

湖北工业大学

# 毕 业 设 计（论 文）

题 目 基于增量聚类的网络热点发现研究

姓 名	<u>徐璐</u>
学 号	<u>1110831116</u>
所在学院	<u>经济与管理学院</u>
专业班级	<u>11 信管 1</u>
指导教师	<u>黄炜</u>
日 期	<u>2015 年 5 月 30 日</u>

## 摘 要

伴随移动互联网时代的来临，人们从外界获取信息的方式更为多元，每时每刻在互联网产生的数据量呈指数级增长，我们正面临着“信息过载”的困境，因此如何从这些海量互联网数据中追踪当下的热点事件的传播与演化正是我们所要研究的重点。本文所论述的网络舆情预警信息系统结合了网络爬虫、分词处理、增量聚类、数据可视化等技术，使之能够实现对国内主流门户网站等的新闻信息获取及文本聚类分析，借助于相应的技术手段，实现对网络热点话题发现的功能。为了能利用互联网信息时变性的特点，本文所述的网络舆情预警信息系统采用的是基于 T-Single Pass 增量聚类算法实现的文本分析模块。该算法与传统静态聚类方式不同的是，它通过以互联网信息发表时间为序列来进行数据流式的增量聚类，利用了每一步聚类计算的结果，从而实现对先前聚类的动态更新。通过将该系统中增量聚类模块与传统的基于静态聚类的 Simple K-Means 算法进行相互比较之后可以发现，该算法在对网络热点事件的挖掘中具有明显优势。

本文是参与指导老师的国家自然科学基金项目“微博环境下实时主动感知网络舆情事件的多核方法研究”（批准号：71303075）的工作成果。

**关键词：**舆情分析，Web 挖掘，增量聚类

## Abstract

With the advent of the mobile Internet era, the way people get information from the outside world is more and more diverse, the amount of data generated on the Internet every time grows exponentially, we are faced with the "information overload" dilemma, so how to track and detect the current hot events and evolution of relevant topics from these massive Internet is what we want to research next. The network public opinion information system which is discussed in this thesis combines the web crawler, word processing, incremental clustering, and data visualization technology. It can realize the function of data analysis of the mainstream domestic portal websites such as news and information acquisition and text clustering. With the help of the corresponding technology, it can realize the function of the hot topics detection. In order to make use of the Internet information's feature of time variation, the text analysis module of this network public opinion warning information system described in this article based on T-Single Pass incremental clustering algorithm. What this algorithm different from the traditional static clustering algorithm is that the method use Internet information's feature: time sequence, to analyze data flow of incremental clustering, by making use of the every step of clustering calculation results, it can update the previous clustering result dynamically. By comparing the incremental clustering algorithm and the traditional simple k-means algorithm, we can draw the conclusion that this algorithm has obvious advantages in the network hot event mining.

This thesis is the achievement of participating in my instructor's National Natural Science Foundation project (Grant Number: 71303075), "network public opinion research multicore approaches events in real-time micro-Bo environment initiative perception".

**Keywords:** Public Opinion Analysis, Web Data Mining, Incremental Clustering

## 目 录

摘 要 .....	I
Abstract .....	II
目 录 .....	III
1 引言 .....	1
1.1 选题背景 .....	1
1.2 研究意义 .....	1
2 研究综述及论文组织结构 .....	3
2.1 国内外研究综述 .....	3
2.2 本文写作思路 .....	4
2.2.1 本文创新之处 .....	4
2.2.2 论文组织结构 .....	4
3 基于 T-Single Pass 的增量聚类算法 .....	6
3.1 词频向量获取 .....	6
3.2 T-Single Pass 算法流程 .....	7
4 网络舆情热点发现系统 .....	10
4.1 互联网数据源获取 .....	10
4.1.1 网络爬虫 .....	10
4.1.2 预处理与分词 .....	11

4.2 增量聚类 .....	14
4.3 舆情预警 .....	18
5 系统设计与实施 .....	21
5.1 详细设计 .....	21
5.2 系统运行 .....	24
5.2.1 系统简介 .....	24
5.2.2 运行环境 .....	25
5.2.3 运行演示 .....	26
5.3 实验结论 .....	33
6 总结与展望 .....	35
6.1 全文总结 .....	35
6.2 研究展望 .....	35
参考文献 .....	37
致 谢 .....	38

