

基于数据挖掘的微博情感分析研究报告

华东师范大学 计算机科学与软件工程学院

小组成员： 谢 斌 10142510237

向 茜 10142510236

王 鑫 10142510229

陈伟文 10152510217

目录

一、研究背景 3

二、主要研究内容 4

三、数据采集 5

3.1微博开放平台API 5

3.2 关键词搜索爬取微博数据 8

四、数据预处理 13

4.1中文分词 13

4.2停用词去除 14

4.3词性选择 15

4.3.1特征选择： 15

五、基于频繁项集和语义的微博情感聚类算法实现 19

六、结果分析和性能评估 20

6.1获取频繁项集： 20

6.2基于语义隶属原则的初始簇重叠消除： 20

6.3基于语义相似度的凝聚层次聚类： 21

七、总结 22

# 一、研究背景

根据中国互联网信息中心进行整理统计的数据，到2015年之前，国内网民有接近7亿人，互联网普及率高达47.9%，这意味着互联网用户越来越庞大。

社交网络是指人们相互之前通过各类关系建立的社会网络结构，指的是社会行为者间关系的集合，即社交与网络的融合，人们通过网络这个载体进行相互连接，进而形成具有一定特色的群体，如微博、微信等。

如今，随着互联网的深入发展，移动端技术逐渐成熟，社交网络已经融入了人们的日常生活，越来越多的网民选择社交网络作为获取信息的一种方式，并通过一些网络交流平台抒发个人情感、点评社会热点等等。互联网能够在极短的时间内凝聚网民意见，并在一定程度上引导舆论走向，诱发民众的一系列行为并对社会造成一定的影响。微博近年来成为大众使用最频繁的社交网络平台，个人、团体、企业等各种形式的账户扩展了微博用户的多样性，也增加了微博上信息的数量以及种类，在繁杂大量的信息中有很多值得挖掘、分析的数据，情感便是其中之一。

舆情是“舆论情况”，代表在某特定社会空间内，随着社会事件的发展、推进与深入，广大群众对事件的社会态度，社会看法等。舆情汇总了绝大多数群众对当今社会出现事件所特有的态度、立场、见解和情感等。

由于社交网络的匿名性，大多数信息都可以真实反映用户当前真实情感，也因为社交平台使用方法简单，信息量充足多样，能够吸引大量用户参与热点事件的讨论，这样的实时互动更能够体现用户的情感变化以及走向，也正式是因为社交网络舆情涵盖的内容很广泛，涉及到政治、经济、娱乐、生活等各个方面，及时地监测用户情感、分析并加以控制才显得如此重要。

微博，作为新媒体呈现形式，微博含有的大量的短文本，能够反映这些用户对某特定产品、品牌的喜好程度，也能够反映对社会热门事件的个人看法，能够企业进一步提高自己的产品，帮助政府正确引导舆论。微博具有较为活跃的用户量，庞大用户带来海量的数据信息，若仅仅依靠传统的人孔方法进行整理统计，无法高效地提取有价值的信息，迫切的需要利用数据挖掘技术对海量短文本数据进行分析与处理。

# 二、主要研究内容

微博情感分析的主要任务是分析微博信息流中的用户情感。本项目是针对微博文本集进行研究，故需要将文本集进行一系列划分，其中各个微博短文本都从属于同一情感，基于同类情感的微博短文本间具有较大的相似度，不同类情感的微博短文本间相似度较小这一原则，所以本项目采用短文本聚类的方法进行研究。因此，本文主要目标是结合微博短文本的显隐性情感特征，设计一个有效的短文本聚类算法，可以进行有效的情感分析。本项目在实施过程中主要进行了以下工作。

1. 通过分析微博情感分析研究现状，结合微博情感分析的目标，将情感进行细粒度分析，分为高兴、喜好、愤怒、悲伤、厌恶、惊讶、恐惧七种情感，将微博短文本进行细粒度情感分析，更有利于社交网络舆情分析。
2. 通过分析微博短文本的特点，以及微博的显隐性情感特征，设计并构建了表情符号词典，情感词词典，提出了基于微博显性情感特征的微博特征选择方法，充分考虑微博显性情感特征，将这些显性情感特征进行频繁挖掘，得到的特征项更好地保留原始微博的情感特征。
3. 分析了传统基于频繁项集聚类算法，在微博短文本聚类问题上的缺陷，提出基于频繁项集语义隐性情感特征聚类方法，首先，采用包含显性情感特征的频繁项集构建初始簇；然后引入《知网》语义库实现基于语义层次的簇间重叠削减；最后，通过凝聚式层次聚类合并微博情感簇，获取最终的情感簇。同时，还通过实验对比来验证算法的有效性。
4. 搭建了微博情感分析与监控系统，通过对雄安新区成立事件的微博数据集进行情感分析，验证了文本方法对微博情感的分析与监控能力。

