspot of Weibo, we can provide users with higher-quality hotspot information service, and reach the goal of hotspot pursuit and public opinion monitoring.

This paper is exactly the design and implementation of a Weibo real-time hotspot analysis system. This paper investigates Weibo analysis systems, real-time hotspot analysis methods and short text classification algorithms at home and abroad, then designs the classification algorithm suitable for Weibo real-time hotspots, realizes the automatic classification of Weibo hotspots. After that, this paper designs the functions, the framework and the database structure of this system, divides the system into five modules: Data acquisition module, data storage module, machine learning module, data analysis module and front-end display module. Finally, the whole system is realized with bootstrap front-end frame and web.py back-end framework.

Keywords: Weibo real-time hotspot Data analysis Text classification Bayesian model TF-IDF algorithm

目 录

表目录

图目录

1. 绪论

1.1. 研究背景和动机

随着互联网的飞速发展，大量新兴的网络媒体涌现，人们也逐渐习惯于从网络媒体获取热点资讯。在这样的大环境下，新浪微博以其信息的实时性、丰富性，信息形式的多样性吸引了大量的用户，成为海量信息的集散地。在此基础上，新浪微博通过机器筛选+人工筛选的方式，提取并归并被用户大量搜索的词条并按搜索指数排序，建立了每分钟更新一次的实时热搜榜。该榜单能够较为客观地反映实时发生的热点事件。在实际生活中，有大量热点事件通过微博用户的传播和搜索，迅速形成热点进入热搜榜单并为更多人所知。也正是因为微博热搜榜的这种特点，使得它的实时性和信息量甚至超越了很多专业的新闻媒体。目前，微博实时热搜榜已成为很多用户获取最新热点资讯的重要渠道。但与此同时，微博实时热搜榜作为质量高、代表性强的资讯数据，并没有被有效的存留和分析。实时热搜榜每分钟一次的刷新频率保障了信息的实时性，但无法反映一段时间内舆情的变化情况、热点事件的关键时间节点等有价值的信息。此外，微博热搜榜目前只支持按热度排序，没有对热搜词条进行分类，用户无法选择自己感兴趣的类别浏览热搜词条。

在调研新浪微博时，笔者注意到新浪微博还有一个热门微博页面。该页面的微博条目是新浪微博根据微博的评论数、转发数、点赞数以及微博博主的用户知名度遴选而来。登上热门微博页面的微博，往往已经累积了一段时间的评论数、转发数、点赞数，因此实时性不如微博热搜排行榜。但是，新浪微博会通过机器筛选+人工筛选的方式，为热门微博分类，用户可以在热门微博页面按类别筛选热门微博并进行浏览，这一功能是热搜榜所不具备的。笔者设想，使用分门别类的热门微博文本数据可以进行机器学习，得到文本自动分类的模型，并对微博热搜榜里的热点词条进行自动分类，从而提高用户体验。

在上述背景下，笔者认为，通过对微博热点进行分析，可以得到每个热点出现的时间和热度随时间变化的规律，并对其自动分类，最终结果既可以用于舆情监测，也能够为用户提供更为优质的热点资讯服务。本文最终实现的就是这样一个微博实时热点分析系统。

1.2. 国内外研究现状

1.2.1. 国内微博分析系统现状

目前国内已经有一些投入实际使用的微博分析系统，但还没有定位和本文实现的系统完全一样，对微博热搜排行榜进行分析的系统。在本小节中，笔者调研了目前已投入实际使用的两个微博分析系统：知徽分析平台和 PKUVIS 微博可视化分析工具。

- **知徽分析平台**

知微是一个可视化的微博传播分析平台，它利用自然语言处理、网络结构分析、数据可视化等技术，对单条微博的传播特征进行全面分析。知徽平台的各模块功能如表 1-1 所示。[1]

表 1-1 知徽平台模块功能表

模块名称	功能名称	功能说明
总览	整体评价	本消息在曝光量（未去重的转发者粉丝加和）、用户总评（用户活跃度、粉丝量等指标的加权平均）、情感值 （正负情感）、内容分析（消息传播深度）四个维度，与行业标准的比较
	消息传播各项指标	包括用户质量、水军比例、短链点击数等的总体概述
	微力值	综合该消息的传播深度、广度及参与用户各项指标加权后得出的微博影响力总体评价
传播分析	转发时间趋势	各时段转发量数值及相应的参与意见领袖（KOL）
	关键账号	带来二次转发最多的前十个微博账号，及其传播路径
	转发层级分析	显示各层的转发数量
参与者信息	地域分析	转发者的所在省份，显示各省份的转发人数、比例、在全国的排名
	微博来源	转发者使用各客户端的比例分布
	认证・男女	转发者的性别比例、认证类型比例
	粉丝质量	所有转发者的粉丝的各区间分布
	活跃用户	极活跃用户、较活跃用户、活跃用户、不活跃用户的数量（根据大量采集的用户行为划分区段）
引爆点	引爆点	十大关键传播账号
短链分析	链接地址	所分析的微博中含有的短链接
	点击数	短链接被点击次数

	分享数	短链接在微博上的被分享次数
	评论数	短链接在微博上的被评论次数
	Referer 来源	前五大短链点击来源
	点击地域分布	各省份点击该短链的用户的数量和比例
	点击/转发比例	各省份的点击用户数和转发用户数比例
水军分析	总体分析	给出无水军、疑似水军、轻度水军、重度水军判别
	水军危害	文字说明微博营销中频见水军的危害
	营销账号水军分析	给出该条微博转发者中的营销账号的水军比例
内容分析	情感值	该条微博转发中呈现出的正能量、中性能量、负能量数值
	关键词/字	给出转发语中的高频词，并分别按序给出正面高频词和负面高频词，并可查看正面/负面高频词的提及次数，以及 提及该关键词的微博
	提及关键词的转发	包括转发者、转发时间、微博地址

知徽提供的数据分析服务是从单条转发量大于 30 的微博链接着手，并根据微博的转发量收取资费。从知徽的功能模块可以看出，知徽通过爬虫技术追踪到了所有转发待分析微博的用户，并分析转发后的微博的评论情感倾向、转发微博的用户的特征等，甚至进一步追踪二级转发该微博的用户，使待分析的文本从一条简单的微博，扩展为一个庞大的文本网络。由此可见，在对微博进行分析时，链接、转发、评论等都能够作为原有微博语义的扩充。

- **PKUVIS 微博可视化分析工具**

北京大学 PKUVIS 微博可视分析工具 (WeiboEvents) 是北京大学可视化与可视分析研究组开发的微博传播分析工具。它通过直观的视图清晰地呈现出一个事件中微博转发的过程，让用户能够迅速地发现事件中的关键人物、关键微博、重要观点，同时通过可视化的方式帮助用户更好地分析新浪微博中事件的发生与发展过程。用户在与页面交互的过程中，不仅可以看到故事发展的全貌，还可以像一个真正的侦探一样发掘出故事背后的秘密。[2]