# 1 引言

## 1.1 选题背景

近几年来，时代见证了互联网特别是移动互联网的良好发展态势，公众对信息的获取能力越来越强，从 10 年前的门户网站及后来各大 BBS 论坛的兴起，再到后来以人人网、QQ 空间为代表的基于 PC 互联网时期的平台，直至今天流行的微信、微博等广受人们喜爱的移动互联网新型互动平台。大数据时代，我们时刻暴露在各种良莠不齐的信息海洋之中。每一种新型互联网技术的更新都伴随着对普通民众生活各方面的极大影响，现如今，以互联网为信息传播平台的方式越来越受到人们的欢迎。

现如今，我们所接触到的行业都已经被不同程度的“互联网化”。CNNIC 在 2014 年 7 月份发布的有关数据显示：截止到统计结束时间，我国的网民总数已经超过了 6 亿人。今天，互联网作为一项民众生活中的基础设施在我国城乡居民当中的普及率已经超过了 46.9%，其中，使用移动终端设备（如 iPad、搭载 iOS、Android 等操作系统的智能手机）进行上网的用户数量首次比使用传统方式（如家用 PC、笔记本）上网的网民数目更多。智能手机以其相对比较低廉的价格、极佳的便携性等优点在普通民众中的渗透率不断激增，与此同时，随着微博、微信等新媒体平台的出现，以及今日头条等具有新闻个性化推荐功能的 APP 成了人们获取信息的重要来源，这也就意味着对于某一重大事件的爆发人们将会在极短的时间内获知。如果对某事件歪曲解读，结果会十分恶劣。

据有关报道分析，大部分普通群众对于某一特定热点事件本身的辨识力并不高，对该事件所持有的情感态度也并不明确，但是通过少数意见领袖的观念引导，借由互联网低门槛的传播，大众的社会情绪很容易被煽动，从而导致热点事件的病毒式传播，更有甚者歪曲事件本身。它是一定时段内相关事件通过互联网的方式进行扩散的，此类事件的爆发往往反映着当时社会群众的情感认知与舆论导向，普通群众在接受相关信息输入或者经由意见领袖的引导，很容易对事件本身的性质丧失正确的判断力，如果不加以及时捕捉并采取相应措施来控制，严重的话可能造成社会动乱。

网络舆情研究是一门综合性和实践性很强的领域，它涉及了 Web 数据挖掘、机器学习以及社会科学等领域。在这样一种背景下，为了及时控制舆情的蔓延，该研究就很有必要了。目前我国舆情热点事件有着如下几个显著特点：（1）相应事件的突发数量呈现递增的趋势；（2）网络上的“分派现象”愈加明显：由各个意见领袖引导的各类不同思潮呈现出不断争鸣的态势；（3）基于微博、微信、以及新闻媒体客户端的传播趋于独立。

## 1.2 研究意义

由于互联网信息呈现出数量巨大、时变性强、非结构化明显等特点，那么如何对这些已有信息进行有效分析一直是国内外学者研究的重点，聚类在整个网络舆情信息系统中都占据着非常重要的位置。聚类通过相应算法将特定数据对象与其本身相似性最大的组合在

一起。

但是，传统的聚类算法只能获取数据的静态聚类结果，而互联网信息具有非常明显的时变性，在某一时间维度流行的话题和热点事件很可能在接下来的时间范围里发生很大的变动，静态聚类在面向互联网信息的分析中已不能满足要求。为了掌握话题随时间而发生变化的特点，在对传统增量聚类算法 Single Pass 的基础之上进行改进，本文所述的舆情分析信息系统增量聚类模块引入了时间维度，即依次将每条互联网新闻文本以信息产生时间为基准的数据流形式输入到增量聚类模块中，依次同之前已经计算的结果进行比较，并实时动态更新已有的簇中心，进而实现基于数据流式的增量聚类算法 T-Single Pass。

## 2 研究综述及论文组织结构

### 2.1 国内外研究综述

在有关网络舆情的相关研究、热点话题发现系统的设计，以及核心算法方面，众多学者都有发表一些独到的见解与成果。

高燕飞等定义了<sup>[20]</sup>什么是增量聚类，以及对于聚类的几个常见类别。他在现有研究成果的基础之上提出如果想要大幅度减少算法时间复杂度和内存开销，可以尝试只对下一步新增加的数据进行聚类。该算法利用特性向量的表示方式进行聚类，实验数据显示，该算法与一般的聚类方式进行对比，算法所消耗的时间明显降低。但是并没有给出某一种具体的算法实现。

