

# 广州中医药大学

## 医学信息工程学院

### 本科毕业论文

题    目： 基于 K-means 算法的校园微博热点话题  
发现系统

姓    名： 郭伟匡

学    号： 2014081029

专业年级： 2014 级

指导老师：

指导教师单位：

论文提交日期：      年    月    日

论文答辩日期：      年    月    日



# 广州中医药大学学位论文原创性声明

本人郑重声明：所呈交的学术论文，是个人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人或集体，均已在文中以明确方式标明。本人完全意识到此声明的法律后果由本人承担。

学位论文作者签名：

签字日期：



## 摘 要

微博由于其“短平快”的信息生产能力和快速传播能力，已经广泛流行于高校学生的日常生活中。但微博上的负面舆情信息给社会、学校和个人带来巨大的危害。由于微博的多而快特点，无法依赖人工对相关信息进行收集、筛选和发掘热点话题。因此研究并开发校园微博热点话题发现系统，对高校舆情工作有重要的意义。

本文从微博独有的短文本特征及国内外相关微博研究出发，通过对校园微博进行分类处理后使用 K-means 聚类算法对校园微博短文本聚类，并改进热度计算公式，通过话题热度提取校园微博热点话题，实现对校园微博热点话题的监控。本文通过几个模块设计并实现了校园微博热点话题发现系统，包括微博数据爬取模块、微博数据预处理模块、微博热点话题分析模块、微博热点话题展示模块等模块。最后以广州中医药大学的生活类微博—广中医 I 栋为研究对象，对校园微博各模块功能及相关技术进行介绍，并对相关模块进行测试验证，分析校园微博热点话题特点，总结系统的优点和不足，提出下一步改进的设想。

**关键词：** 校园微博    K-means    热点话题



## ABSTRACT

Because of its "short and fast" information production capability and rapid dissemination capability, MicroBlog has become widely popular in the daily life of college students. However, the negative public sentiment information on microblog has brought great harm to society, schools and individuals. Due to the multiple and fast characteristics of microblog, it is impossible to rely on manual collection of relevant information to screen and explore hot topics. Therefore, researching and developing hotspot discovery system on campus microblogs is of great significance to the public opinion work in colleges and universities.

This article starts with the unique short text feature of microblog and related microblog studies at home and abroad. After classifying campus microblogs, we use K-means clustering algorithm to cluster short texts on campus microblogs and improve heat calculation formulas. Through the hot topic of campus microblog hot topic extraction, to achieve the monitoring of campus microblogging hot topics. This paper designs and implements a campus microblog hot topic discovery system through several modules, including microblog data crawling module, microblog data preprocessing module, microblog hot topic analysis module, and microblog hot topic display module. The University of Medicine's Life Microblog –Guangzhongyi I dong is the subject of the study. It introduces the functions and related technologies of the campus microblog modules, tests and verifies the relevant modules, analyzes the characteristics of the campus microblogging hot topics, and summarizes the advantages and disadvantages of the system. Put forward the idea of further improvement.

**Keyword:** Campus Micro-Blog K-means Hot topic detection





## 目 录

摘 要 .....	I
ABSTRACT .....	III
第 1 章 绪论 .....	1
1.1 国内外研究现状与意义 .....	1
1.2 本文创新点 .....	2
1.3 论文写作思路 .....	2
第 2 章 相关技术介绍 .....	5
2.1 网络爬虫技术 .....	5
2.2 中文分词技术 .....	5
2.3 特征选择及权重计算 .....	6
2.3.1 特征选择 .....	6
2.3.2 特征权重计算 .....	7
2.4 文本表示 .....	8
2.4.1 布尔模型 .....	8
2.4.2 概率模型 .....	9
2.4.3 向量空间模型 .....	9
2.5 文本聚类算法 .....	9
2.5.1 距离算法 .....	10
2.5.2 K-means 聚类算法 .....	10
2.5.3 二分 K-means 聚类算法 .....	11
第 3 章 校园微博热点话题发现系统设计与实现 .....	13
3.1 系统设计目标及要求 .....	13
3.1.1 系统设计目标 .....	13
3.1.2 系统设计要求 .....	14
3.2 系统详细架构设计 .....	14
3.3 系统功能模块设计与实现 .....	15
3.3.1 微博数据获取模块 .....	15
3.3.2 微博文本预处理模块 .....	18
3.3.2 校园微博热点话题发现模块 .....	21
第 4 章 系统功能测试 .....	25

4.1 系统运行环境和参数 .....	25
4.2 实验数据及处理 .....	25
4.3 系统可视化界面 .....	27
4.3.1 数据获取界面 .....	27
4.3.2 热点话题排行榜 .....	27
4.3.3 热点话题热度直方图 .....	28
4.3.4 敏感词展示 .....	28
总 结 与 展 望 .....	31
参 考 文 献 .....	33
致 谢 .....	35
附 录 .....	37

## 第1章 绪论

### 1.1 国内外研究现状与意义

随着互联网的快速发展,Internet已经成为当今时代信息传播的主要途径。据中国互联网络信息中心(CNNIC)发布的第41次<sup>[1]</sup>《中国互联网络发展状况统计报告》显示我国网民达7.72亿,互联网普及率为55.8%,而手机网民规模达7.53亿,手机网民占比达97.5%,其中新浪微博月活跃用户达到3.76亿。移动互联网的快速发展,使得以新浪微博为首的社交应用平台快速成为信息传播的重要途径。然而,有利就有弊,各种网络负面舆情信息充斥在微博、微信、QQ等社交平台。对于高校而言,由于微博其具有“短平快”以及传播途径多等特点,已经成为高校学生获取并传播信息的主要途径。

微博是近几年来迅速发展的社交媒体与信息交流平台。用户可以通过该平台发布文本信息、图片、短视频等多媒体信息。因此流行于各大高校,各种校园重大事件或者突发事件都是通过该平台迅速传播出去。但由于微博的使用方式简单,任何人都可以在微博里面传播信息,其中就包括各种虚假新闻、谣言反动等信息的传播,容易造成谣言的滋生和传播,造成恶劣的社会影响。对于高校而言,对学校的声誉造成影响。校园微博作为高校学生传播信息的主要渠道,是热点话题发现必不可少的重要环节。通过发现微博热点话题,及时掌握校园中正在传播或议论的热点话题,并进行正确的舆论引导,可以最大限度控制不良话题的发展。因此,设计实现一套校园微博热点话题发现系统是有意义的,可以作为以后高校舆情管理人员的重要监控工具。

国内外很早就已经有学者研究微博热点话题相关内容。在国外,Jing Guo等人提出的Frequent Pattern stream mining算法<sup>[2]</sup>,DARPA发起的话题发现与追踪(TDT)项目<sup>[3]</sup>,Salton和Wong等人提出的向量空间模型,实现了将文本转化成空间向量的处理<sup>[4]</sup>,Cataldi M等人基于时间和社会术语的Twitter热点话题检测<sup>[5]</sup>,F de Villiers等人提出使用无监督聚类算法构造基于主题的Twitter列表,并使用TF-IDF等相

似度度量算法来评估K-means和AP聚类算法聚类后的结果,通过LDA生成Twitter主题列表<sup>[6]</sup>。在国内,张东霞基于微博热点话题基础下通过改进相关聚类算法实现一套舆情监控系统<sup>[7]</sup>,陈彦舟和曹金璇基于微博大数据进行挖掘、分析,实现对舆情热点话题的发现及追踪<sup>[8]</sup>,张亚男基于LDA模型对文本建模并改进K-means算法提高了热点话题发现的准确性<sup>[9]</sup>。李磊对传统的K-means和BIRCH聚类方法进行改进,并结合两种算法来发现微博热点话题<sup>[10]</sup>。孙胜平提出基于向量空间模型结合SP&HA聚类算法用来发现热点话题,并改进Single-Pass聚类算法,在最后的标题合并阶段使用改进的凝聚式层次聚类算法,提高话题发现质量<sup>[11]</sup>。