PKUVIS 系统主要分析单条微博的传播，分析内容有传播分析图、关键词提取、认证比例、转发层级、性别比例、用户地域分布、用户微博数时间走势

等。如图 1-1 所示就是 PKUVIS 的主界面。在图中可以清晰地看出一条微博被多级转发之后形成的传播网络，传播热度随时间的变化规律，关键词等信息。

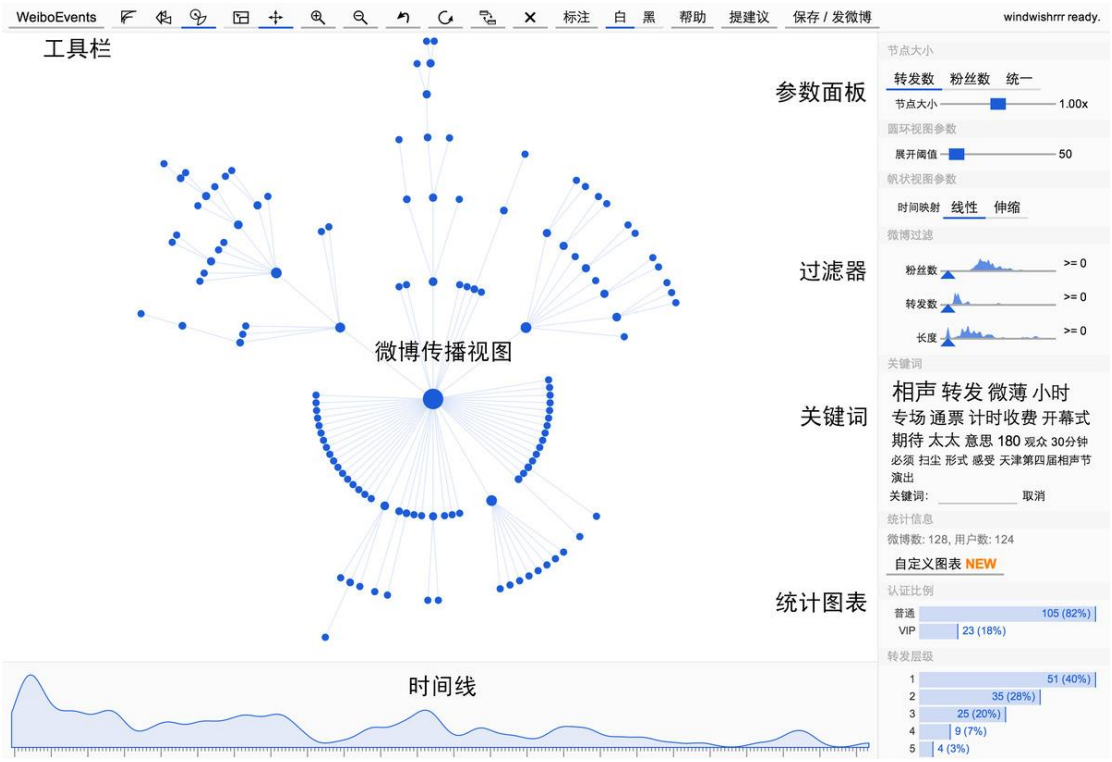

图 1-1 PKUVIS 系统主界面

1.2.2. 国内外实时热点分析方法研究现状

实时热点指的是当下正在发生的热点事件或者被广泛讨论的热点话题。实时热点的表现形式多种多样，例如各类新闻、新浪微博的实时热搜榜榜单、北京大学未名 BBS 上的十大热门话题等。国内外对实时热点的分析方法多种多样，笔者调研了国内外一些分析实时热点的论文并在本小节予以介绍。

- **国外研究现状**

FD Martino_Affanb 等人采用 Fuzzy c-means (FCM)聚类算法，对热点事件进行主题聚类，并得到热点话题。该算法允许一个热点事件属于多个不同的类，能够充分考虑到自然语言文本的类别成分的混合性、复杂性。[3]

H Senaratne 等人在分析 Twitter 中的热点事件时，引入了在一定时间内，用户发 Twitter 的地理位置的变化情况和 Twitter 文本内容的情感变化情况，对热点事件在时间、空间上的演变进程进行了深入的分析。[4]

- **国内研究现状**

邹盼湘在研究网络舆情热点的自动提取与分析时，采用了命名实体识别技术，对文本进行切词和词性标注，并从中抽取出时间、地点、人名、事件内容等关键信息并将其关联起来，实现了对网络热点的自动提取。[5]

张寿华等人采用 TF-IDF 算法从文本中提取出了热点事件的关键词，并根据这些关键词构建向量对热点事件进行聚类得到热点话题。然后根据一个热点话题的聚类文章数、参与人数、评论数目计算出该话题的热度，作为热点话题的评价指标。[6]

1.2.3. 国内外短文本分类技术研究现状

- **文本分类技术发展历程**

文本分类指的是基于待分类文本集的内容，对文本集按照一定的分类体系或标准进行分类标记。一般而言，文本分类任务由计算机根据一定规则自动完成。[7]

短文本分类是文本分类的一种，通常是针对微博、论坛帖子、网络留言及回复、手机短信、即时聊天记录等进行分类。因为文本内容少，只有寥寥几十字到 100 字，短文本特征并不明显，所以常用的基于统计与向量空间模型的方法分类效果不好。针对这种情况，在进行短文本分类时，需要对短文本进行语义延伸和特征扩展，同时通过人工介入的方法来提高准确率。

在早期的文本分类中，主要采用了信息检索技术中经典的布尔模型对文本进行分类。这样，表示文本和类别的特征一般较少，分类的准确率不高，无法达到处理大规模真实文本的实用目的。后来，随着对自然语言处理及人工智能技术的研究日渐深入，能够反映关键词权重、量化文本相关度的向量模型开始出现并取代布尔模型。在这一过程中，TF-IDF 算法被提出并得到广泛应用。TF-IDF 是一种统计方法，用以评估一个字或词对于一个文件集或一个语料库中的其中一个文本单元的重要程度。字词的重要性随着它在文本单元中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，最终每个词的重要性能够被量化表示，从而构建文本向量。[8]

随着自然语言处理学科的形成和发展，曾经一度被当作信息检索问题的文

本分类问题已经被视为机器学习的一种特定形式进行研究。而在信息检索领域被广泛运用的 TF-IDF 算法，也被运用于文本特征选择、词权重计算，进而运用于更为广泛的文本分类问题。

在目前的文本分类研究中，较为常用的手段是采用基于统计学习的方法（包括 TF-IDF）抽取文本特征，采用数据挖掘的经典分类算法进行类别学习。几乎所有重要的机器学习算法在文本分类领域都得到了广泛应用。例如基于最小二乘拟合的回归模型、最近邻分类、贝叶斯概率模型、决策树模型、神经网络模型、支持向量机模型、最大熵模型和隐马尔可夫模型等都能够被用来构造文本分类器。

近几年随着 Twitter、微博等网络媒体的兴起，短文本的数据量不断增加，国内外从对短文本分类的研究也不断增加。笔者调研了国内外短文本分类技术的研究现状并在本小节予以介绍。

- **国外研究现状**

Sriram B 等人在对 Twitter 的文本进行特征提取和分类时，引入了 Twitter 作者的信息，扩充了 Twitter 文本的特征量，提高了分类效果。[9]

Sankaranarayanan J 等人从 Twitter 中获取新闻事件，引入了 Twitter 的评论信息和转发信息，扩充了 Twitter 文本的特征量，并通过手工标注的方式，标注出了扩充后的文本中的新闻事件，然后用机器学习的方式使他们的系统具备了从 Twitter 中自动抽取新闻的能力。[10]

从国外的研究中可以看出，特征扩展是短文本分类技术的重点，人工介入也是提高短文本分类效果的重要手段。

- **国内研究现状**

王细薇等人用 FP-Growth 算法挖掘训练集特征项与测试集特征项之间的关系，然后用得到的关联规则对短文本测试文档中的概念词语进行特征扩展。[11]

Xue B 等人用 Word2Vec 算法，读入切词后的新浪微博文本进行训练，得到的 Word2Vec 模型可以判定词与词之间的相关关系并输出一个词的关联词语，例如输入“北京”，输出“首都”。通过这种方法，可以对微博文本进行同义词扩充，从而达到特征扩展、提高分类效果的目的。[12]

新浪微博的热门微博页面所提供的分类功能，也并不只是基于单条微博的内容，而是考虑了发微博的大 V 用户本身的用户属性，并且默认了大 V 用户的大部分微博都可以归到同一类。在实际操作中，新浪微博预先将用户的所有微博合并并对用户分类，在之后对热门微博进行分类时就可以将博主的用户属性纳入特征考虑，从而避免了单条微博特征稀疏的问题。[13]