# 三、数据采集

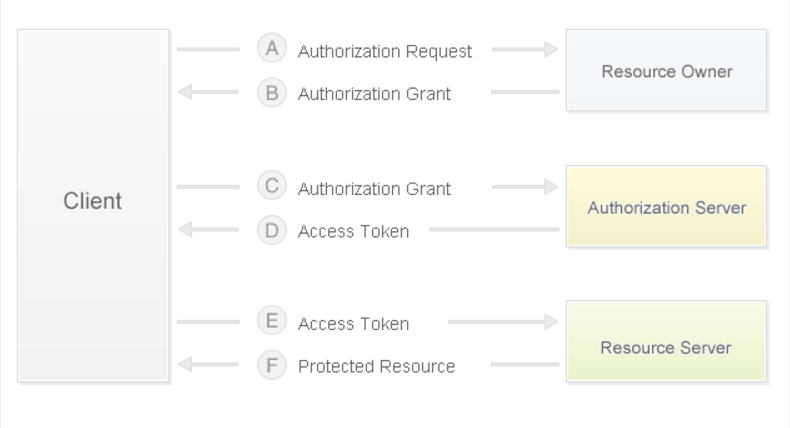
## 3.1微博开放平台API

新浪微博提供了一系列可供外部程序使用的接口（API），其中包括内容型接口、关系型接口、工具型接口、数据标签以及微博内的框架和组件。通过这些接口与服务，开发者易获得微博的相关用户资源和内容资源，也便于获得所需的微博的结构化信息。

1)OAuth2.0

微博开放接口的调用，如发微博、关注等，都是需要获取用户身份认证的。目前微博开放平台用户身份鉴权主要采用的是OAuth2.0。较1.0相比，整个授权验证流程更简单更安全，也是最主要的用户身份验证和授权方式。

关于OAuth2.0协议的授权流程可以参考下面的流程图，其中Client指第三方应用，Resource Owner指用户，Authorization Server是授权服务器，Resource Server是API服务器。



用户先访问客户端，向资源拥有者发送一个授权临时令牌请求；当资源所有者收到此授权请求，向客户端发送一个访问请求；客户端会将访问许可及私有证书，发送给授权服务器，并向授权服务器发送临时令牌请求；授权服务器对客户端提交信息进行验证，授予一个临时令牌；客户端获得临时令牌后，将用户引导至服务提供方的授权页面请求用户授权。在这个过程中将临时令牌和客户端的回调连接发送给服务提供方，用户在服务提供方的网页上输入用户名和密码，然后授权该客户端访问所请求的资源。授权成功后，服务提供方引导用户返回客户端的网页。客户端根据临时令牌从服务提供方那里获取访问令牌。服务提供方根据临时令牌和用户的授权情况授予客户端访问令牌。客户端使用获取的访问令牌访问存放在服务提供方上的受保护的资源。

2)返回数据格式

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式。它基于ECMAScript (w3c制定的js规范)的一个子集，采用完全独立于编程语言的文本格式来存储和表示数据。简洁和清晰的层次结构使得 JSON 成为理想的数据交换语言，易于阅读和编写，同时也易于机器解析和生成，并有效地提升网络传输效率。

使用微博开放平台提供的API访问资源，服务端将以JSON格式返回请求的数据信息，直观且易于解析。

数据示例

{

"statuses": [

{

"created\_at": "Fri Feb 24 15:18:31 +0800 2012",

"id": 3416614810943471,

"mid": "3416614810943471",

"idstr": "3416614810943471",

"text": "与大家分享我所喜爱的照片！#ABC晒新年# 。",

"source": "微活动</a>",

"favorited": false,

"truncated": false,

"in\_reply\_to\_status\_id": "",

"in\_reply\_to\_user\_id": "",

"in\_reply\_to\_screen\_name": "",

"thumbnail\_pic": "http://ww3.sinaimg.cn/thumbnail/5f0eb04atw1dq4ir5bztkj.jpg",

"bmiddle\_pic": "http://ww3.sinaimg.cn/bmiddle/5f0eb04atw1dq4ir5bztkj.jpg",

"original\_pic": "http://ww3.sinaimg.cn/large/5f0eb04atw1dq4ir5bztkj.jpg",

"geo": null,

"user": {

"id": 1594798154,

"idstr": "1594798154",

"screen\_name": "刘麦",

"name": "刘麦",

"province": "34",

"city": "16",

"location": "安徽 亳州",

"description": "新一代世界小童星。",

"url": "http://blog.sina.com.cn/liumaiduo",

"profile\_image\_url": "http://tp3.sinaimg.cn/1594798154/50/5614782838/1",

"profile\_url": "liumaiduo",

"domain": "liumaiduo",

"weihao": "",

"gender": "m",

"followers\_count": 314,

"friends\_count": 555,

"statuses\_count": 1913,

"favourites\_count": 1,

"created\_at": "Sat Jun 11 00:00:00 +0800 2011",

"following": false,

"allow\_all\_act\_msg": false,

"geo\_enabled": true,

"verified": false,

"verified\_type": -1,

"allow\_all\_comment": false,

"avatar\_large": "http://tp3.sinaimg.cn/1594798154/180/5614782838/1",

"verified\_reason": "",

"follow\_me": false,

"online\_status": 1,

"bi\_followers\_count": 290,

"lang": "zh-cn"

},

"annotations": [...],

"reposts\_count": 0,

"comments\_count": 0,

"mlevel": 0,

"visible": {

"type": 0,

"list\_id": 0

}

},

...