## 1.2 本文创新点

目前新浪微博平台,已经存在热点话题的发现和推荐功能,但这些功能对校园微博热点发现与舆情监控方面存在不足之处。

校园微博有针对性的舆情监控目的,而微博平台没有提供针对校园微博的热点发现,因此校园微博热点话题发现系统对于高校舆情监控有着关键的意义。本文基于K-means聚类算法实现一套校园微博热点话题发现系统,通过改进相关算法等步骤提高了获取热点话题的准确性,并实现可视化界面操作获取微博数据和热点话题等操作,方便舆情人员快速获取舆情热点。

## 1.3 论文写作思路

第一章 绪论,介绍国内外同类课题的研究现状、论文创新之处和论文的研究内容。

第二章 相关关键技术的介绍。介绍了网页爬虫技术、中文分词、特征权重计算和特征选择、文本表示模型(VSM)及K-means聚类算法等关键技术理论。

第三章 校园微博热点话题发现系统的总体设计与实现方案。详细介绍系统各功能模块的设计与实现,包括微博数据获取模块、微博数据预处理模块、微博热点话题发现模块、热点话题展示模块。

第四章 基于广州中医药大学的校园微博对系统进行功能验证和使用介绍。

结论：总结本文的工作，指出目前在研究上存在的问题和不足，并给出下一步研究的方向及相关改进算法工作。



## 第2章 相关技术介绍

### 2.1 网络爬虫技术

网络爬虫(web crawler),也叫网络蜘蛛(spider),是一种按照特定规则,用来自动浏览并获取万维网上资源的程序或脚本。网络爬虫程序也是搜索引擎中的重要组成部分。目前网络爬虫主要分为几种类型:通用网络爬虫、主题网络爬虫、增量式网络爬虫、深层网络爬虫。

### 2.2 中文分词技术

中文分词,即是 Chinese Word Segmentation,指的是将汉字序列切成一个个单独的词。由于计算机无法直接处理自然语言文本,因此需要对微博文本进行分词并构建数据模型,这也是文本挖掘的基础。

中文分词技术属于自然语言处理技术的范畴。目前现有的分词算法有:基于字符串匹配的分词、基于理解的分词和基于统计的分词三大类。

表 2-1 分词优劣对比

分词方法	字符串匹配分词	理解分词	统计分词
歧义识别	差	强	强
新词识别	差	强	强
需要词典	需要	不需要	不需要
需要词典	否	否	是
需要语料库	否	是	否
需要规则库	容易	难	一般
算法复杂性	成熟	不成熟	成熟
技术成熟度	容易	难	一般
实施难度	一般	准确	较难
分词准确性	快	慢	一般
分词速度	差	强	强

在实际使用上中文分词技术仍然存在着问题,中文文本不同于英文

文本，英文文本单词之间有空格作为分隔符，而中文只有句子，段落有分隔符，因此文本分词要复杂很多。对于微博文本而言，由于存在大量的网络新词，例如围脖、豆你玩、高富帅等新词。这些网络新词、缩略语、谐音词给分词处理带来了新的挑战。

目前中文分词技术已经取得很大的进展，典型的中文分词工具有中科院的汉语词法分析系统 ICTCLAS、SCWS、IKAnalyzer、NLPIR、jieba 等。本文的校园微博热点话题发现系统使用的是 jieba 中文分词工具。

## 2.3 特征选择及权重计算

### 2.3.1 特征选择

特征选择是为了构建模型而选择相关特征子集的过程。特征选择指的是从原始多维数据集中选取  $K$  个最有效的特征使系统达到最优化。使用特征选择可以剔除不相关或冗余特征，降低数据集合的维度，提高模型准确性，使得分析特征、训练模型的时间更短。

微博文本特征选择也是构建向量空间模型的前提条件。特征选择的原理图如图 2-3 所示，其具体步骤如下：

- (1) 产生过程，选取特征子集。
- (2) 评价函数，评价特征子集好坏。
- (3) 停止准则，当评价函数值达到一个阈值后停止搜索。
- (4) 验证过程，验证特征子集的有效性。

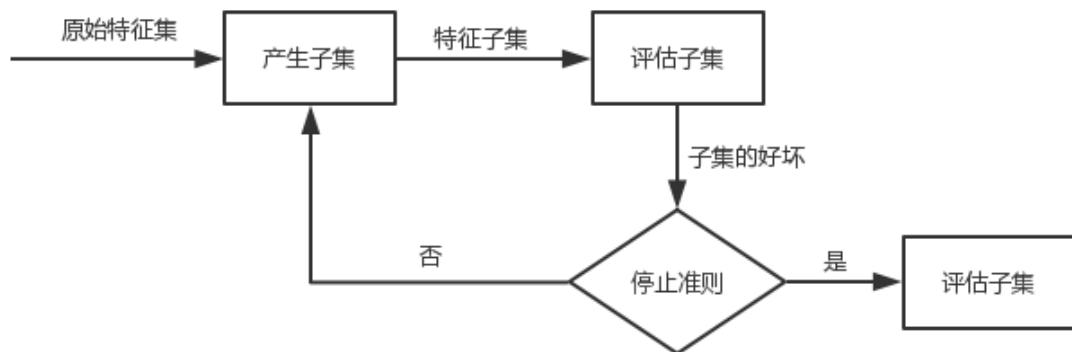


图 2-3 特征选择过程

常见的特征选择方法有文档频率、互信息、信息增益法、期望交叉熵等。



### 2.3.2 特征权重计算

特征权重计算指的是为特征空间中文本向量的每一维确定合适的数值，该数值体现了对应特征在文本中的重要程度，通常用于评估某一特征词对于文本主题的重要性。常见的做法就是统计文本的词频，根据算法计算出每一个特征合适的权重值。常见的特征权重计算方法有：布尔权重、频度权重、TF-IDF 权重等。

#### (1) 布尔权重

布尔权重也叫二值权重，是比较简单的权重计算方法。特征  $t_i$  在文本  $d_j$  中的权重为：

$$w_{ij} = \begin{cases} 0 \\ 1 \end{cases} \quad (2-1)$$

其中  $w_{ij}$  表示特征  $t_i$  的权重值，如果特征  $t_i$  在文本  $d_i$  中出现过，则值为 1，否则为 0。布尔权重比较简单，容易失去特征在文本中代表的意义。

#### (2) 频度权重

频度权重即是以特征词频作为权重，特征频度的定义：特征项  $t_i$  在文本  $d_j$  出现的次数，则特征频度权重公式如下所示：

$$w_{ij} = tf_{ij} = tf(t_i, d_j) \quad (2-2)$$

该方法的思想是：特征在文本出现次数越多，该特征就越重要，因此在该文本中权重越大。

#### (3) TF-IDF 权重计算

TF-IDF (Term Frequency-Inverse Document Frequency) 方法是文本分类中应用最多的权重计算方法，是一种用于信息检索与文本挖掘的常用加权技术，用来评估一字词对一个文件集或者一个语料库中一个文件的重要程度。TF 指的是词频，即是某一个给定词语在该文本中出现的频率。TF 公式如下：

$$\text{词频 (TF)} = \frac{\text{某个词在文章中出现次数}}{\text{文章的总词数}} \quad (2-3)$$

IDF 指的是逆文档词频，即是语料库中所有文档总数与语料库所包含该词的文档数量的比值。IDF 公式如下所：

$$\text{逆文档频率 (IDF)} = \log\left(\frac{\text{语料库文档总数}}{\text{包含该词的文档数} + 1}\right) \quad (2-4)$$