1.3. 本文研究工作

本文的主要研究内容是微博实时热点分析系统的设计与实现。本文实现的热点分析系统是基于微博热搜词条随时间的变化情况，对热搜词条进行舆情分析，并根据热搜词条对应的微博搜索结果，对热搜词条进行文本扩充和特征扩展，进而实现对热搜词条的自动分类。系统的自动分类模块本身也独立出来，作为一项系统功能，提供文本输入端口，对用户输入的文本进行切词和类别成分分析。系统通过 web.py 框架，实现了 Python 语言的后台环境，可以借助 Python 海量的工具包支持复杂的数据分析功能。系统前端采用 Bootstrap 网页框架和 Echarts 图表框架，实现了完全数据驱动的系统前端展示页面。

本文设计并实现了微博爬虫，能够抓取微博实时热搜榜单和榜单上每个热搜词条对应的搜索结果并存入数据库。通过在服务器部署爬虫，本文实现了对微博榜单的定时抓取，保证了用于分析的数据的实时性和历史数据的连续性、完整性。

本文研究工作的重点是自动文本分类器的构建和训练。本文从新浪微博的热门微博页面获取分门别类的数据，并沿用热门微博的分类类目构建训练集，利用 TF-IDF 算法计算词权重，提取文本特征，尝试采用“词权重和”模型、朴素贝叶斯模型等不同的机器学习算法构建分类器，实现系统的自动分类功能。

1.4. 论文组织结构

第一章是本文的绪论，介绍了本文研究的背景和动机、国内外研究现状，并简要介绍了本文的研究工作。

第二章是本文的研究重点，介绍了文本分类的算法和具体的数据准备、数据挖掘工作。

第三章介绍了本文实现的微博热点分析系统的整体设计，包括系统框架、数据库结构、系统模块等。

第四章介绍了系统的具体实现，包括用到的编程语言、框架和工具包，并展示了系统效果。

第五章是本文的总结。

2. 微博热点分类方法

本章节的主要内容是文本分类算法的选取、训练数据的准备、文本分类器模型的训练与测试。最终，本章节基于文本分类器模型的测试结果确定了对微博热点的分类方案。

2.1. 文本分类算法

文本分类指的是基于待分类文本集的内容，对文本集按照一定的分类体系或标准进行分类标记。一般而言，文本分类由计算机根据一定规则自动完成。

本文采用的文本分类算法属于机器学习算法，即计算机需要训练集来“学习”分类的规则。这就需要确定从训练集中抽取特征的方法和计算机学习分类规则的“学习方法”，它们就是文本分类算法中的特征抽取方法和分类器模型。

2.1.1. 特征抽取方法

在对中文文本进行分类时，很重要的步骤就是将中文文本量化为可以计算超平面距离和进行比较的数字向量。本文不对特征选择算法做深入的探究，直接采用实际生活中非常常用的 TF-IDF 算法，得到的结果是每个词在不同类别下的权重。

2.1.2. 分类器模型

本文采用了两种分类器模型进行训练，并比较分类效果。

第一种模型是朴素贝叶斯分类器模型。朴素贝叶斯模型发源于古典数学理论，是一种十分简单的分类算法，具有较为稳定的分类效率。朴素贝叶斯分类器模型的思想是对于给出的待分类项，求解在此项出现的条件下此项属于各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。朴素贝叶斯模型默认的前提假设是，在给定分类项作为前提条件时，各个属性之间相互条件独立，概率可以直接相乘。

第二种模型是笔者自己设计的应用于实际非常简单方便的“词权重和”模型。在特征抽取阶段，借助 TF-IDF 算法，可以得到每个词在各个类别下的权重，以此建立词典如表 2-1 所示，并将其作为分类器模型。

表 2-1 TF-IDF 词权重表

词	类别	权重
婚戒	婚庆	0.01334
厨房	家居	0.051711
厨房	美食	0.03472

对于给定的待分类文本，对其建立字典，字典的键值为各个类别，各键值对应的字典值均设为 0。然后对文本进行切词，取切词结果中在分类器模型中出现的词，查询其在模型各类别下的权重，然后将文本对应的字典的对应类别的值相加相应的权重值。最后，取对应字典值最大或前几大的几个字典键值，作为对待分类文本的预测结果。

2.2. 数据准备

本文实现了功能稳定的爬虫，能够从新浪微博的热门微博页面抓取各个类目下的微博，将微博条目和对应的类别保存到本地的文本文档，从而获得分类器的训练集。获取到的数据如表 2-2 所示。

表 2-2 文本分类训练数据

类别	微博内容
社会	#视野#【惊险！实拍雷电降临 火光洒满整条街】5 月 11 日傍晚，辽宁沈阳和平大街，行车记录仪捕捉到惊人一幕——天色突变，一道白光闪过，雷电劈到大街上，炸出火花，洒满整条街。L 实拍雷电劈到大街 炸出火花洒满整条街
瘦身	提臀健身法，每个动作要领都有详细的解说和动作示范， 非常全面，速度马住了！L 健身氧吧的秒拍视频
政务	国家互联网应急响应中心官方权威发布【关于防范 Windows 操作系统勒索软件 Wannacry(比特币勒索病毒)的情况通报】O 网页链接

然后，在设计文本分类器的类目时，本文基本沿用了热门微博的分类体系。新浪微博的热门微博页面共有 49 个类别，但其中的“视频”类别不适合作为短文本的类别，其中的“本地”类别获取到的都是北京本地的数据，不具有普适性，因此，最终确定分类器有以下 47 个类别，如表 2-3 所示。

表 2-3 文本分类类别表

类别编号	类别名称	类别 URL
1	社会	http://d.weibo.com/102803_ctg1_4188_-_ctg1_4188
2	国际	http://d.weibo.com/102803_ctg1_6288_-_ctg1_6288
3	科技	http://d.weibo.com/102803_ctg1_2088_-_ctg1_2088
4	科普	http://d.weibo.com/102803_ctg1_5988_-_ctg1_5988
5	数码	http://d.weibo.com/102803_ctg1_5088_-_ctg1_5088
6	财经	http://d.weibo.com/102803_ctg1_6388_-_ctg1_6388
7	股市	http://d.weibo.com/102803_ctg1_1288_-_ctg1_1288
8	明星	http://d.weibo.com/102803_ctg1_4288_-_ctg1_4288
9	综艺	http://d.weibo.com/102803_ctg1_4688_-_ctg1_4688
10	电视剧	http://d.weibo.com/102803_ctg1_2488_-_ctg1_2488
11	电影	http://d.weibo.com/102803_ctg1_3288_-_ctg1_3288
12	音乐	http://d.weibo.com/102803_ctg1_5288_-_ctg1_5288
13	汽车	http://d.weibo.com/102803_ctg1_5188_-_ctg1_5188
14	体育	http://d.weibo.com/102803_ctg1_1388_-_ctg1_1388
15	运动健身	http://d.weibo.com/102803_ctg1_4788_-_ctg1_4788
16	健康	http://d.weibo.com/102803_ctg1_2188_-_ctg1_2188
17	瘦身	http://d.weibo.com/102803_ctg1_6488_-_ctg1_6488
18	养生	http://d.weibo.com/102803_ctg1_6588_-_ctg1_6588
19	军事	http://d.weibo.com/102803_ctg1_6688_-_ctg1_6688
20	历史	http://d.weibo.com/102803_ctg1_6788_-_ctg1_6788
21	美女模特	http://d.weibo.com/102803_ctg1_2288_-_ctg1_2288
22	美图	http://d.weibo.com/102803_ctg1_4988_-_ctg1_4988
23	情感	http://d.weibo.com/102803_ctg1_1988_-_ctg1_1988
24	搞笑	http://d.weibo.com/102803_ctg1_4388_-_ctg1_4388
25	辟谣	http://d.weibo.com/102803_ctg1_6988_-_ctg1_6988
26	正能量	http://d.weibo.com/102803_ctg1_7088_-_ctg1_7088
27	政务	http://d.weibo.com/102803_ctg1_5788_-_ctg1_5788
28	游戏	http://d.weibo.com/102803_ctg1_4888_-_ctg1_4888
29	旅游	http://d.weibo.com/102803_ctg1_2588_-_ctg1_2588
30	育儿	http://d.weibo.com/102803_ctg1_3188_-_ctg1_3188
31	校园	http://d.weibo.com/102803_ctg1_1488_-_ctg1_1488
32	美食	http://d.weibo.com/102803_ctg1_2688_-_ctg1_2688
33	房产	http://d.weibo.com/102803_ctg1_5588_-_ctg1_5588
34	家居	http://d.weibo.com/102803_ctg1_5888_-_ctg1_5888
35	星座	http://d.weibo.com/102803_ctg1_1688_-_ctg1_1688
36	读书	http://d.weibo.com/102803_ctg1_4588_-_ctg1_4588
37	三农	http://d.weibo.com/102803_ctg1_7188_-_ctg1_7188
38	设计	http://d.weibo.com/102803_ctg1_5388_-_ctg1_5388
39	艺术	http://d.weibo.com/102803_ctg1_5488_-_ctg1_5488