张小明、李舟军等提出了<sup>[13]</sup>借助于增量聚类的手段来研究话题的自动发现，这种方式的目的在于极大程度上提高检测话题时的效率。该方法还可以自动发现出语料库中相关话题的数量，对于特征权重值的计算也进行了相应的改进使之更符合要求，为了提高文本聚类的准确性，该方法采用了自适应的方式来提炼语料库中主题识别能力相对比较强的文本特征，并且在聚类的时候借助于 BIC 来获得主题类的数量。与此同时，通过借助话题本身的延续性这一特点来对文本语料进行预聚类，这样一来话题检测的效率就会大大提高。

潘敏、王明文等<sup>[1]</sup>利用文本的簇特征来进行增量聚类的方法，这种算法是使用广泛使用的 K-Means 聚类算法来预先对最开始的数据集产生初始聚类。在聚类的时候把先前上一步的聚类得到的每个类的文本数量、质心、方差以及均值等都先保存下来，以此作为相应簇的簇特征。当接下来对新数据进行聚类的时候，就可以使用之前的簇特征来对新增数据集产生二次聚类。然而在处理互联网数据时没有考虑其本身的特征。

王丹等人提出了<sup>[10]</sup>通过改进的层次聚类算法来对网络中发表的帖子进行聚类分析的方式来获取相关热点话题，基于此概念对词集提出了高权重的概念。利用这种设计理念实现了相关的增量聚类算法，通过相关的实验数据测试发现，这种算法能够较好匹配动态数据，而且算法复杂度也降低了很多。但是对帖子主题和内容获取的权重没有涉及。

税仪冬等<sup>[18]</sup>在针对增量聚类过程中出现的主题模型不是很准确、误检率与漏检率会随着聚类过程的进行而被放大的问题，提出了将文本周期性分类和 Single-Pass 聚类进行有机结合的方式来进行热门话题的识别与追踪。它的思想是当网络爬虫采集到的数据积累到了一定的数量之后，就对之前已经产生聚类的文本按特定的周期来进行分类，这样可以提高话题簇的精确度，这样后续热点事件发现的精度就会大大提高。

于翔等提出了<sup>[12]</sup>基于网格的数据流聚类方法，该方法使用动态划分的策略对新的数据空间展开研究，使其能够完成对网格单元结构和统计信息的增量更新。在这样一个理论基础上，构思了基于动态网格划分的增量聚类算法，这种算法除了有和传统的网格聚类算法一样的高效率优点之外，最终也可以得到更为准确的结果。却没有涉及新闻类数据的时效性。



在国外相关的研究中, Agostino Forestiero 提出了<sup>[17]</sup>基于 multi-agent 的数据流式聚类方法来对相似数据点进行分组。Issei Sato 等提出了<sup>[21]</sup>基于 LDA 的改进模型, 为了处理同一时刻接收到的文档, 该算法不需要对所有数据存储器旧的分析结果, 因此可以极大提高处理大规模数据时的效率。

对于网络舆情热点话题的发现, 学术界有许多优秀的聚类算法, 在本论文的指导思路方面与系统概念设计方面, 以上这些优秀学者的研究成果都为最终本系统的设计与实现提供了很好的思路。但是分析以往的网络热点话题发现系统的设计就会发现, 它们大多只是单纯的遵循数据获取、静态聚类、产生结果这三个步骤, 并没有考虑数据本身的一些特征, 例如对于同一篇文档按照不同的次序进行聚类往往结论也不一样, 或者是对于一篇新闻文本中的同一词语, 如果它在该文本中出现的先后顺序不一样, 也往往说明它们的重要性会有比较大的差别。因此本论文所述的网络舆情预警信息系统所利用到的时间值通过相应 API 解析互联网新闻页面中的标签属性, 将其融入到增量聚类算法当中去。