],

"total\_number": 2543821

}

3)局限性

微博开放平台提供了集成的功能接口，易于开发者获取所需数据，虽然高效但是有着局限性。首先，获取新浪微博API分为普通接口和高级接口，而我们所需的高级接口需要申请才可以调用，但申请流程繁琐，且审核时间过长；其次，微博API返回的记录有数量限制，所获取的信息有限，因此，决定放弃新浪微博API，转而使用抓取网页后解析的手段获取数据。

爬取的数据将存放在MongoDB中，进行进一步的处理。

## 3.2 关键词搜索爬取微博数据

　　网络爬虫是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL

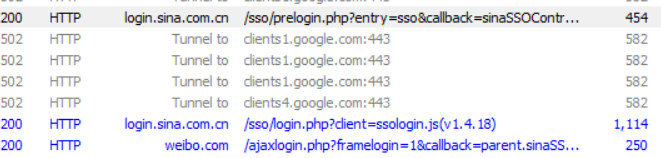
新浪提供了一个高级搜索功能，为我们爬取数据提供了一个很好的切入点。



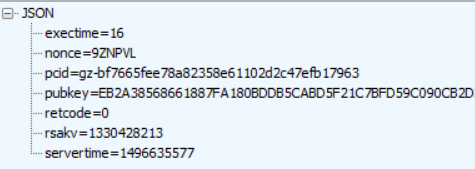
1)模拟登陆

　　高级搜索功能，需要用户登陆后才可以使用，所以首先需模拟登陆新浪微博。

PC 登录新浪微博时，在客户端用js预先对用户名、密码都进行了加密，而且在POST之前会GET一组参数，这也将作为POST\_DATA的一部分。在模拟登录之前，首先观察浏览器在登录过程中Fiddler抓到的包，在/sso/login.php打开之前会先使用GET方式请求/sso/prelogin.php。