TF-IDF 就是 TF 和 IDF 的乘积，TF-IDF 公式如下：

$$\text{TF-IDF} = \text{词频 (TF)} \times \text{逆文档频率 (IDF)} \quad (2-5)$$

该算法的基本思想是：如果某个词语在一篇文档中出现次数比较多，并且该词语在整个语料库中的其它文档中出现次数比较少，那么该词语对于该篇文章的主题相关性比较大，即是通过该词来对文档进行文本分类等操作。本文使用张静<sup>[12]</sup>改进的算法来进行特征提取，该方法是将语料库所有微博文本分词后集合在一起，然后计算每个关键词出现的次数作为词频，总词数即是语料库所有文本分词后的关键词总数。改进后的算法公式如下：

$$\text{TF} = \frac{\text{词频}}{\text{总词数}} \quad (2-6)$$

其中词频为每个词出现的次数，总词数为所有分词后的关键词总数。

$$\text{IDF} = \log\left(\frac{\text{总词数}}{\text{该词频数} + 1}\right) \quad (2-7)$$

最终根据每个关键词的 TF 和 IDF 值计算出权重值来指导特征的选择。

## 2.4 文本表示

文本表示指的是将文本数据进行处理后成为程序可以处理的数据形式。目前文本表示模型有布尔模型、向量空间模型（VSM）及概率模型等。

### 2.4.1 布尔模型

布尔模型是基于特征性的严格匹配模型，根据特征是否存在文档，特征项的属性为 true 或者 false，若待建模文本出现相应特征项，则特征属性为 true，否则为 false。该模型简单实用，速度快，但文本表示很不精确，不能反映特征项对于文本的重要性。

### 2.4.2 概率模型

概率检索模型基于概率排序原理，以词与词和词与文档间的概率关系为内容进行检索。基本思想是：给定一个用户查询，若搜索系统能在搜索结果排序时按照文档和用户查询的相关性由高到低排序，那么这个搜索系统的准确性是最优的。

### 2.4.3 向量空间模型

向量空间模型（Vector Space Model）把文本内容的处理简化成向量空间的向量计算，并以空间上的相似度表达语义上的相似度。该模型已经成为文本挖掘技术使用最多的文本表示模型之一。

向量空间模型将每个文本表示成一个  $n$  维的向量，每个特征都会计算相应的权重，这些  $n$  维特征权重构成一个文本，表示该文本的主题内容。该模型通过向量表示文本内容，设语料库中  $M=\{d_1, d_2, d_3, \dots, d_n\}$ ，每个文档的特征项表示如下

$$d_i = (w_{i1}, w_{i2}, \dots, w_{in}) \quad (2-8)$$

其中  $d_i$  为第  $i$  个文本的特征权重向量， $w_{ij}$  表示在文档  $i$  中，第  $j$  个词的权重值。

## 2.5 文本聚类算法

聚类是一种数学统计分析方法，指的是按照某个特定标准把一个数据集分割成不同的类或簇，使得一个簇内的数据相似度达到最优，不同簇间的差异性达到最大，是一种无监督学习的机器学习算法。

传统的聚类算法有很多种方法，图 2-5 是各种聚类算法的类别：

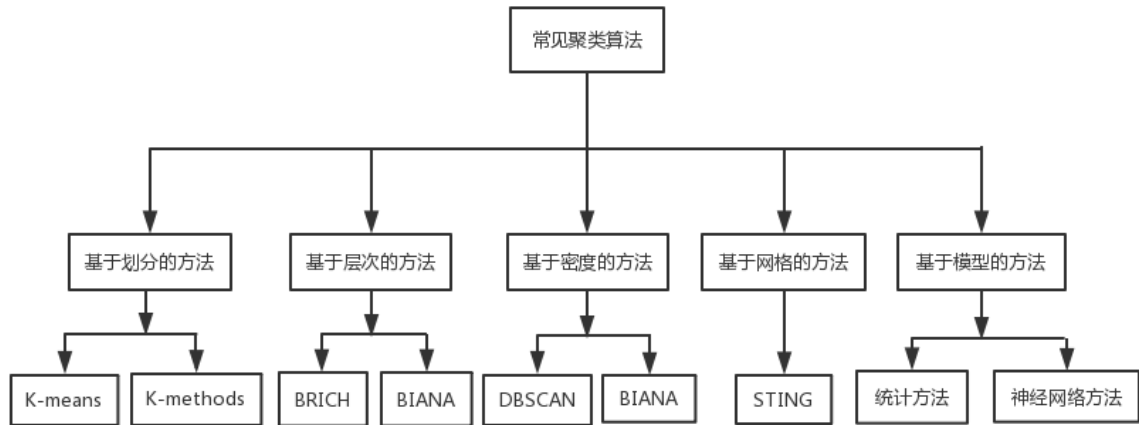


图 2-5 主要的聚类算法

各种聚类算法都有其特定的应用场景，本文使用的是基于划分的聚类算法 K-means，因此只介绍 K-means 算法以及二分 K-means 算法的相关原理。

### 2.5.1 距离算法

在介绍 K-means 聚类算法之前，先说明 K-means 算法中使用到的相关距离算法，距离算法目的是为了度量不同样本数据之间的相似度。目前常用的机器学习距离算法主要有：余弦距离、欧氏距离、曼哈顿距离等。下面介绍两种常见的距离算法的相关定义：

#### （1）余弦距离

余弦距离，即是余弦相似度，是通过两个向量的夹角余弦值来度量它们之间的相似度。夹角余弦取值范围为 $[-1, 1]$ 之间，两个向量的夹角越小，说明两个向量之间越靠近，也就是它们之间相似度越高。

#### （2）欧式距离

欧式距离，即是欧几里得距离，指的是欧几里得空间中两点间的直线距离。在欧几里得空间里，点  $x=(x_1, \dots, x_n)$  和点  $y=(y_1, \dots, y_n)$  之间的欧式距离为：

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \quad (2-9)$$

### 2.5.2 K-means 聚类算法

K-means 聚类算法是一种典型的划分聚类算法，采用距离作为相似

性的评价指标，即认为两个对象的距离越近，其相似度就越大。**K-means** 算法目的是：把  $n$  个点划分到  $k$  个聚类中，使得每个点都离它最近的簇心对应的聚类。

**K-means** 聚类算法的关键在于设置  $k$  个聚类中心。算法过程如下：

- (1) 从  $N$  条数据文档中随机选取  $k$  个数据文档作为簇心。
- (2) 对剩余的每个数据文档通过距离计算算法计算该文档到每一个簇心的距离，并将该文档归类到距离最近的簇心类中。
- (3) 根据簇内数据，重新计算  $k$  个聚类簇心。
- (4) 迭代 (2) ~ (3) 步骤，直到新的簇心与原来的簇心相等或者小于指定阈值，算法才结束。

**K-means** 聚类算法是以簇为中心的，能够很好代表聚类中心的对象，而且时间复杂度比较低，在处理数据量大时效果好。但是，**K-means** 聚类算法也有缺点，就是需要人为指定  $k$  值，不同的  $k$  值聚类出来结果容易出现很大的区别。并且 **K-means** 算法对噪点比较敏感，容易进入局部最优解，导致聚类效果不明显。

### 2.5.3 二分 K-means 聚类算法

传统的 **K-means** 聚类算法结果容易受到初始簇心的影响，簇心选取不好很容易得到的是局部最小值。介绍二分 **K-means** 算法前介绍一个误差平方和 (SSE) 的定义：是用来度量聚类效果的一个指标，SSE 计算的就是一个簇中每个点到簇心的、平方差。SSE 越小，聚类效果越好。