## 2.2 本文写作思路

### 2.2.1 本文创新之处

由于互联网信息具有明显的时变性、不稳定性, 为了研究热点话题随时间变动的特征, 在基于传统经典增量聚类 Single Pass 算法基础之上进行了少许更改。本文所论述的内容采用了这样一种处理方式: 将互联网新闻文本按照时间的先后顺序进行排列, 信息发布时间越靠前的新闻文本就越早进入到聚类模块, 然后依照此顺序进行流式聚类, 此处假定大多数互联网新闻的发布时间与对应的热点事件的爆发点接近, 因此对有关新闻的聚类分析即可反映该事件的相关演化特征, 然后依次以文本数据流的形式作为输入进行增量聚类, 用新接收的数据来对已有的话题簇进行更新。与此同时, 结合这样一个事实, 即大多数新闻的标题对相应事件所对应的新闻文本内容已经具有较高程度的概括性, 故本文将标题进行切分得到的关键词与从新闻文本语料正文内容当中提取到的关键词进行不同程度的加权处理, 从而使得最后的结论更能反映事件本身的特点。

话题热度值的计算标准如公式 2.1 所示:

$$\text{Hot Spot} = \lambda \text{Weight Of Title} + (1 - \lambda) \text{Weight Of Content} \quad (2.1)$$

在该舆情热度值计算式中,  $\lambda$  为平衡新闻标题与新闻内容的权重系数,  $\lambda=1$  时, 仅考虑标题的话题表述贡献值;  $\lambda=0$ , 仅考虑内容的话题表述贡献值。此处,  $\lambda$  的值应该根据实验的需要来进行取值,  $\lambda$  值太大不能体现新闻内容的特征, 会有数据稀疏的缺点;  $\lambda$  值太小则会掩盖新闻标题本身传递的信息。在实际的测试数据中, 常常将  $\lambda$  设置为 0.1 到 0.2 左右较为合适。

### 2.2.2 论文组织结构

本文一共分为六个章节，其中：

第一章为本论文所述领域的研究背景与理论意义部分，主要介绍了该课题在互联网新环境下对应的研究意义与背景，详细描述了当前该领域的一些新特征。

第二章为国内外学者的研究现状综述，主要介绍了其最新的研究成果，以及对这些已有聚类方式和网络热点话题发现系统设计的简要点评。

第三章介绍了本系统中增量聚类算法相关的一些数学理论，包括著名的 TF-IDF 算法，本文中用来比较不同新闻文本相似度的计算标准——余弦相似度，以及本文基于 Single Pass 增量聚类算法的修改后版本，T-Single Pass 算法。

第四章介绍了本套网络热点发现系统所用到的一些技术，包括基于 JSoup 技术的网络爬虫，关于常见的聚类算法的简要介绍，以及本系统所使用的数据可视化技术手段等等。

第五章主要论述了网络舆情预警信息系统具体实现模块的相关细节以及系统运行流程，包括数据获取与预处理模块、基于 T-Single Pass 的增量聚类模块、作为信息可视化处理的舆情预警模块，以及系统界面展示、用户管理功能模块、相关的数据加载以及最终的运行结果等。

最后一章总结了本文的内容与还需要值得修改与完善的地方，以及本人对网络舆情研究、热点话题发现领域相关范畴的展望。

### 3 基于 T-Single Pass 的增量聚类算法

#### 3.1 词频向量获取

对于从互联网上获得的新闻文本并不能直接作为聚类的输入，首先经由分词系统进行分词处理（本文论述的网络舆情预警信息系统采用 lucene 分词系统的 API），并根据我们的实验需要移除停用词，经过初步预处理步骤之后再利用 TF-IDF 算法统计相关词的词频大小，TF-IDF 是常常用来度量特定词语对应于某文本内容在整个文本数据集当中重要性大小的手段。它能够很好地衡量词语对某一文本的话题表征。经过该步骤处理后每篇文本分别对应于一个如公式 3.1 所示的向量：

$$\text{document} - \text{featureVector}(a_1, w_1, a_2, w_2, \dots, w_i, t_i) \quad (3.1)$$