40	时尚	http://d.weibo.com/102803_ctg1_4488_-_ctg1_4488
41	美妆	http://d.weibo.com/102803_ctg1_1588_-_ctg1_1588
42	动漫	http://d.weibo.com/102803_ctg1_2388_-_ctg1_2388
43	宗教	http://d.weibo.com/102803_ctg1_5688_-_ctg1_5688
44	萌宠	http://d.weibo.com/102803_ctg1_2788_-_ctg1_2788
45	婚庆	http://d.weibo.com/102803_ctg1_1788_-_ctg1_1788
46	法律	http://d.weibo.com/102803_ctg1_7388_-_ctg1_7388
47	收藏	http://d.weibo.com/102803_ctg1_8189_-_ctg1_8189

2.3. 模型训练与测试

在读入训练数据后，笔者取了各个类别下 80%的数据构成训练集，20%的数据构成测试集，这样就得到了有类别标注的训练数据和有标准答案的测试数据，并且保证了测试数据不在训练数据中。然后，分别用本章第一节中介绍的朴素贝叶斯模型和笔者自己设计的“词权重和”模型训练文本分类器，再对测试集进行特征抽取和类别预测，计算两种模型类别预测的正确率。两种分类器模型所采用的训练数据和测试数据相同，训练数据的数据量为 29340 条，测试数据的数据量为 7335 条，得到的结果格式相同，如下表 2-4 所示。

表 2-4 预测分类结果表

微博内容	预测类别	实际类别	正误
{直发板短发卷发}用直发板完成的蓬松短发卷发~ L 艾薇 Daily 的秒拍视频 CR.INS	美妆	美妆	1
#犬夜叉# 女子本弱，为母则刚！	育儿	动漫	0
聊天做游戏，就是好欢喜 请刷一波 666，0 今晚 20:30@浙江卫视中国蓝 继续欢乐颂～#张晓谦欢乐颂 2#	电视剧	电视剧	1

最终，经过数据测试，朴素贝叶斯模型分类的正确率为 79.45%，“词权重和”模型预测的正确率为 68.19%。

2.4. 微博热点分类方案

单从分类器模型训练和测试的结果来看，朴素贝叶斯模型的分类效果要优于笔者自己提出的“词权重和”模型。但与此同时，笔者发现，从人工识别的角度，很多测试数据（微博文本）本身是含有多个类别的成分的。例如“刘涛主演欢乐颂”，既可以被分类为“明星”，也可以被分类为“电视剧”。但朴素贝叶斯模型只返回一个可能性最大的分类结果“电视剧”，而“词权重和”模型可

以将文本在不同类别下的词权重和进行排序，按词权重和从大到小的顺序输出多个分类结果。例如，同样对于文本内容“刘涛主演欢乐颂”，“词权重和”模型的输出结果为：“电视剧：0.45，电影：0.09，美女模特：0.09，明星：0.06”，可以出这句话的多种类别成分都在分类结果中得到了体现。

基于微博内容本身可能存在的混合分类情况，笔者决定将两种分类器模型都应用于系统中，采用贝叶斯模型训练文本分类器，为系统提供自动分类功能并返回唯一确定类别，同时采用“词权重和”模型具体分析文本所包含的所有类别成分，使系统在分析属于混合分类的文本内容时也能有不错的表现。

在向系统中集成我自己设计的“词权重和”模型分析文本的类别成分时，为了尽量避免文本内容中的个别词对文本内容类别成分带来的影响，笔者改变了模型的返回结果，从返回所有词权重和不为 0 的类别类别，改为返回词权重和大于最大值的 2/3 的类别列表，从而保证在对文本内容进行分析时，系统输出的类别成分标签与原文本内容关联性是较高的。

3. 系统设计

在本章节中，笔者对微博实时热点分析系统进行了详细的设计，具体包括系统功能设计、系统框架设计、系统的数据库结构设计和系统模块的设计。

3.1. 系统功能

本系统定位于对微博热搜排行榜中的热搜词条进行挖掘和分析，并将分析结果可视化地呈现给用户，因此，在设计系统框架、数据库结构和系统模块时，需要考虑以下功能的集成：

- 微博热点（热搜词条）自动分类

该功能的实现首先需获取到热搜词条对应的微博搜索结果，并将这些与词条相关的微博搜索结果作为原热搜词条语义和特征的扩展，将它们结合为一个数据量更大的中文文本个体，然后需要应用本文第 2 章的研究成果——中文文本分类器，这就要求通过机器学习得到的分类器模型能够以功能模块的形式集成在系统中并被随时调用。

- 在线数据分析平台

该功能同样是基于第 2 章实现的文本分类器，不同之处在于，该功能是向用户提供数据分析的端口，待分析数据由用户提供，数据量往往只有寥寥几十个字，无法像分类器对系统内的微博热点分类时那样有现成的数据进行特征扩展。因此，需要优化文本分类器在处理短文本时的表现，并尝试通过微博搜索、百度搜索关键词的方式进行特征扩展，优化数据分析平台的表现。

- 热点或微博内容抽取

本系统数据源全部来自于新浪微博，因此需要设计相应的爬虫，能够从微博获取数据并抽取所需字段，然后存入数据库的对应数据表中，供系统查询和分析。在实际操作中，从微博获取数据可能会存在一系列的问题，如爬虫抓取频率过高会被封号封 IP，数据库数据太多会导致查询速度缓慢等，这些都是后续实现时需要考虑的。

- 数据驱动的热点随时间变化趋势的图表

想要实现分析结果的可视化，合适的数据可视化工具是必要的。需要注意

的是数据可视化工具与 HTML 前端、Python 后端交互的可行性：当待展示的数据从 Python 后端传入数据可视化工具中，应做怎样的处理，才能使其在 HTML 前端以图表的形式可视化的展现。

- 数据驱动的时间线视图

每个热点都有自己的时效性，对单个热点，可以以热度随时间变化的趋势图展现其时效性的变化规律，但对系统中的所有热点，就需要设计一个合适的页面，向用户展示一条连贯的时间线，时间线上的每个时间点发生了什么热点事件，从而达到舆情监测和提高用户体验的目的。

- 数据驱动的文本词云图

对于文本数据，图像化、可视化的文本词云图呈现方式能够提高系统的用户体验。这就要求系统在 Python 后端对文本内容进行切词和词频统计，在去除停用词后，系统会调用词云工具包，根据词频调整每个词的大小，生成词云图并保存到系统图片目录，然后在前端展示给用户，实现文本可视化。

- 热点、微博内容按关键词检索功能和分面组配的筛选功能

当系统中的数据足够多时，分面组配功能和检索功能就成为了系统的标配，在具体的实现过程中，应考虑如何让系统前端高效地与数据库发生查询交互，让用户可以快速找到自己需要的结果。