二分 **K-means** 聚类算法的主要思想是：首先把所有点当成一个簇，然后将该簇一分为二。之后选择能最大程度降低聚类代价函数（误差平方和）的簇划分为两个簇。以此下去，直到簇的数目等于用户指定的  $k$  为止。算法过程如下：

- (1) 将所有数据点当成一个簇。
- (2) 当簇数量小于  $k$  值时，对每一个簇计算误差平方和。
- (3) 在给定的簇上进行 **K-means** 聚类算法 ( $k=2$ )。
- (4) 选择使得误差最小的簇进行划分操作。



## 第3章 校园微博热点话题发现系统设计与实现

### 3.1 系统设计目标及要求

#### 3.1.1 系统设计目标

本文以高校学生微博用户为研究对象,实现对高校学生微博热点话题进行及时获取、分析和监控,并根据热点话题类别进行图表方式展示。系统可视化界面可以实现操作一套热点话题发现的流程,并且实时对校园微博热点话题进行监控,从而提供微博舆情的预警。

由于高校微博用户群体用户量大,产生的数据量也随之增大,因此需要一个实时并发的系统不断获取微博数据并分析数据。

##### 功能性目标:

##### (1) 微博数据获取模块

系统获取的文本来源于新浪微博,用户可以根据自己需求,通过用户可视化界面操作配置需要监控的微博号,并获取该校园微博号的文本数据,支持设置爬取数据页数,一页有十条微博内容,后续可实现支持配置不同微博号内容存入不同表设置。

##### (2) 微博文本预处理模块

获取数据是通过网页爬虫直接获取 HTML 文本,因此需要对文本进行提取,提取微博发布内容、发布时间、评论数、点赞数等。进一步对提取结果进行清洗,去除无效或无意义的微博内容及没有评论的微博,对文本进行分词处理及特征选择及提取等操作后构建向量空间模型。

##### (3) 微博热点话题发现

短文本聚类是微博热点话题发现系统最重要的模块,聚类结果对于热点话题发现的准确性有很大影响。该模块只需在可视化界面上操作获取热点话题步骤就可以获取热点话题。

##### (4) 热点话题可视化界面

在上面(3)步骤执行获取热点话题步骤后,可在界面上以图表形式展示热点话题信息及相关热点关键词,并实现敏感词展示。

### 3.1.2 系统设计要求

校园微博热点话题发现系统是集数据获取、文本处理、文本挖掘等模块，网页爬虫、中文分词等多种技术在一起的系统，所以对各功能模块的可用性要求比较高。为了实现对校园微博的热点话题发现，系统应该具备以下特性：

（1）实用性和稳定性。系统应该以实用为目的，选取合适的软硬件环境搭建系统，并保证系统长时间正常运行，而且需在发生故障发生后，能快速恢复系统。

（2）精准性。数据的来源和提取应该保证准确性，保证获取热点话题和舆情监控的准确性。

（3）可维护性和可扩展性。微博数据的不断增长，导致系统各模块不可避免出现问题。当出现问题时，就需要系统具备可维护性，快速定位问题并修复。考虑到系统用户的增加，系统流量大而导致系统无法可用，就需要系统具备可扩展性，可迁移到更高配置的软硬件环境下。

由于校园微博用户群体大，微博数据增长量也随之剧增，所以应该合理设计系统各个功能模块，保证每个功能模块的可用性，并能保证系统的抗压能力。

## 3.2 系统详细架构设计

根据设计目标与系统设计，本文研究的校园微博舆情监控系统有四大模块，分别是校园微博文本获取模块、微博文本预处理模块、校园微博热点话题发现模块、校园微博热点话题展示模块。系统整体功能架构图如图 3-1 所示：



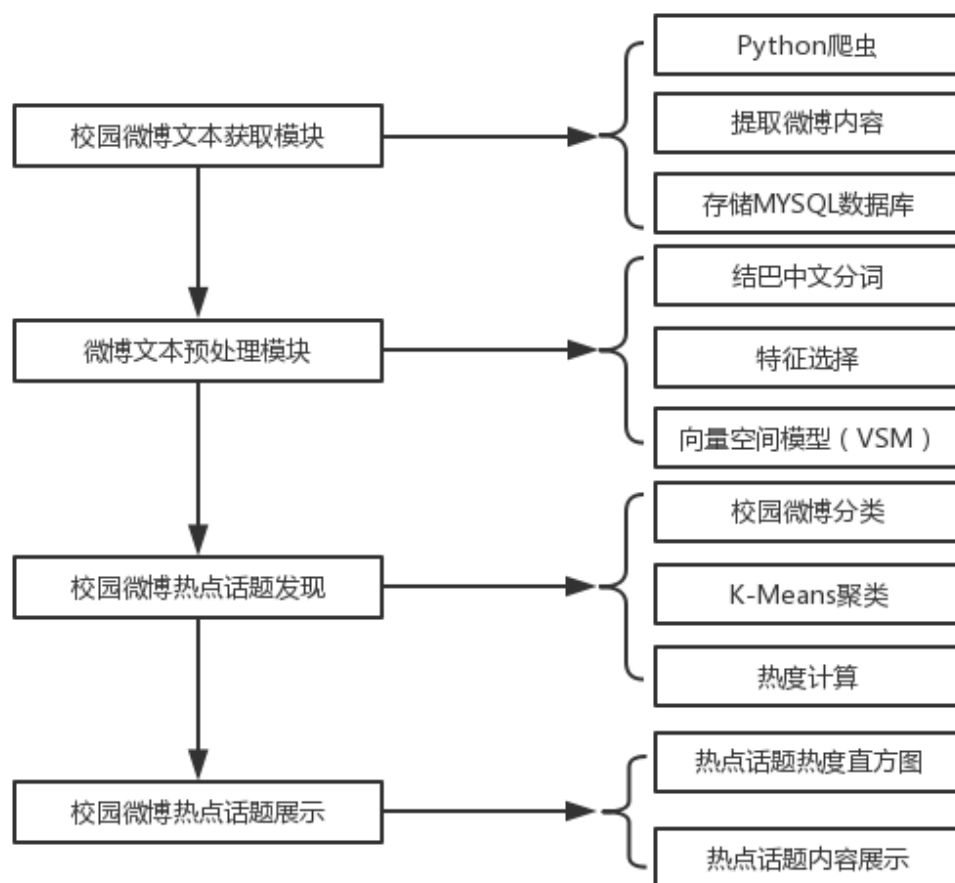


图 3-1 校园微博舆情监控系统架构图

校园微博热点话题发现系统主要分为四大模块。微博文本获取模块主要负责使用 Python 爬虫技术获取微博 HTML 文本并进行提取过滤后，存储文本到 MySQL 数据库中。微博文本预处理模块主要负责将文本分词并进行特征权重计算和特征选择后，进行构建向量空间模型等预处理操作。微博热点发现模块主要负责使用 K-means 聚类算法对文本进行聚类等操作，并计算相应的热度值。热度话题展示模块主要负责展示热点话题相关信息，并提供一个可视化界面供用户操作。

### 3.3 系统功能模块设计与实现

#### 3.3.1 微博数据获取模块

对校园微博进行舆情监控，首先应该设计一个微博数据获取模块，用于校园微博数据获取，并对获取数据进行提取微博内容。获取微博数据有两种方式，一种是使用网页爬虫，另外一种就是调用官方微博提供