其中  $w_i$  表示对应的文档中的第  $i$  维属性值。

下面来介绍一下如何将一篇新闻文本映射到其对应的特征向量：

TF 为词语在文本中出现的频率，频率越高则重要指数越高。某词语出现在文本中的频率如公式 3.2 所示：

$$TF_{ij} = \frac{n_{ij}}{\sum n_{ij}} \quad (3.2)$$

IDF 是对于某一个特定的词语在语料库中的重要性的一种度量标准，IDF 的主要设计思想是：如果含有特定词的文档数量越少的话，IDF 值就会越大，就表明该词语同其他类别相比的话有很好的区分能力，如公式 3.3 为其计算标准：

$$IDF = \log \left( \frac{\sum n_{ij}}{\text{num} + 1} \right) \quad (3.3)$$

其中， $\sum n_{ij}$  表示新闻文本总数量，num 表示含有该词的文本数量。

某特定词语如果数据集中的文本内频繁出现，分母值就越大。但是为了防止出现 0 作为分母的情况，所以通常需要进行略微处理。例如某篇新闻中，一些地名由于具备成为热度词语的可能性，故并不能将其视作停用词而处理掉。但是在某些文档中出现频率高并不能代表这一类文本的簇特征，所以此处需要对其进行加权，得到相对比较平衡的处理方式就是 IDF 算法的设计初衷，经过此步骤的处理，具有高权重 TF-IDF 值的词语就是我们所希望获得的关键词。TF-IDF 算法如公式 3.4 所示：

$$TF - IDF = TF \times IDF \quad (3.4)$$

经过 TF-IDF 词频计算，就可以获取每篇新闻文档的词频向量值，这样就将一片新闻文本映射到了其对应的特征向量。

### 3.2 T-Single Pass 算法流程

基于 T-Single Pass 增量聚类算法，它通过预先设立一个由用户给定的阈值 Threshold 作为基础，然后在与其他文本按顺序依次对比得到相似度，若是结果大于该阈值，表明相互比较的文本属于同一类，并且在此基础上动态更新已有的类中心，该算法利用了先前聚类的结果来实时的对聚类结果进行更新，因而实现了动态增量聚类，弥补了传统的静态聚类不能利用之前聚类结果的缺点，T-Single Pass 的这种算法设计理念在对流式数据的聚类处理方面具有很多优越性。

为了计算文档之间的相似程度如何，常常借助 Cosine Similarity 来作为其衡量标准。若这两篇文档经过计算的结果为 0，则说明向量互相垂直，则对应的两篇新闻文本没有相关性，若为-1，则说明完全相反，若为 1，则说明相似度最高，则认为对应的两篇新闻文本所要表述的主题完全一致。由于新闻文本内容较长，会存在维数较高的情形，对于需要进行比较其内部主题相似度的两篇文档而言，余弦相似度这一比较算法可以作为这方面很好的衡量标准。

如图 3.1 所示，不同文本的相似度就可以映射到其对应的特征向量的夹角  $\theta$ ， $\theta$  值越大，说明相似性越低； $\theta$  越是趋向于 0，说明它们越相似。

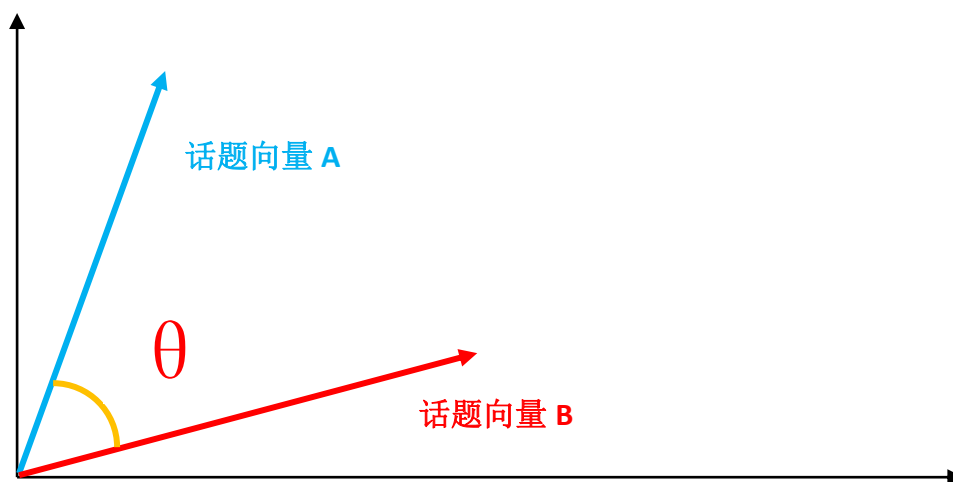


图 3.1 文档相似性的向量度量

在相似度评价方面还有其他衡量尺度。但是在文本挖掘相关的方面，余弦相似度往往被用来作为某特定集群内部相似度的衡量尺度。因此本文所述的网络舆情预警信息系统需要进行比较的文本相似度也采用这一算法。具体计算公式如公式 3.5：

$$\text{CosSim} = \frac{\bar{A} \times \bar{B}}{|\bar{A}| |\bar{B}|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (3.5)$$