在Fiddler中，打开Responese body查看返回数据详情



可以看出，这是一个json数据，并且携带了几个参数，我们关心的有以下四个：

·servertime

·nonce

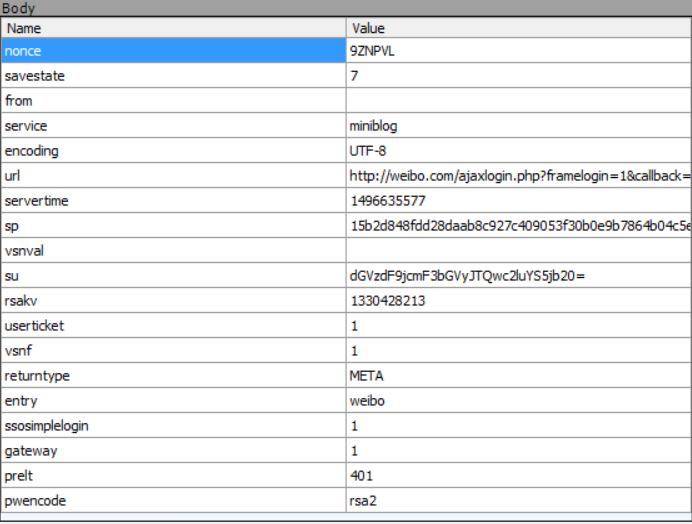
·pubkey

·rsakv

之所以认为这几个参数比较重要，那是因为后面对 password 的加密需要用到，得到这几个参数之后，POST方式请求login.php页面

### 

可以发现页面有如下参数



在参数列表中，需要我们指定的参数有下面几个：

·servertime

·nonce

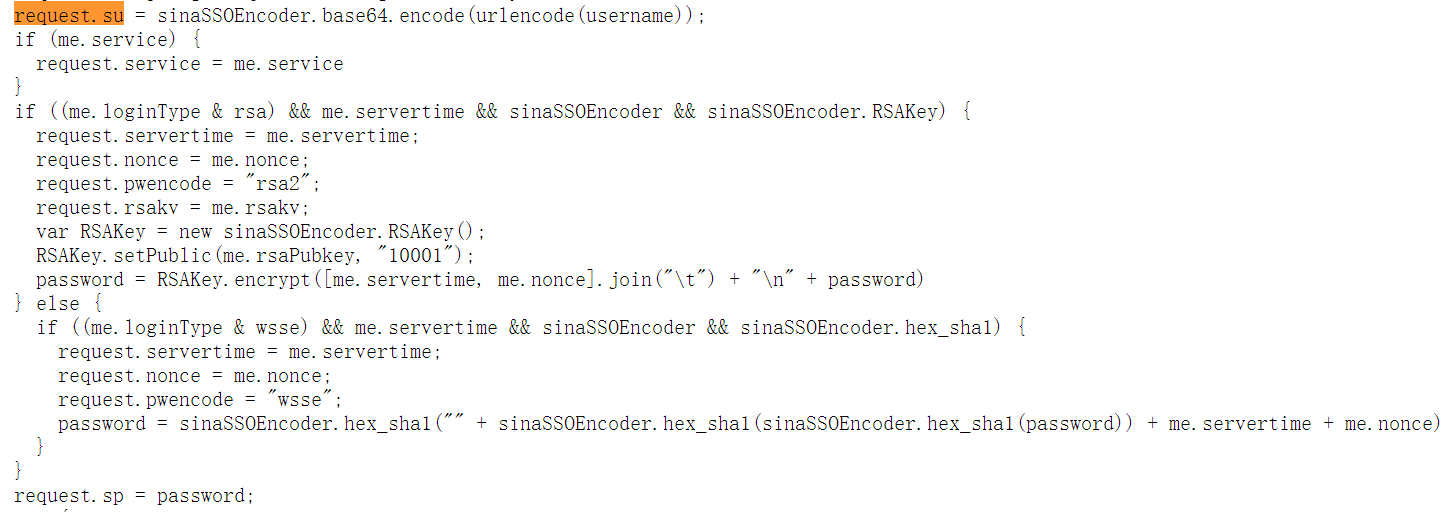
·rsakv

·sp：加密后的密码

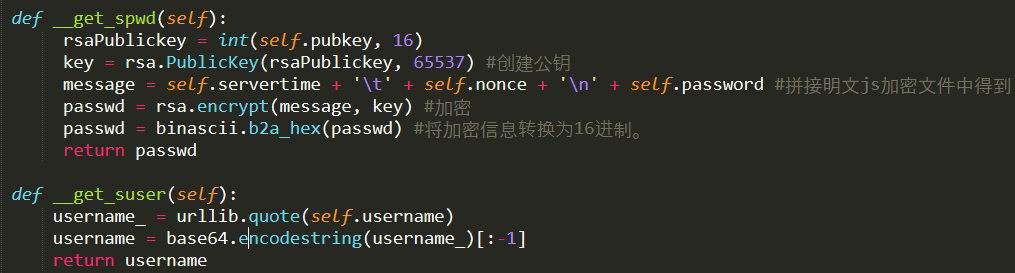
·su：加密后的用户名

对于参数nonce、servertime、rsakv，都可以从第一步中的prelogin.php中直接获取，而sp和su则是经过加密后的字符串值。因此，下一步就是要找出su和sp所代表的含义以及它们的加密方式。

打开微博登陆页面，查看源码，可以找到ssologin.js文件，搜索username和password，经分析后得知su为加密后的用户名，sp为加密后的密码。加密方式如下：



微博对于username的加密规则比较单一，使用的是Base64加密算法，而对password的加密规则比较复杂，使用的是RSA模块，但加密的逻辑比较多。根据上面的代码，可以看出password加密是这样的一个过程：首先创建一个rsa公钥，公钥的两个参数都是固定值，第一个参数是登录过程中prelogin.php中的pubkey，第二个参数是加密的js文件中指定的”10001”（这两个值需要先从16进制转换成10进制，把“10001”转成十进制为“65537”）。最后再加入servertime和nonce进行rsa加密，最后将加密信息转化为16进制。



2)构建URL

登陆新浪微博，进入高级搜索，如图输入，之后发送请求会发现地址栏变为如下：http://s.weibo.com/weibo/%25E6%25AF%2595%25E4%25B8%259A&region=custom:31:1000&typeall=1&suball=1&timescope=custom:2017-06-01:2017-06-05&Refer=g

解析如下：

固定地址部分：http://s.weibo.com/weibo/

关键字二次UTF-8编码：%25E6%25AF%2595%25E4%25B8%259A

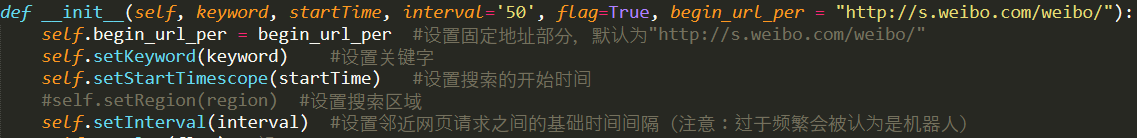
微博类型: typeall=1

微博内容包含: suball=1

搜索地区：region=custom:31:1000  
搜索时间范围：timescope=custom:2017-06-01:2017-06-05

可忽略项：Refer=g

某次请求的页数：page=1（第一页可不加）



将上述URL中的每一部分进行相应的拼接，即可构造访问的目标URL。

3)解析数据

打开高级搜索网页的源代码，分析页面元素，定位至字符串<script>STK && STK.pageletM && STK.pageletM.view({"pid":"pl\_weibo\_direct"处即为搜索到微博界面的unicode码。