的 API 接口。

使用网页爬虫的优势在于爬取数据不受限制，并且获取数据比较全面，缺点是获取到的都是 HTML 文本，需要进一步对文本进行提取内容并处理等操作，提取数据处理步骤繁琐且花费时间长。而调用官方 API 接口的优势在于，获取数据比较方便，并且不用进行提取或处理等操作，速度较快，而缺点是爬取数据频率有限制，无法短时间内获取大量微博内容。本文使用的是网页爬虫获取微博数据方式，所以需要以下几个步骤：

#### 模拟登录微博：

由于微博对爬虫比较敏感，所以需要模拟登录微博，并获取账号 Cookie，使用该 Cookie 对后续微博内容进行爬取。本文采用的是微博网页版进行数据爬取，因此模拟登录流程不是特别复杂，无需验证码之类。

模拟登录的流程如下所示：

（1）用户输入用户名、密码，构建请求 Headers。

（2）发起 POST 登录请求。

（3）若请求返回状态为 200，则将该次请求的 Cookie 存入 Redis 内存数据库，后续获取微博文本只需从 Redis 中取出 Cookie，在请求中带上 Cookie 即可。

（4）存入缓存的 Cookie 会在定时任务里判断 Cookie 是否失效，若失效则重新模拟登录微博，并再次刷新 Cache 中的 Cookie。

#### 获取微博文本内容：

网页爬虫获取下来的文本是 HTML 文本，存在着各种不相关的信息，本文使用 Python 的 BeautifulSoup 库从 HTML 文件中提取指定文本内容。可以使用 BeautifulSoup 直接定位到 DOM 树的节点，将每条微博文本的文本内容、发布时间、评论个数、点赞个数等信息提取出来，并将这些数据存入 MySQL 数据库。数据表字段如图 3-2 所示：


名	类型	长度	小数点	不是 nul	虚拟	键
id	int	11	0	<input checked="" type="checkbox"/>	<input type="checkbox"/>	 1
title	varchar	255	0	<input type="checkbox"/>	<input type="checkbox"/>	
pub_time	varchar	255	0	<input type="checkbox"/>	<input type="checkbox"/>	
comment_num	int	255	0	<input type="checkbox"/>	<input type="checkbox"/>	
like_num	int	11	0	<input type="checkbox"/>	<input type="checkbox"/>	
url	varchar	255	0	<input type="checkbox"/>	<input type="checkbox"/>	

图 3-2 数据表字段

微博文本获取模块主要流程如下：

（1）使用 Python 爬虫技术。

主要是使用 Python HTTP 库 Requests 带上 Cookie 获取 HTML 文本。为了避免频繁爬取微博，触发微博的反爬虫机制。这里简单实现一个避免触发反爬虫的策略：在获取微博内容时随机更换请求头的 User-Agent，爬取指定页数后随机睡眠 1~2 秒，使用多个账号 Cookie 随机选择使用，这几个步骤基本可以保证不会触发到微博的反爬虫机制。触发了微博的反爬虫机制后会被封 IP，为了避免该情况，后期需要加入代理 IP 模块，通过不断切换 IP 爬取，这个策略是十分有效的，缺点在于免费代理 IP 的可用性低。

（2）使用多进程+协程并发模式。

多进程加协程策略并发获取微博文本，缩短文本获取时间。使用 Python 内置库 multiprocessing 来实现多进程并发，协程使用 grequests 库，该库与 Requests 是同一种库，只是使用 Gevent 对 Requests 进行封装。Gevent 是基于 greenlet 的异步并发网络库，该库大大加快了微博数据获取的速度，并且比线程更轻量级、而且在一个协程挂掉之后对进程没有其它影响。爬虫模块相关代码如下：

表 3-1 爬虫并发模块代码

---

```
1. from multiprocessing import Pool
2. def run_crawl_multiprocess(start_page, end_page, pool=4):
3.     p = Pool(pool)
4.     page_num = end_page - start_page + 1
5.     interval = page_num // pool
6.     for page in range(1, page_num+1, interval):
7.         p.apply_async(run_async_crawl, args=(page, page+pool))
8.     p.close()
9.     p.join()
```

---

分析：

通过上面代码，引入了 Python 的 multiprocessing 进程池，通过设置进程池个数来实现控制并发个数，这部分代码实现了多进程并发获取微博文本。

---

1. 创建进程池，并分配爬取任务给 pool 数的进程。
2. apply\_async 函数作用就是分配任务给进程池中进程，每个进程处理 interval 页数爬取任务。
3. 第 8 行 p.close() 关闭进程池，使其不能接受新任务。
4. 最后 p.join() 作用是主进程阻塞等待子进程退出。

### （3）微博文本判重策略。

使用布隆过滤器实现过滤重复微博文本，防止重复存入数据库。提取每条微博文本的 URL 并使用正则表达式提取文本标识，布隆过滤器使用 Redis 缓存布隆信息，使用 Redis 的 bit 类型存储文本的布隆值。在存入数据库前进行布隆判断是否已经存在，若存在则跳过保存，否则存入数据库。

微博文本获取并存储数据库流程如图 3-3 所示：



图 3-2 校园微博文本获取模块流程图

### 3.3.2 微博文本预处理模块

在获取到微博文本后，需要对微博文本进行预处理。该模块主要负责微博文本的分词处理和过滤、特征项选择、向量空间模型（VSM）的构建这几步操作。该模块的流程图如图 3-3 所示：

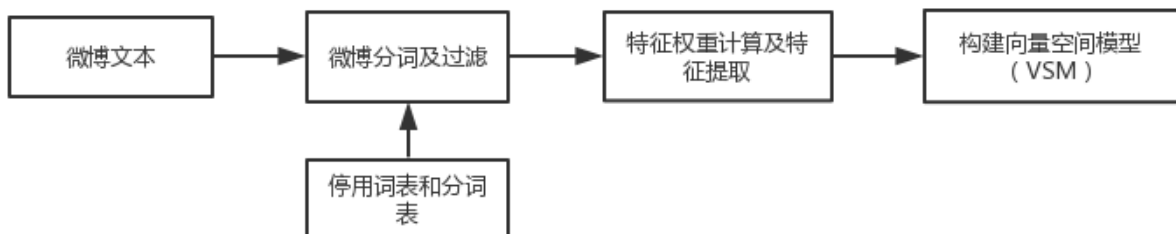


图 3-3 微博文本预处理模块流程图

（1）微博文本过滤。由于微博文本经常附带着很多表情符号、超链接之类，这些信息对于话题的发现和提取是没有意义的。其次，微博

文本具有随意性，经常出现微博文本字数少，经常出现只是发几个字，包含主题信息比较少的微博。本文使用的过滤策略：过滤掉微博长度为10的微博文本，该类文本长度过短，分词后包含文本信息过少；过滤掉微博评论数为零的微博文本，该类文本由于没有评论内容，因此文本信息相对过少。

(2) 分词处理与过滤。该模块使用 jieba 分词系统，该分词工具有分词精度好、分词速度快和支持处理字符等功能。jieba 分词的相关特点：

- ① 支持三种分词模式，分别是精确模式、全模式、搜索引擎模式。
- ② 支持自定义词典，自行添加新词。例如一些校园内特定的简称，可以加入自定义词典，保证更高的准确性。
- ③ 支持自定义停用词表，过滤掉一些对微博热点话题发现无意义的词。

在分词后进一步过滤掉文本分词后词语数量少于5的文本，该文本分词后信息过少。

表 3-2 加载 jieba 词典并过滤

---

```
1. import jieba
2. jieba.load_userdict(get_jieba_dict_path("user_dict.txt"))
3. def participle_text(text):
4.     seg_list = jieba.cut(text, cut_all=False)
5.     filter_content = set(seg_list) - stop_words
6.
7. def load_stop_words():
8.     """加载停用词表
9.     """
10.    stop_words = set()
11.    stop_words_path = get_jieba_dict_path("stop_dict.txt")
12.
13.    with open(stop_words_path, 'rb') as fp:
14.        for line in fp.readlines():
15.            stop_words.add(line.decode('utf-8').strip('\n'))
16.    return stop_words
```