3.2. 系统框架

3.2.1. 整体框架

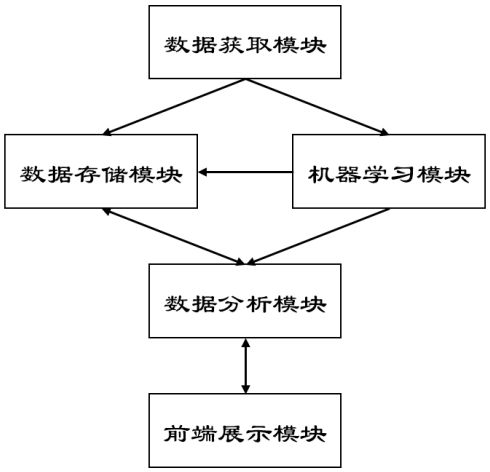

图 3-1 系统整体框架图

系统的整体框架结构如图 3-1 所示。其中箭头的走向代表系统内数据的传输方向，单箭头表示数据单向传输，双箭头表示数据双向交互。

3.2.2. 具体框架

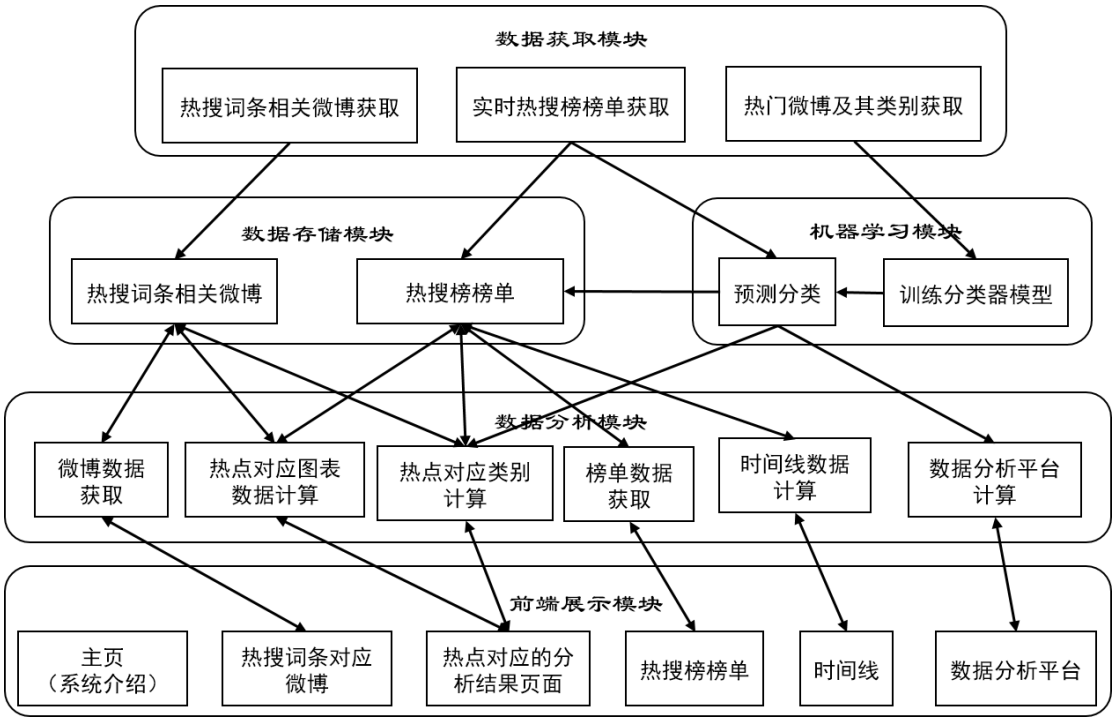

图 3-2 系统具体框架图

系统的具体框架如图 3-2 所示，其中箭头的走向代表系统内数据的传输方向，单箭头表示数据单向传输，双箭头表示数据双向交互。各个模块的具体功能描述参见本章第 4 节。

3.3. 数据库结构

本系统的“weibo”数据库中包含两个数据表“realtimehot”和“weiboitem”，分别存储微博实时热搜榜的榜单数据和榜单中的热搜词条所对应的微博搜索结果中的微博条目。两个表的字段及含义见表 3-1 和表 3-2。

表 3-1 数据库 realtimehot 表（实时热搜榜榜单表）

字段	类型	描述
id	int(11)	唯一标识编码
rank	int(11)	榜单排名
link	text	搜索结果链接

type	varchar(255)	类别
keyword	text	热搜词条
freq	int(11)	搜索指数
date	datetime	获取日期

表 3-2 数据库 weiboitem 表（热搜词条对应微博条目表）

字段	类型	描述
id	int(11)	唯一标识编码
hotspot	varchar(255)	热搜词条
content	text	对应微博内容

以上两个表囊括了实现系统功能所需的绝大多数数据内容，两个表的数据通过“realtimehot”表的“keyword”字段和“weiboitem”表的“hotspot”字段关联，使得一个热搜词条可以对应多条相关微博，实现了对热搜词条的语义扩充和特征扩展。其他未被数据库存储的数据包括文本分类器模型的训练数据（以文本文档的形式保存在系统文件中）和训练后的分类器模型（以文件的形式保存在系统文件中）。

3.4. 系统模块

根据本文第 3 章第 2 节的系统框架设计方案，本系统包括数据获取模块、数据存储模块、机器学习模块、数据分析模块和前端展示模块，共计 5 个模块。各个模块之间通过数据交互实现系统功能。

3.4.1. 数据获取模块

本系统数据源全部来自于新浪微博，因此需要设计相应的爬虫，能够从微博获取数据并抽取所需字段，然后存入数据库的对应数据表中，供系统查询和分析。数据获取模块需要获取三类数据：微博实时热搜榜的榜单，实时热搜榜榜单上的词条分别对应的搜索结果页面中的微博条目，热门微博页面的微博及其对应的类别。其中前两类数据直接传入数据存储模块的数据库中保存，最后一类数据则以文本文档的形式保存到本地，作为机器学习模块的训练集。这三

类数据都需要部署爬虫来实现。

本系统需要获取连续的时间序列中的实时热搜排行榜榜单，用来分析热搜词条随时间变化的规律，生成图表和时间线，因此抓取热搜榜榜单的爬虫应确保能够不间断地定时运行，并将抓取到的榜单按照对应字段存入系统数据库的“realtimehot”表中，目前设置的抓取间隔为 10 分钟，每次抓取到的榜单含有 50 个热搜词条。

在设计爬虫实现抓取热搜榜榜单上的词条对应的搜索结果页面的功能时，笔者遇到的主要问题是，榜单每 10 分钟就会自动获取一次，每次的榜单含有 50 个词条，分别抓取这 50 个词条对应的搜索结果就需要让爬虫打开 50 个页面，而新浪微博的反爬虫保护策略是能够检测到连续的抓取行为并封锁 IP 的，因此必须在抓取页面时设置时间间隔，经笔者测试，想要不被封 IP，所需的时间间隔至少为 30 秒，这样抓取 50 个页面的时间开销就达到了 1500 秒以上，约 25 分钟，远大于榜单 10 分钟一次的刷新频率。因此，笔者优化了该爬虫的抓取步骤，设计了如下解决方案：

- 从数据库中获取到所有还没有对应微博条目的热搜词条及其搜索结果链接；
- 以热搜词条为字典键值、搜索结果链接为字典的值建立字典，相同的热搜词条对应同样的搜索结果链接；
- 对字典中的每个热搜词条，抓取热搜词条对应链接页面的所有微博条目；
- 将抓取结果按照热搜词条的内容，执行多次插入数据库“weiboitem”表的 SQL 语句，“hotspot”字段为热搜词条的内容，对应实时热搜榜的实时热搜词条，“content”字段为抓取到的微博。