T-Single Pass 算法详细流程如下：

**Algorithm:** T-Single Pass Clustering

**Input:** feature Vector

**Output:** Clustering result and Hot words

1. 初始数据处理：包括对从互联网上爬取下来的新闻文本规范化、分词处理、去除停用词、TF-IDF 计算权重，根据计算得到的权重值映射得到特征向量。
2. 按新闻发表时间进行排序，首先读取第一篇新闻文本对应的特征向量进内存，将它作为第一个簇的簇中心。
3. 分别读入每一篇新闻的特征向量，依次与每一类话题进行余弦相似度计算，若相似度大于之前给定的阈值，则将该新闻归并入该类，同时更新话题向量。若小于它，就建立新类。

图 3.2 所示为 T-Single Pass 聚类流程。

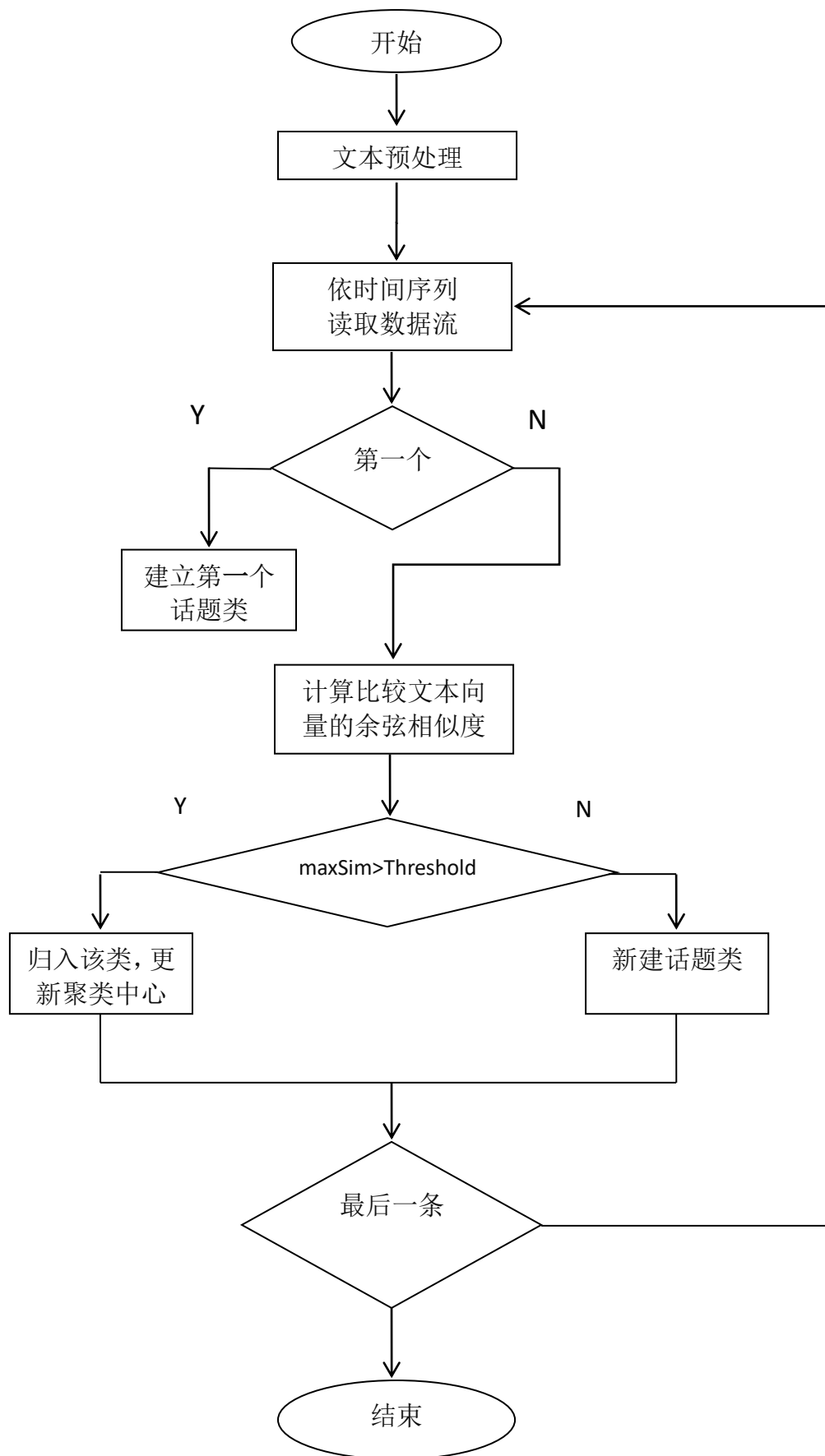


图 3.2 T-Single Pass 聚类流程

## 4 网络舆情热点发现系统

### 4.1 互联网数据源获取

#### 4.1.1 网络爬虫

由于互联网信息海量、结构繁杂，因此必须借助计算机来辅助信息的收集。网络爬虫（Computer Spider）可以根据需要以某种机制自动获取互联网页面的信息，常常被用在搜索引擎之中用来获取 URL-地址映射。