结点内容示例：

<a class=\"W\_texta W\_fb\" nick-name=\"\u9896\_\u5e7b\u7075\u4e4b\u9f99\" href=\"http:\/\/weibo.com\/234731093?refer\_flag=1001030103\_\" target=\"\_blank\" title=\"\u9896\_\u5e7b\u7075\u4e4b\u9f99\" usercard=\"id=5156170667&usercardkey=weibo\_mp&refer\_flag=1001030103\_\"\t\tsuda-data=\"key=tblog\_search\_weibo&value=weibo\_ss\_1\_name\" class=\"name\_txt W\_fb\">

结点内包含用户名，微博地址，微博内容等信息，使用xpath解析后可得到相应的结果。

# 四、数据预处理

微博短文本处理后会产生大量的特征项，仅通过1万条微博进行文本处理，选取情感描述性较好的形容词和名词，得到近2万的特征项，若将这些特征项直接用于后续处理，会严重影响算法效率，因此必须对文本进行预处理，包括中文分词、词性标注选择、停用词去除，以及特征选择。

图4-1 文本预处理流程图

## 4.1中文分词

中文分词是将汉字序列进行一系列切分工作，形成单个相对独立的词汇。文本处理中关键问题就是中文分词，分词效果会直接影响文本分析算法的精准度以及效率。英文文本中各个单词相互间以空格进行分割，中文文本词中没有一个类似空格的直接分界符，故中文文本分析与英文文本分析相比，更为复杂。

中文自动分词技术日渐成熟，在几十年的研究中，出现了许多使用较广的中文分词系统，本次实验我们小组采用中国科学院计算所研制的汉语词法分析系统ICTCLAS系统，该系统整合了中文分词、词性标注、新词识别等多项技术，分词精度高达98.45%，是当前使用最为广泛的汉语此法分析器。

ICTCLAS利用了层叠马尔科夫模型，融合汉语词语分析各个环节，在保证分词准确率的同时确保了系统的效率。主要思想如图4-2：

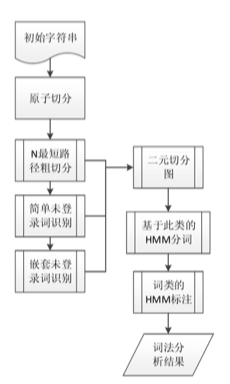


图4-2 ICTCLAS的中文词法分析

## 4.2停用词去除

在处理文本数据过程中，将某些代表性不强的字或词进行过滤，从而可以节省存储空间和提高搜索效率，过滤的文本就是停用词（STOP WORD）。这类词往往出现频率较高但对于文本标引作用低，若文本特征项中包含停用词，导致两个属于不同类别的文本因为包含停用词而难以区分，因此从中文分析得到的词语集，进行分离停用词操作是非常必要的，这个过程就叫做停用词去除。

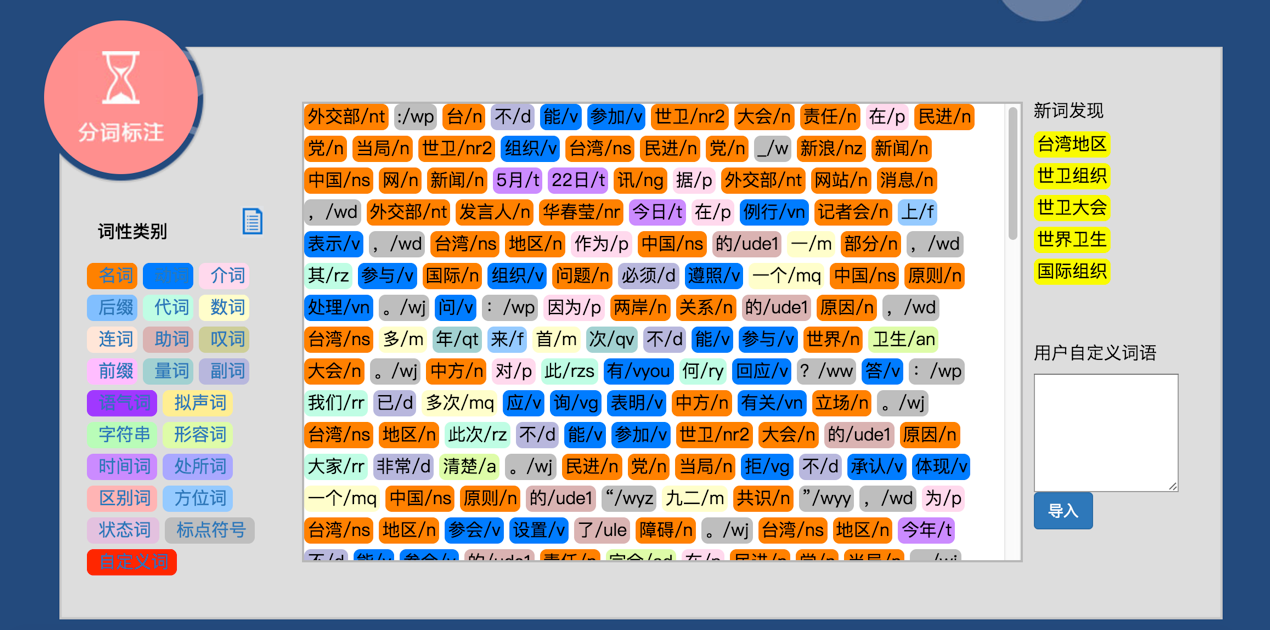
常用生成停用词表的方法是将词项按照文档集频率，是基于每个词项在文档集中出现的频率，将出现的频率进行降序排列，然后手动标注语义内容与主题相关度较低的高频词，将此类高频词整理为停用词，再查询词汇是否属于停用词列表。