通过以上解决方案，大大减少了爬虫请求 URL 的次数，增加了每次请求 URL 获取到数据后执行 SQL 插入数据库语句的次数。这样，只要热搜榜榜单不会突然出现大量的从没出现过的新热搜词条，那么在每次获取到的实时热搜榜的 50 个热搜词条中，需要抓取微博条目的新热搜词条就只有几个到十几个，微博条目的获取速度就能够跟上热搜词条榜单 10 分钟一次的更新速度，从而保障了数据的实时性和系统的稳定性。

关于热门微博页面的微博及其对应的分类，获取到的数据是为自动文本分类模型提供训练数据，因此并不需要实时抓取，部署在本地即可。抓取到的数据内容如表 3-3 所示，格式为文本文档。

表 3-3 抓取到的热门微博数据表

类别	微博内容
社会	#视野#【惊险！实拍雷电降临 火光洒满整条街】5 月 11 日傍晚，辽宁沈阳和平大街，行车记录仪捕捉到惊人一幕——天色突变，一道白光闪过，雷电劈到大街上，炸出火花，洒满整条街。L 实拍雷电劈到大街 炸出火花洒满整条街
瘦身	提臀健身法，每个动作要领都有详细的解说和动作示范， 非常全面，速度马住了！L 健身氧吧的秒拍视频
政务	国家互联网应急响应中心官方权威发布【关于防范 Windows 操作系统勒索软件 Wannacry(比特币勒索病毒)的情况通报】O 网页链接

3.4.2. 数据存储模块

本系统的数据存储模块即为“weibo”数据库，数据库类型为 MySQL，数据库结构在本章第 3 节中已经介绍过了。本模块的输入数据为两种 SQL 语句：第一种是 INSERT（插入）语句，输入者为数据获取模块，插入的内容为数据获取模块获取到的微博实时热搜榜榜单和热搜词条对应的搜索结果中的微博条目，分别存入“realtimehot”表和“weiboitem”表中；第二种是 SELECT（查询）语句，输入者为数据分析模块，查询内容为按条件选取热搜榜单表中的条目和微博表中的条目。本模块的输出数据为两个数据表的条目，输出对象是系统的数据分析模块，在数据存储模块接收到数据分析模块输入的 SELECT（查询）语句之后，会输出相应的 SQL 查询结果，供数据分析模块进行数据分析。

3.4.3. 机器学习模块

本模块在本文第 2 章中已经有了详细的介绍。机器学习模块与系统的其他模块相比相对独立，不需要与数据存储模块的数据库发生交互，在经过一次机器学习训练后，所得的分类器模型可以一直为系统所用，直到系统需要更新机器学习的训练集或算法。

本模块的输入数据分为两类：第一类是数据获取模块获取到的热门微博页面的微博及其对应的分类，本模块会以该输入数据为训练集，通过机器学习的工具包和算法，执行分类器的训练，并将训练后的模型保存为系统内的本地文件；第二类是数据分析模块传入的待分类文本，本模块会按照特定的特征抽取算法将输入的文本处理为向量，然后读入本地保存的分类器模型计算其所属分类，然后向数据分析模块输出文本所属类别。

本模块的输出数据也分为两类：第一类是模块在读入训练集完成模型训练后输出的分类器模型，包括朴素贝叶斯分类器模型，文本向量转换模型，以及 TF-IDF 词权重模型；第二类是模块读入待分析的文本后返回的分析结果，包括基于朴素贝叶斯模型的文本所属类别，基于 TF-IDF 词权重模型的文本包含的类别成分列表。

3.4.4. 数据分析模块

数据分析模块是整个系统的后端的核心，用户在系统前端看到的所有数据都来自于数据分析模块。数据分析模块会根据前端的请求，对数据存储模块输入 SQL 查询语句，得到相应的查询结果，然后按照需求进行特定的数据分析，并将分析后的数据传回系统前端进行表格、折线图、雷达图、折线图、词云图等可视化的展示。

简单的数据分析操作是从数据库的数据表查询出全部数据并传入前端。复杂的数据分析操作包括根据前端的分面组配条件，生成复杂 SQL 查询语句并从数据库取出相应数据，以及调用机器学习模块的算法计算文本分类等。

3.4.5. 前端展示模块

前端展示模块是系统的门面，用户能直接接触到的页面都来自前端展示模块，本模块的设计直接影响用户体验。考虑到系统实现的各种功能，将前端展示模块划分为以下 5 个页面：

- **主页（系统介绍）** 系统内相对 URL：“/”

主页上展示的是笔者对系统内容的介绍，包括系统框架、系统简介、系统模块、系统功能、特别致谢和用到的工具等内容，此外，笔者将本文的电子版放置在了系统主页上。

- **实时热点** 系统内相对 URL：“/hotspot”

实时热点页面上展示的是实时的热搜词条排行榜。在主页上用户可以查看历史数据，搜索关键词，按词条排名、词条名称、词条热度进行分面组配式的筛选，点击词条进入微博的相应搜索结果页面，点击词条后的“在线分析”链接进入对该词条的分析结果页面。

- **热点相关微博** 系统内相对 URL：“/weibo”

热点相关微博页面展示的是按时间倒序排列的热搜词条及其对应的搜索结果中的微博，单个词条会对应多个微博条目。在该页面用户可以搜索关键词，点击词条进入微博的相应搜索结果页面，点击微博条目后的“在线分析”链接进入对该词条的分析结果页面。

- **时间线** 系统内相对 URL：“/timeline”

时间线页面展示的是时间轴上按时间倒序排列的热点事件，系统记录的热点事件为每个热搜词条第一次出现在系统中的时间，以及热搜词条在系统中达到热度峰值的时间。用户可以点击热搜词条进入对该词条的分析结果页面。

- **在线数据分析平台** 系统内相对 URL：“/data”

数据分析平台上展示的是一个简单的文本输入框。用户可以在文本框内输入想要进行数据分析的文本，点击“分析”按钮进行分析。分析结果为该文本所包含的关键词和所属类别，系统内的分类体系为新浪微博的热门微博页面的 47 个类别。

- **对单个热搜词条的分析结果页面** 系统内相对 URL：“/热搜词条”

该页面展示的是热搜词条的热度随时间的变化规律，该热搜词条对应的类别和微博条目，以及该热搜词条包含的关键词词频统计结果（已通过添加对应微博条目进行了文本扩充），还会基于词频自动生成热搜词条对应的词云图。由于有多个页面可以链接到本页面，因此本页面提供了“返回上一页”链接，让用户可以快速返回刚刚所在的页面。

4. 系统实现

根据第 3 章的系统设计方案，笔者采用 bootstrap 前端框架和 web.py 后端框架实现了整个系统。在本章节中，笔者介绍了实现后的系统，包括系统链接、部署环境、系统页面展示等。

4.1. 系统链接

本系统已部署在阿里云服务器，并绑定了国际域名，用户可通过网址 http://weibohotanalyze.online 进行在线访问。服务器租赁有效期为 2017 年 7 月 1 日，在有效期前，系统会在服务器端自动维护，以保障系统数据的实时性和连续性。

4.2. 部署环境

本系统可以部署在 Windows 或 Linux 系统环境中（实际部署的系统环境为 Windows Server 2008），需要部署 MySQL 数据库和 Python2.7 编程环境，需要对 Python 安装 sklearn 机器学习工具包，jieba 中文分词工具包，wordcloud 词云包，web.py 网站框架包，硬盘空间越大越好，以存储海量数据库数据。

4.3. 系统页面展示

本系统的前端页面包括主页（系统介绍）、实时热点页面、热点相关微博页面、时间线页面、在线数据分析平台页面和对单个热搜词条的分析结果页面共计 6 类页面，每类页面的链接和实现效果在本章节中有具体介绍。