---

分析：

通过上面代码，引入 jieba 分词库，并加载自定义词典和停用词表对微博文本进行分词过滤处理。

1. jieba.load\_userdict 作用是加载用户自定义词典，提高分词准确性。
2. 然后使用 jieba.cut 对文本进行分词，cut\_all=False 是指定以精确模式进行分词。
3. 接下来就是停用词过滤，先从 load\_stop\_words 函数获取自定义的停用词表，并把所有停用词放入 set 集合中。

过滤文本分词中的停用词，只需要让文本分词和停用词进行集合差相减，得到的结果就是过滤掉停用词后的文本分词。

### （3）特征提取及权重计算

特征项是微博文本分词后的某个词，是能够体现文本主题的词。对微博文本分词后进行提取，提取目的是为了降低数据的改维度并提取出能够反映微博文本主题的特征项。这里使用张静研究并进的特征提取方法，该方法是把所有微博文本分词后的关键词集合在一起，并把关键词出现的次数作为词频，根据改进的 TF-IDF 公式计算各特征权重，根据权重指导特征项的选择。特征提取及权重计算处理流程如下：

- ① 将语料库所有的分词后的词集合在一起。
- ② 迭代分词集合，取出一个词，进行统计词频。
- ③ 若该词没有出现，则将词频为 1，否则为原词频+1。
- ④ 全部迭代完成后，计算所有特征项的 TF-IDF 值。
- ⑤ 根据特征项的 TF-IDF 值指导特征项的选择。

表 3-4 TF-IDF 权重计算代码

```
1.def tf_idf(self):
2.    """ TF * IDF
3.    计算所有关键词的 tf-idf 权重值
4.    """
5.    self.tf()
6.    words_num = len(self.tf_dict)
7.    for word, value in self.tf_dict.items():
8.        self.tf_idf_dict[word]=float(value* float(math.log(words_num
```

---

```
/ value + 1)))
```

```
9. return self.tf_idf_dict
```

---

分析：

通过上面代码，计算微博文本每个特征项 TF-IDF 的权重。

1. self.tf() 函数获取每个特征项的 TF 值。

2. 迭代 tf\_dict 集合，计算每个特征的 TF-IDF 权重值。

---

#### （4）向量空间模型表示

向量空间模型（VSM）的表示，是基于 TF-IDF 计算得到的特征项权重来实现的。本模块使用特征权重排序后的结果，根据情况选取前几十特征项作为向量空间模型的基础。向量空间模型表示的就是每条微博文本的信息，基于该模型可以通过欧式距离算法计算文本之间的相似度。向量空间模型结构如下：

$$s = (w_1, w_2, w_3 \dots w_{n-1}, w_n)$$

$$w_1 = (0, 0.1, 0.12 \dots 0.34, 0.84)$$

其中  $s$  为整个语料库， $w_1$ 、 $w_2$ 、 $w_n$  等都是每条微博文本分词后对应的特征权重向量。

### 3.3.2 校园微博热点话题发现模块

校园微博热点话题发现模块是该系统的核心模块，该模块发现热点话题的准确性直接影响到舆情监控的效果，是热点话题计算热度的前提条件。该模块主要包括：校园微博分类模块、K-means 聚类算法模块、热度计算模块。

#### （1）校园微博分类模块

由于校园微博主要用户群体为大学生，通常都是发布关于学校、校园生活、买卖交易、情感之类的话题。本文通过对校园微博的研究，决定在使用聚类算法之前先对整个微博文本语料库进行分类处理，然后再对每个分类分别进行聚类算法，这样可以大大增加热点话题的准确性。分类流程如下：

① 维护一份分类训练集，用于训练分类。

② 使用 Scikit-learn 机器学习库中的 TF-IDF 算法构建向量空间模型，并将文本向量模型对象化，方便后续生成向量空间模型。

③ 构建测试集的向量空间模型，执行多项式贝叶斯算法进行文本

分类，并把不属于该类的文本正确分类。

## （2）微博短文本聚类

本文使用的是 K-means 聚类算法对前面构建向量空间模型进行聚类分析。第 2 章已经介绍过 K-means 聚类算法的原理和优缺点。由于 K-means 聚类算法关键在于初始簇心的选择和 k 值的选择，这些都会影响 K-means 聚类的效果。下面是微博文本中 K-means 聚类算法的流程：

- ① 加载之前步骤处理得到向量空间模型。
- ② 随机或者人工设置初始簇心。
- ③ 调用 K-means 算法并指定距离算法，一般使用欧式距离算法。
- ④ 遍历向量空间模型，使用距离算法计算文本间的相似度。
- ⑤ 把文本归类到距离最近的簇心类别中，并更新 k 个簇心。
- ⑥ 迭代直到簇心没有变化，退出算法。

通过 K-means 聚类后，获取文本的标签后把文本归于对应的话题中。

## （3）微博热点话题热度计算

经过上面步骤聚类算法，初步得到了话题类，但这些类中话题还不能直接代表热度话题，还需要对每一个类进行热度计算，得出最大热度的话题类，该类才能作为校园微博热点话题。

判断一个话题类是否是热点话题，应该根据特定的热度公式计算该类别的热度。热度计算不能单独以该话题类中微博文本数量决定，因为微博还有评论功能。通常情况下越受关注的话题评论越多，点赞数越多，由于校园微博的特点，一般情况下微博转发情况比较少。所以应该根据每条微博的评论数和点赞数等数据根据公式计算出热度计算。热度计算的公式如下：

$$hot_i = \log(like + 1) + comment \quad (3-1)$$

经过热度计算后选取最大热度的话题类即是该段时间内的热度话题。

## （4）微博热点话题展示。

得到微博热点话题后，系统提供一个展示热点话题的界面。校园舆情监控人员可以通过可视化界面快速获取舆情热点话题信息，还可以通过图表形式获取当段时间的热点话题的主题关键词等，同时还提供一个敏感词界面，让舆情人员快速获取一些紧急的事件，例如抑郁、自杀之类的敏感词，方便舆情人员快速发现校园热点事件和掌握舆情动态。可



视化界面使用的技术及功能如下：

① 爬取微博文本界面。可由舆情人员自动配置爬取微博号，配置爬取页面等操作。

② 微博热点话题发现界面。系统的聚类分析、热度发现模块在界面上不可见，直接通过界面提示配置即可生产热点话题相关信息。

③ 敏感词展示界面。提供敏感词识别并以图表形式展示敏感词相关信息。



## 第 4 章 系统功能测试

### 4.1 系统运行环境和参数

本系统在开发过程需要相应的软硬件环境，开发测试使用的是本机环境，不同的软硬件环境对系统的稳定性以及流畅性都有很大影响。开发环境配置如下：

硬件环境：内存 12G、硬盘容量：100G、CPU: 4 核

系统环境：64 位 Ubuntu16.04 系统

数据库：MySQL、Redis

开发语言与工具：Python、Vim、Pycharm

框架与相关库：Scikit-Learn、Flask、Numpy、Requets、BeautifulSoup、Celery、Gunicorn

生产环境与开发环境配置只是在硬件和系统上有区别，其它区别不大。生产环境使用的是阿里云服务器，系统及相关模块全部部署在阿里云服务器，生产环境配置如下：

硬件环境：内存 1G、硬盘容量：100G、CPU: 2 核

系统环境：64 位 Ubuntu14.04 系统

数据库：MySQL、Redis

其它数据：同生产环境

### 4.2 实验数据及处理

本系统以广州中医药大学的一个生活类微博号——广中医 I 栋作为数据来源，获取了一个月内的微博内容作为实验数据。实验过程中，过滤掉一部分不符合要求的微博文本，例如将微博文本长度小于 10，评论个数小于 2 全部过滤掉，因为文本太短或者评论太少本身就对热点话题的发现模型产生噪点影响。