停用词去除的目的在于可以使文本的表示更加准确，并降低特征空间的维度，从而提高后续处理效率和精确度。

本试验中我们小组采用jieba分词去除停用词。

## 4.3词性选择

中文文本通常由名词、动词、形容词、副词构成，而我们表达情感通常用描述性比较强的形容词，名词，副词等，所以直接利用ICTCLAS系统的词语词性标注，并进行描述情感较为强烈的词性词语，从而进一步提高分析的效率和准确率。

****

### 4.3.1特征选择：

微博是一种允许用户及时查看、更新、公开发布简短文本（通常少于140字）的微型博客形式，所以微博在实质上是具有特殊结构的短文本，如果采用文档频数方法，会忽略某些代表性较强的稀缺词，造成分类的精度；采用信息增益的方法，则会给特征集带来一定的噪声，不适合情感分析。

通过对微博文本进行分析，我们发现微博中某个特征项是具有情感特征的，则可认为该特征项在一定程度上反应了文本的情感趋势，需要将这特征项进行保留，这些词汇在研究中也被称为显性情感特征。在微博中，显性的情感特征包括表情符号和情感词等。这里，我们小组考虑当代青少年丰富的网络语用词，将网络用于也加入到显性情感特征中。

**4.3.1.1表情符号：**

表情符号指在网络媒体社交中使用特定的表情符号，在网络社交潜移默化的影响下，表情符号虽然就是简单的某个图案，但是它也承载了用户的情绪表达，让社交网络变得更为直观与生动，表情符号在网络社交中必不可缺。

表情符号在微博中使用广泛，如随机抽取几千条微博，包含表情符号的接近20%，根据相关数据，新浪微博用户更倾向于使用单个表情符号。

新浪微博提供的表情符号是用中括号包含的文本表达，这些表情的内容丰富，直接形象贴切地表达了用户的情感状况。实验中采用新浪微博自带的表情构建情感符号库，根据七类情感划分原则，将情感分为高兴、喜好、愤怒、悲伤、恐惧、厌恶、惊讶这七个类别。对这些情感符号统一化处理，建立情感符号表，然后通过查表的方式判断该情感符号是否属于情感符号表，若是则提取情感符号，通过转换成情感类别后写入情感特征表。

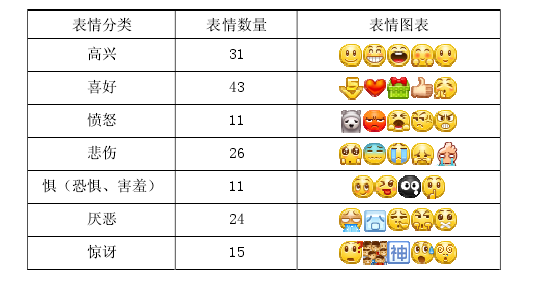


表4-1 情感类别和每个类别的典型表情符号

**4.3.1.2情感词**

微博文本中还包含着具有情感意义的词语或者俗语，表达着用户的心情，对微博情感分析也有着关键的作用，因此情感词典是微博情感分析的基础工作。

中文情感词汇复杂，仅仅基于词的词性进行情感词选择，这一方法不是很科学，如名词“垃圾”，这类名词往往具有一定的负面情感，但大多数名词本身并不具有情感色彩，大量选择会降低分类性能。本次实验采用打脸立功大学信息检索研究室提供的中文本体资源，将人的情感划分为7大类，21小类，这7类情感可以较为准确的体现用户情感状况。情感词汇本体统计整理了27466个词语，包含14986个成语熟语，网络词语78个，如表4-2所示，但对于当前微博中出现的网络词语，俗语等的统计收录不够完善