4.3.1. 主页（系统介绍）

本页面是整个系统的首页，最终实现效果如图 4-1 所示。本页面的网址链接为 http://weibohotanalyze.online。

图 4-1 主页（系统介绍）页面

主页上展示的是笔者对系统内容的介绍，包括系统框架、系统简介、系统模块、系统功能、特别致谢和用到的工具等内容，此外，笔者将本文的电子版放置在了系统主页上。

4.3.2. 实时热点

本页面展示的是新浪微博的实时热搜榜单，最终实现效果如图 4-2 所示。本页面的网址链接为：http://weibohotanalyze.online/hotspot。

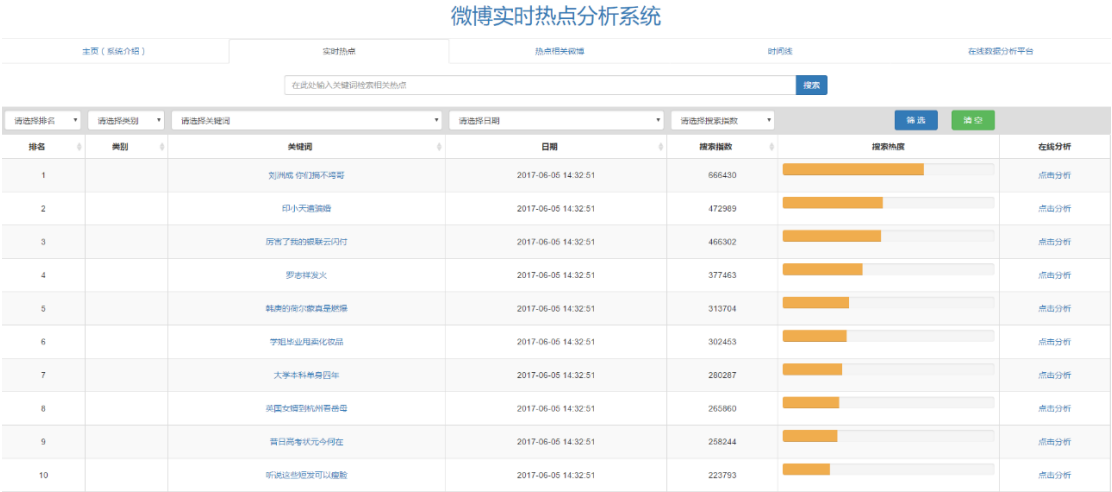

图 4-2 实时热点页面

在实时热点页面上用户可以查看热搜词条的历史数据，搜索热点关键词，以及按词条排名、类别、词条名称、词条热度进行分面组配式的筛选，点击词条会在新标签页打开微博的相应搜索结果页面，点击词条后的“在线分析”链接进入对该词条的分析结果页面。

4.3.3. 热点相关微博

本页面展示的是按热度倒序和时间倒序排列的实时热点对应的相关微博条目（来自微博搜索热搜词条的搜索结果页面），每个热搜词条会对应多条微博，最终实现效果如图 4-3 所示。本页面的网址链接为：http://weibohotanalyze.online/weibo。

图 4-3 热点相关微博页面

在热点相关微博页面，用户可以搜索关键词，点击词条会在新标签页打开微博的相应搜索结果页面，点击微博条目后的“在线分析”链接进入对该词条的分析结果页面。

4.3.4. 时间线

本页面展示的是时间轴上按时间倒序排列的热点事件，最终实现效果如图 4-4 所示。本页面的网址链接为：http://weibohotanalyze.online/timeline。

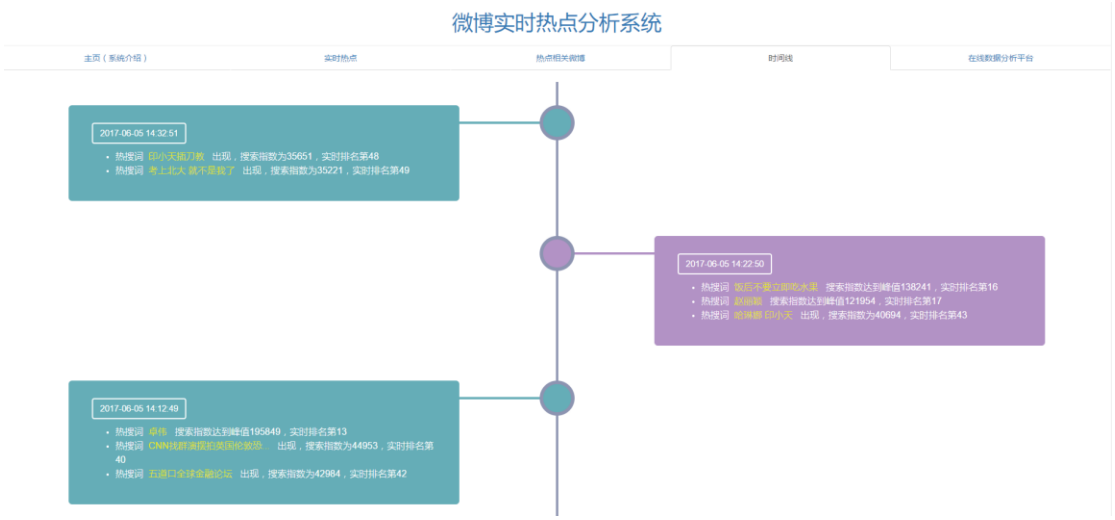

图 4-4 时间线页面

时间线页面记录的热点事件为每个热搜词条第一次出现在系统中的时间，以及热搜词条在系统中达到热度峰值的时间。用户可以点击热搜词条进入对该词条的分析结果页面。

4.3.5. 在线数据分析平台

本页面是系统的数据分析平台，支持用户输入自定义文本进行数据分析，最终实现效果如图 4-5 所示。本页面的网址链接为：http://weibohotanalyze.online/data。

图 4-5 在线数据分析平台页面

数据分析平台上展示的是一个简单的文本输入框。用户可以在文本框内输入想要进行数据分析的文本，点击“分析”按钮进行分析。分析结果为该文本所包含的关键词和所属类别，系统内的分类体系为新浪微博的热门微博页面的 47 个类别（详见本文第 2 章第 2 节“数据准备”）。

4.3.6. 对单个热搜词条的分析结果页面

本页面展示的是时间轴上按时间倒序排列的热点事件，最终实现效果如图 4-6 所示。本页面的网址链接为：http://weibohotanalyze.online/热搜词条，其中的“热搜词条”代表系统中的热搜词条内容而非固定链接。例如，图 4-6 对应的网址链接为：http://weibohotanalyze.online/昔日高考状元今何在。

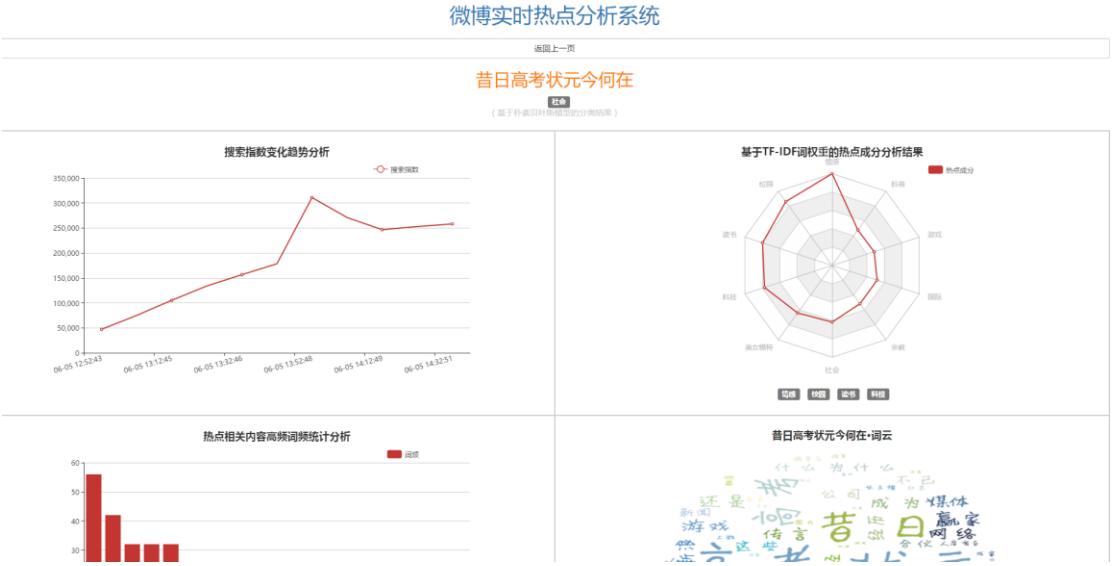

图 4-6 对单个热搜词条的分析结果页面

在对单个热搜词条的分析结果页面中，展示了热搜词条的热度随时间的变化规律，该热搜词条对应的类别和微博条目，以及该热搜词条包含的关键词词频统计结果（已通过添加对应微博条目进行了文本扩充），还会基于词频自动生成热搜词条对应的词云图。由于有多个页面可以链接到本页面，因此本页面提供了“返回上一页”链接，让用户可以快速返回刚刚所在的页面。