根据第三章的微博文本分类模块，先使用维护的分类词表对微博文本进行分类，然后维护一个已经训练好的训练集并对归类错误的文本正

确归类，归类成功后的文本可用于后面步骤的文本预处理。

根据第三章的微博文本预处理步骤对文本进行分词及停用词过滤等操作，并维护一份分词表和停用词表，用来提高分词的准确率。图 4-1 是分词后的微博文本：

中医 感激 栋 蓝色 广 一卡通 发 中药 寻 照片 模糊 药剂 联系方式 学院 卡套 学号 作揖 I 失主 甚  
中医 拿走 越来越 栋 找出 广 作案 频率 高 力量 生气 钱 吃白食 楼主 此人 平台 希望 室友 付 同学 外卖  
想办 保卫处 通行证 办理 港澳 学校 流程 出入境 户口 登记 迁 证件 请问  
问 广州 医药 中药 联系 学校 单位 实习 学院 帮忙 转行 相关  
挑战杯 老师 项目 想到 课题 先 参加  
饭堂 乱打 饭菜 价钱 学校 感觉 难吃 真心 贵  
现行 意愿 小组 栋 人有 外卖 抓 跑下去 电话 痛恶 偷 嘴脸 好想 抓个 接到 点个 飞毛腿 成立 至极

图 4-1 分词后文本

然后对分词后的文本进行 TF-IDF 权重计算。本文使用的是改进的 TF-IDF 算法，把语料库所有关键词集合在一起，然后计算每个关键词的权重值，并根据 TF-IDF 权重值指导特征的选择，特征选择后构建向量空间模型，对于微博短文本而言，一般选择 TF-IDF 权重值前五十的特征作为向量空间模型的维度。如下图 4-2 所示，由于维度过大，图中向量空间模型的列有所减少。矩阵中每一行代表着一条微博文本的对应的特征向量值：

```
0.917 0.0 0.313 0.0 0.23 0.225 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.126 0.0 0.1
0.917 0.323 0.0 0.0 0.23 0.0 0.0 0.0 0.168 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.917 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.917 0.0 0.0 0.253 0.0 0.0 0.208 0.0 0.0 0.0 0.0 0.0 0.138 0.0 0.0 0.0
0.0 0.323 0.0 0.253 0.0 0.0 0.208 0.208 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.917 0.0 0.313 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.313 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.126 0.12 0.0
0.917 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.168 0.0 0.0 0.0 0.0 0.0 0.12 0.0
0.917 0.323 0.0 0.0 0.23 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.138 0.0 0.0 0.0
0.0 0.323 0.0 0.0 0.23 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.917 0.323 0.0 0.0 0.23 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.23 0.0 0.0 0.0 0.156 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.917 0.0 0.313 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.208 0.0 0.0 0.0 0.0 0.138 0.0 0.0 0.0
0.917 0.0 0.0 0.253 0.0 0.0 0.0 0.0 0.168 0.0 0.0 0.0 0.0 0.0 0.0 0.0
```

图 4-2 向量空间模型 (VSM)

接下来使用聚类算法对文本进行聚类分析，使用 K-means 算法对文本空间向量进行相似度计算并归于对应的类别。

## 4.3 系统可视化界面

系统可视化界面主要分为几大功能展示，用户登录界面，爬取微博数据界面、热点话题分析发现界面、各分类热度展示界面、敏感词展示模块等。由于该系统主要功能在于热点话题发现方面，对相关的登录等界面其它功能不做解释。

### 4.3.1 数据获取界面

通过输入要获取数据的微博号的微博链接，并设置爬取相关页数，系统在后台自动执行爬取微博的操作，期间不需要任何的配置。如下图4-1所示：



爬取链接

起始页数

终止页数

点击爬取

图 4-1 微博数据获取界面

### 4.3.2 热点话题排行榜

系统主页面包括当前时间段的热点话题排行榜、热点话题内容等模块。如下图4-2所示，展示的是热点话题排行榜的前八位热点话题事件，并且展示热点话题的前七位关键词与热度值：

当前时间段的热点话题如下：

排名	热点事件类别	热点事件关键词	热度值
1	情感第1类	喜欢 女生 男生 男朋友 分手 朋友 问	1068.91
2	学校新闻第3类	学校 请问 三元里 学生 封路 有人	760.18
3	买卖交易第3类	评论 私聊 出 有意 价格 买 有意者	705.77
4	求助第1类	有人 请问 出 买 香港 大学城 求	574.56
5	校园生活第2类	同学 宿舍 外卖 偷 栋 请问 师兄	429.56
6	情感第3类	喜欢 男生 女生 男朋友 推荐 女朋友 感觉	366.96
7	情感第2类	喜欢 女生 男生 感觉 男朋友 女朋友 朋友	348.7
8	校园生活第3类	同学 宿舍 师兄 外卖 养 大学城 喜欢	334.82

图 4-2 热点话题排行榜

### 4.3.3 热点话题热度直方图

其中图表展示模块展示了各类别中聚类后的最大热度直方图，并展示最大热度类别的微博内容，如下图 4-3 所示：

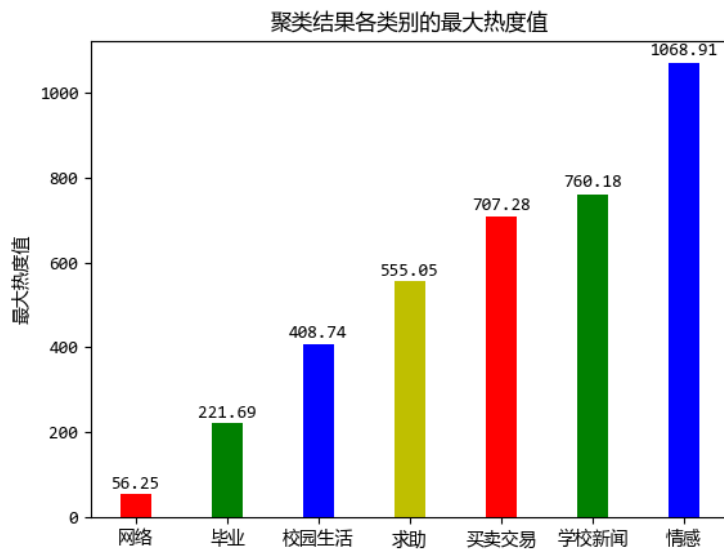


图 4-3 热度直方图

### 4.3.4 敏感词展示

后台维护一份敏感词表，用来匹配微博文本中存在的敏感词，一般校园微博存在心理健康、校园突发事件、校园安全等需要舆情人员监控的敏感类别。本系统实现后台定时任务来监控相关敏感词并实时更新在