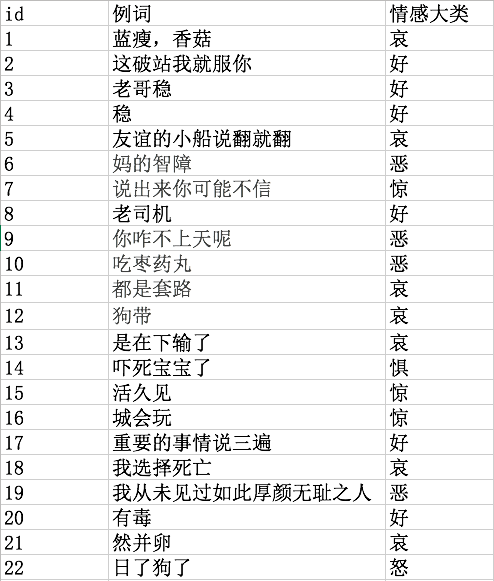
本实验采用《知网》的情感词典，并通过社交网络对当前流行的网络词汇进行整理。建立一个情感词典的情感词表。通过情感词汇匹配，判断分词产生的词汇是否为情感词，若是则提取该词汇，并写入情感特征表。部分示例：

表4-2 选用的中文本体资源情感分类表摘要

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 情感大类 | 情感类 | 例词 |
| 1 | 乐 | 快乐(PA) | 喜悦、欢喜、笑眯眯、欢天喜地 |
| 2 |  | 安心(PE) | 踏实、宽心、定心丸、问心无愧 |
| 3 | 好 | 尊敬(PD) | 恭敬、敬爱、毕恭毕敬、肃然起敬 |
| 4 |  | 赞扬(PH) | 英俊、优秀、通情达理、实事求是 |
| 5 |  | 相信(PG) | 信任、信赖、可靠、毋庸置疑 |
| 6 |  | 喜爱(PB) | 倾慕、宝贝、一见钟情、爱不释手 |
| 7 |  | 祝愿(PK) | 渴望、保佑、福寿绵长、万寿无疆 |
| 8 | 怒 | 愤怒(NA) | 气愤、恼火、大发雷霆、七窍生烟 |
| 9 | 哀 | 悲伤(NB) | 忧伤、悲苦、心如刀割、悲痛欲绝 |
| 10 |  | 失望(NJ) | 憾事、绝望、灰心丧气、心灰意冷 |
| 11 |  | 疚(NH) | 内疚、忏悔、过意不去、问心有愧 |
| 12 |  | 思(PF) | 思念、相思、牵肠挂肚、朝思暮想 |
| 13 | 惧 | 慌(NI) | 慌张、心慌、不知所措、手忙脚乱 |
| 14 |  | 恐惧(NC) | 胆怯、害怕、担惊受怕、胆颤心惊 |
| 15 |  | 羞(NG) | 害羞、害臊、面红耳赤、无地自容 |
| 16 | 恶 | 烦闷(NE) | 憋闷、烦躁、心烦意乱、自寻烦恼 |
| 17 |  | 憎恶(ND) | 反感、可耻、恨之入骨、深恶痛绝 |
| 18 |  | 贬责(NN) | 呆板、虚荣、杂乱无章、心狠手辣 |
| 19 |  | 妒忌(NK) | 眼红、吃醋、醋坛子、嫉贤妒能 |
| 20 |  | 怀疑(NL) | 多心、生疑、将信将疑、疑神疑鬼 |
| 21 | 惊 | 惊奇(PC) | 奇怪、奇迹、大吃一惊、瞠目结舌 |

**4.3.1.3网络词汇词典**

随着网络的不断发展，网络上出现了许多网络用词、网络语言等。所以我们小组特意加入网络词汇词典，收集整理100多个网络用词，分别进行情感标注以及统一化处理，建立一个网络词汇的情感词表，通过查表匹配判定词语的情感类别，并写入情感特征表。部分网络用词来源归类：



通过以上的表情符号词典、情感词词典、网络词汇词典，匹配筛选得到可以较好描述微博文本情感信息的特征词集。

# 五、基于频繁项集和语义的微博情感聚类算法实现

传统聚类算法多是采用Bag of words方法，容易引起短文本发生特征高维问题，因此需要挖掘词语间共现关系以及顺序关系，获取代表性较强的特征，进而降低特征维度。因此我们打算采用基于频繁项集的聚类算法。同时，在中文中，一个句子的意思除了考虑特征词，还应充分考虑句子的语义。而语义的研究主要是从语法意义角度去归纳，也就是词汇的各种组合。因此，我们小组采用基于频繁项集和语义的聚类算法

层次聚类分为三个步骤：

1. **频繁项集创建初始情感簇，每个初始情感簇文本都含有频繁项集**

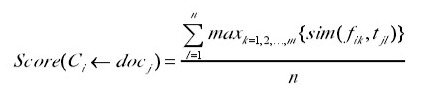
该步骤在[apriori.py](file:///Users/chenweiwen/Documents/ecnu2016/WeiBoDataMining/emotion/apriori.py)中，根据原来微博文本的频繁项集创建初始的情感簇

1.扫描文本集 D，利用词频统计候选项集出现的次数，收集满足最小支持度 min\_s 设定的项集，记为频繁项集；

2.利用产生的频繁 k 项集构造强关联规则，利用频繁 k 项集构造候选(k+1) 项集，反复迭代直至候选(k+1)项集为空。

**二、根据《知网》的中文语义相似度模型，根据最大语义隶属原则分离各个初始情感簇**

该步骤在[overlap.py](file:///Users/chenweiwen/Documents/ecnu2016/WeiBoDataMining/emotion/overlap.py)中，使每个文档隶属于一个情感簇中