5. 总结与展望

5.1. 总结

本文设计并实现了一个微博实时热点分析系统，包括数据获取模块、数据存储模块、机器学习模块、数据分析模块和前端展示模块共计五大模块。本系统能够根据微博热搜词条随时间的变化情况，对热搜词条进行舆情分析，并根据热搜词条对应的微博搜索结果，对热搜词条进行文本扩充和特征扩展，进而实现对热搜词条的自动分类。系统的机器学习模块本身也独立出来，作为一项系统功能，提供文本输入端口，对用户输入的文本进行切词和分析。

本文研究工作的重点是自动文本分类器的构建和训练。本文从新浪微博的热门微博页面获取分门别类的数据，并沿用热门微博的分类类目构建训练集，利用 TF-IDF 算法计算词权重，提取文本特征，尝试采用“词权重和”模型、朴素贝叶斯模型等不同的机器学习算法构建分类器。最终，笔者根据数据实验结果（朴素贝叶斯模型分类正确率优于“词权重和”模型，“词权重和”模型输出的类别数目多于朴素贝叶斯模型），决定将两种分类器模型都应用于系统中，采用贝叶斯模型训练文本分类器，为系统提供自动分类功能并返回唯一确定类别，同时采用“词权重和”模型具体分析文本所包含的所有类别成分，使系统在分析属于混合分类的文本内容时也能有不错的表现。

5.2. 研究不足与改进

在未来的研究工作中，本文所实现的微博实时热点分析系统仍有很大的改进空间。数据采集模块可以优化爬虫的算法或采用分布式爬虫，突破反爬虫限制，提高系统数据的实时性。数据分析模块可以强化数据分析的广度和深度，从而使系统功能更加强大。机器学习模块的算法还有很大的优化空间，可以作为一个单独的研究课题展开更加深入的讨论，提高短文本分类器的分类效果。数据存储模块可以优化数据库的数据表结构，或采用 NoSQL 非关系型数据库，加快数据查询的速度。前端展示模块还可以采用更为丰富的图表形式进行数据的可视化展示，提高用户体验。

参考文献

[1] 知徽传播分析:使用说明[EB/OL]. http://www.weiboreach.com/explain.html, 2017.05.

[2] Ren D, Zhang X, Wang Z, et al. WeiboEvents: A Crowd Sourcing Weibo Visual Analytic System[C]// IEEE Pacific Visualization Symposium. IEEE Computer Society, 2014:330-334.

[3] Martino_Affanb F D, Loia V, Sessa S. Extended fuzzy C-means clustering algorithm for hotspot events in spatial analysis[M]. IOS Press, 2008.

[4] Senaratne H, Ring A, Schreck T, et al. Moving on Twitter: using episodic hotspot and drift analysis to detect and characterise spatial trajectories[C]// ACM Sigspatial International Workshop on Location-Based Social Networks. ACM, 2014:23-30.

[5] 邹盼湘. 网络舆情热点提取与分析[D]. 华南理工大学, 2015.

[6] 张寿华, 丛帅, 尚开雨,等. 网络舆情追踪中热点关键词的提取[J]. 河北大学学报自然科学版, 2012, 32(3):311-315.

[7] Sebastiani F. Machine learning in automated text categorization[J]. Acm Computing Surveys, 2001, 34(1):1-47.

[8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management, 1988, 24(5):513-523.

[9] Sriram B, Fuhry D, Demir E, et al. Short text classification in twitter to improve information filtering[C]// Proceeding of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July. DBLP, 2010:841-842.

[10] Sankaranarayanan J, Samet H, Teitler B E, et al. TwitterStand: news in tweets[C]// ACM Sigspatial International Conference on Advances in Geographic Information Systems. ACM, 2009:42-51.

[11] 王细薇, 樊兴华, 赵军. 一种基于特征扩展的中文短文本分类方法[J]. 计算机应用, 2009, 29(3):843-845.

[12]Xue B, Fu C, Zhan S. A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec[C]// IEEE International Congress on Big Data. IEEE, 2014:358-363.

[13]李传扬. 微博分析系统的设计与实现[D]. 北京邮电大学, 2015.

[14]张洋, 何楚杰, 段俊文,等. 微博舆情热点分析系统设计研究[J]. 信息网络安全, 2012(9):60-64.

[15]王宇. 基于 TFIDF 的文本分类算法研究[D]. 郑州大学, 2006.

[16]王政霄. 基于微博的热点事件挖掘与情感分析[D]. 上海交通大学, 2013.

[17]丘恩. Python 核心编程: 第二版[M]. 人民邮电出版社, 2008.

[18]HarryJ.W.Percival. Python Web 开发:测试驱动方法[M]. 人民邮电出版社, 2015.

[19]刘长龙. Python 高效开发实战[M]. 中国工信出版集团, 2016.

[20]Bootstrap-Table:文档[EB/OL]. http://bootstrap-table.wenzhixin.net.cn/zh-cn/documentation/, 2017.05.

[21]Echarts:配置项[EB/OL]. http://echarts.baidu.com/option.html, 2017.04.

[22]菜鸟教程:Bootstrap[EB/OL]. http://www.runoob.com/bootstrap/bootstrap-tutorial.html, 2017.05.

致　谢

经过一学期的努力我终于完成了自己的毕业论文，这也意味着我在北京大学的求学生涯即将画上一个圆满的句号。本文最终实现的系统历经了整整十次版本更新，这带给我的工作量是前所未有的。一年一度毕业季，北京大学的一切都让我感到珍惜与怀念。借此论文完成之际，我想向在我撰写论文和编写系统的过程中给予我帮助的人们致谢。

首先，我衷心地感谢我的导师王继民教授。作为导师，他向我提供了学习交流的组会平台，给了我很多锻炼能力的机会。作为任课教师，他激发了我学习数据挖掘的热情，让我学到了很多新的知识，同时确立了自己的毕业论文方向和未来发展方向。在指导我完成论文期间，王老师为我答疑解惑，引领我将论文向格式规范、内容严谨的方向修改，这些都花费了王老师很多的宝贵时间，在此向导师表示衷心地感谢！

此外，我还要感谢北京大学信息管理系，这里有学术水平高超而且处处为学生着想的老师们，有来自五湖四海的可爱的同学们。在这里生活的四年，我就像待在一个温暖的大家庭里一样，和同学们一起过了很多节，参加了很多活动。在信管求学的四年，我学到了很多知识和技能，让我最终有能力实现本文设计的系统，体会到了无与伦比的成就感。

最后，我要感谢国内外程序开发团队的编程大神们，你们编写的开源工具包让我能够方便地在自己的系统中集成多种多样的酷炫的功能，你们是我未来努力的方向。

本文参考了大量的期刊与书籍，难免有所疏漏，不能一一注明，敬请原谅并向所有作者和刊物表示诚挚的谢意。同时，本人学术水平有限，难免有纰漏之处，恳请各位老师不吝赐教！