系统敏感词展示页面， 如下图 4-4 所示：

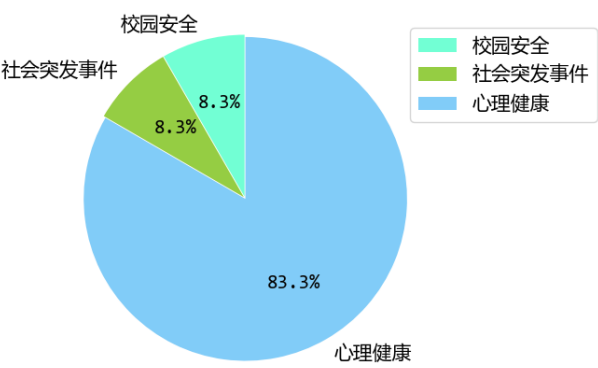


图 4-4 敏感词饼形图

敏感词类型下的微博文本如图 4-5 所示：

舆情话题类型		心理健康		寻找			
舆情类型	敏感词	微博内容	评论个数	点赞数	发布时间		
1	心理健康	抑郁	最近心情特别抑郁，生活很累，虽然也知道还有更辛苦更累的人，但自己总是克服不了自己内心最黑暗的想法	1	3	2018-04-20 13:46:27	
2	心理健康	抑郁	曾经因为舍友对我有恩我容忍她一切脾气任性懒惰自私，后来发现其实都不值得，因为她在拥抱我的同时在背后抽出一刀，曾对这友情的破裂抑郁，好在身边有好友相伴，希望善待他人，不要有伤害他人的想法	1	4	2018-04-25 10:21:29	
3	心理健康	抑郁	怀疑自己得了抑郁，我会去医院的，每一天都在忍受，忍受	10	0	2018-03-30 12:11:03	
4	心理健康	抑郁	心情抑郁你们会干嘛	9	0	2018-03-26 08:29:10	

图 4-5 敏感词类型下文本





## 总 结 与 展 望

本文以校园微博作为研究主体,结合微博的特点,通过研究热点话题相关技术,实现了一套校园微博热点话题发现系统。该系统主要由数据获取、微博数据预处理、K-means 聚类分析等模块组成。该系统为校园舆情人员及时发现热点话题带来了很大的帮助。本文主要内容如下:

(1) 使用 Python 爬虫技术并发获取微博页面,并使用 Python 相关库快速提取微博文本内容。

(2) 观察校园微博的特点和传统文本的区别,根据校园微博短文本的特点进行数据预处理操作。

(3) 针对于向量空间模型的高维度以及微博文本表示的稀疏性,通过改进的 TF-IDF 算法实现有效的降维和特征选择,解决了 VSM 特征向量的稀疏性问题。

(4) 针对于传统 K-means 算法存在的局部最优解问题,改进了 K-means 算法初始簇心选择,提高了 K-means 聚类的准确性。

(5) 根据校园微博的特点,改进了热度计算的算法,提高了获取热点话题的准确性。

(6) 实现热点话题发现系统可通过可视化界面进行操作,方便舆情管理人员从界面上获取热点话题信息。

本文基于聚类算法实现校园微博热点话题发现系统还是有一些不足之处,以下几项需要完善:

(1) 数据预处理模块的特征选择后文本表示稀疏性还是有点高,需要改进特征提取相关方法。

(2) K-means 聚类算法对微博短文本的聚类效果不是很理想,需要换用其它聚类算法,例如 Single-pass 等算法。或者通过两种聚类算法结合方法对文本进行聚类,这是本系统下一步需要研究的方向。

(3) 可视化界面还需要不断改进,目前只是提供一个界面供舆情人员操作并查看相关热点话题信息,但不提供修改相关参数的界面。下一步应该完善相关参数配置界面,使得舆情人员可根据需要配置相关参数达到对某些话题的舆情监控目的。



## 参考文献

- 1 CNNIC. 第 41 次中国互联网络、发展状况统计报告. 中国互联网络信息中心, 2018, 1-2
- 2 Jing Guo, Peng Zhang, Jianlong Tan, Li Guo. Mining Hot Topics from Twitter Streams. *Proedia Computer Science*, 2012, 9(2012), 2008-2011
- 3 Allan J, Carbonell J, et al. Topic Detection and Tracking Pilot Study Final Report. *Darpa Broadcast News Transcription & Understanding Workshop*. San Francisco, USA, 2000:194-218
- 4 Salton G, Wong A, Yang C S. A vector space model for automatic indexing. *Communications of the Acm*, 1975, 18(11):273-280
- 5 Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on Twitter based on temporal and social terms evaluation. *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. ACM, Chicago, USA, 2010:1-10
- 6 De Villiers F, Hoffmann M, Kroon S. Unsupervised Construction of Topic-Based Twitter Lists. *South Africa: Stellenbosch University*, 2012:12-34
- 7 张东霞. 基于高校学生微博的网络热点发现及舆情分析研究: [华南理工大学硕士学位论文]. 广州: 华南理工大学, 2013, 25-30
- 8 陈彦舟, 曹金璇. 基于 Hadoop 的微博舆情监控系统. *计算机系统应用*, 2013, (4): 17-20
- 9 张亚男. 基于混合聚类算法的微博热点话题发现的研究: [杭州电子科技大学硕士学位论文]. 杭州: 杭州电子科技大学, 2017, 23-39
- 10 李磊. 基于新浪微博的热点话题发现系统研究与实现: [复旦大学硕士学位论文]. 上海: 复旦大学, 2012, 30-39
- 11 孙胜平. 中文微博客热点话题检测与跟踪技术研究: [北京交通大学硕士学位论文]. 北京: 北京交通大学, 2011, 29-37
- 12 张静. 基于微博的网络热点话题发现模型及平台研究: [华中科技大学]. 武汉: 华中科技大学, 2010, 16-31



## 致 谢

本文的大部分工作都是在蔡洪民老师的指导下完成。蔡老师严谨的治学态度、开阔的科研思路、丰富的科研经验使我受益匪浅。课题期间，蔡老师帮助我解决了一些选题与技术上的一些难题，并对相关技术提出了一些宝贵的建议，这些建议和指导对我完成这个课题有着很大的帮助。在这里，我衷心感谢蔡洪民老师在我本科论文期间对我论文的相关指导和帮助。此外，感谢医学信息工程学院的全体老师、感谢您们四年来对我的教导和鼓励。

另外，感谢这篇论文所引用过技术知识的各位学者，本文引用了一些研究文献，这些文献对于我论文的思路有着很大的影响与帮助。

最后，我要感谢同学和家人们，是他们对我的鼓励和支持帮助我顺利完成这篇论文。



## 附录

K-means 聚类算法如下：

```
def k_means(self, distance=euclidean_distance):
    """ kmeans algorithim

    :param data_set: 数据集
    :param k: k 个簇心
    :param distance: 距离算法
    :return:
    """
    row = numpy.shape(self.data_set)[0] # 获取行数
    # 初始化一个矩阵， 用来记录簇索引和存储误差平方和(指当前点到簇质
    点的距离)
    cluster_assment = numpy.mat(numpy.zeros((row, 2)))
    # 随机生成一个质心矩阵簇
    centroids = self.rand_cent()
    centroids = self.set_rand_cent()
    cluster_change = True
    while cluster_change:
        cluster_change = False
        for i in range(row): # 对每个数据点寻找最近的质心
            min_dist = numpy.inf # 设置最小距离为正无穷大
            min_index = -1
            for j in range(self.k): # 遍历质心簇，寻找最近质心
                dist_j = self.euclidean_distance(centroids[j, :],
self.data_set[i, :])
                if dist_j < min_dist:
                    min_dist = dist_j
                    min_index = j
            if cluster_assment[i, 0] != min_index:
```

```
cluster_change = True
cluster_assment[i, :] = min_index, min_dist ** 2 # 平方的意义
在于判断聚类结果的好坏

for cent in range(self.k): # 更新质心，将每个簇中的点的均值作为质
心
    index_all = cluster_assment[:, 0].A # 取出样本所属簇的索引值
    value = numpy.nonzero(index_all == cent) # 取出所有属于第
cent 个簇的索引值
    sample_in_clust = self.data_set[value[0]] # 取出属于第 I 个簇的
所有样本点
    centroids[cent, :] = numpy.mean(sample_in_clust, axis=0)
return centroids, cluster_assment
```