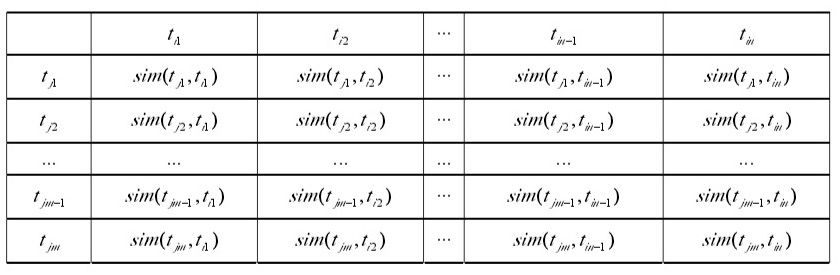
定义一个文档对于一个情感对隶属得分： 

其中sim来自《知网》的中文语义相似度模型

对于一个与多个情感簇关联的文档，计算其对于每个情感簇的隶属得分，取其最大，删去其与剩余情感簇的关联

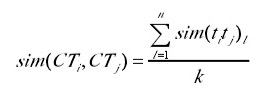
**三、通过定义簇间语义相似度矩阵，完成微博情感簇的层次聚类，并优化得到最终情感簇**

该步骤在[merge.py](file:///Users/chenweiwen/Documents/ecnu2016/WeiBoDataMining/emotion/merge.py)中，合并相似情感簇，得到最终聚类结果

定义两个情感簇的相似矩阵： 

其中sim依然使来自知网的模型

计算两个情感簇的相似度，取其前k大个



然后利用简单的双重for循环计算任意两个情感簇的相似度，若最大的大于阈值，则合并之，否则return

# 六、结果分析和性能评估

情感分析可以采用分类算法和聚类算法，分类算法作为有监督的机器学习算法，需要对文本进行分类，人工标注类别，在数据处理时工作量较大，因此我们小组打算采用聚类算法，无监督的机器学习算法，直接将相似的数据对象划分为不同的簇。

传统聚类算法多是采用Bag of words方法，容易引起短文本发生特征高维问题，因此需要挖掘词语间共现关系以及顺序关系，获取代表性较强的特征，进而降低特征维度。因此我们打算采用基于频繁项集的聚类算法。同时，在中文中，一个句子的意思除了考虑特征词，还应充分考虑句子的语义。而语义的研究主要是考虑不同词性词汇的各种组合。因此，我们小组采用基于频繁项集和语义的聚类算法：首先，基于频繁项集创建初始情感簇，每个初始情感簇文本都含有频繁项集；然后，根据《知网》的中文语义相似度模型，根据最大语义隶属原则分离各个初始情感簇；最后，通过定义簇间语义相似度矩阵，完成微博情感簇的层次聚类，并优化得到最终情感簇。

## 6.1获取频繁项集：

利用Apriori算法，利用频繁项集划分构造初始情感簇，将包含频繁项集词集微博划分为一个簇，得到初始情感簇。但因为一条微博可能包含多个频繁项而被分配到多个初始簇中，导致单条微博存在于多个情感簇中，因此需要消减重叠的部分，将重叠的微博分配给最合适的情感簇。

## 6.2基于语义隶属原则的初始簇重叠消除：

这里定义评分函数来衡量一条微博对于初始簇的情感隶属度。对于每条微博，我们仅仅将其保留在情感隶属度最大的初始簇中。这里用到《知网》的语义库拓展信息进行评分函数的设计与实现。只要对整个初始簇中的文本集扫描一遍进行评估比较便可消除重叠部分，同时删除空簇。

## 6.3基于语义相似度的凝聚层次聚类：

通过之前步骤获得的情感簇数量较多，还可将其归属于“喜怒哀乐惊惧恶”七大类情感，因此对其采用凝聚层次聚类，合并情感簇。这里我们抽取每个簇文本的特征向量，构建相似度矩阵，进行相似度计算，相似度计算公式采用《知网》词汇语义相似度计算公式，知道情感簇书目达到期望值，聚类结束。

# 七、总结

目前项目的数据获取、数据预处理包括分词、停用词去除、词性标注和特征选择的情感符号表、情感词汇表、网络用词表均已完成，数据存储在MongoDB中，正在对文本向量化表示和特征空间构建，同时目前也在算法实现中，因为算法是参考网上相应的博客教程以及国外的一些研究内容，所以实现起来难度较大，正在努力实现中。之后在结果分析时，将对算法分析的准确率与传统的聚类算法进行对比分析。此外，我们希望把该数据挖掘算法应用到实际应用中，准备开发一个简单的微博话题情感分析系统，获取微博相关事件的话题标签，用户点击选择后可以生成对应的针对该话题的用户情感趋势变化图表。估计两个星期的开发周期可以基本完成任务。