网络爬虫需要根据相应的方式去除不相关的 URL 来对互联网页面进行爬取，这样就保证了所要爬取的 URL 同主题之间的较大相关度，接下来依次将有关 URL 链接放入 URL 队列。然后从 URL 队列中获取接下来要操作的 URL，计算机程序会一直在满足相应结束条件之前不断重复之前的过程，当获取内容达到要求时停止整个流程。互联网信息的复杂、多样、海量、非结构化等特性使得我们不可能通过手工、低效率的复制方式来一个个收集相关数据。基于网络爬虫的这些优点，我们可以采纳网络爬虫这项技术来从指定的网站页面获取相应的原始文本信息。

网络爬虫所需要解析的 html 网页结构的文档对象模型如图 4.1 所示。

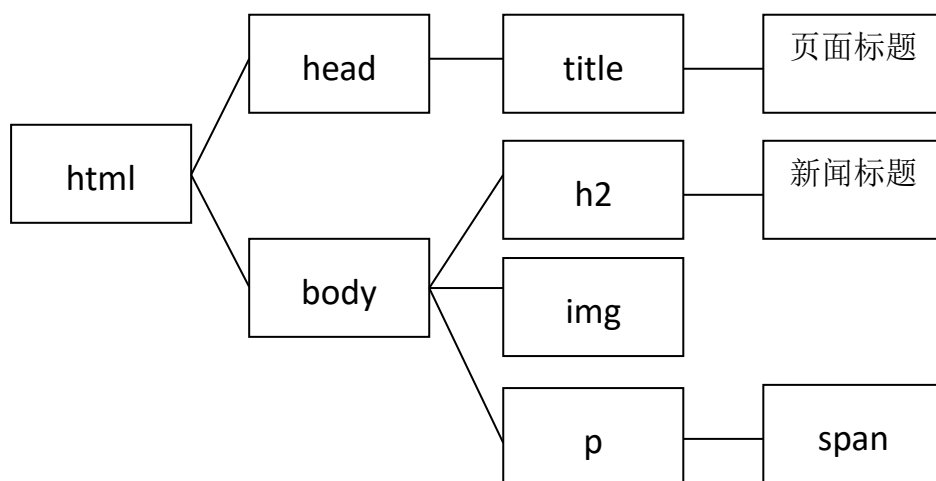


图 4.1 html 文档对象模型

由于 html 是一种高容错性的标记语言，因此 Web 网页开发人员在编写有关页面的时候，如果由于粗心或者书写不规范等原因对于某些标签没有按照标准来进行书写，常见的不规范如标签未闭合，以及没有正确书写属性。在本文完稿之时，尽管 W3C 组织早就已经在半年前发布了最新的 html5 标准，但是各大浏览器对相关属性的支持却是大相径庭，因此网页设计人员必须额外编写不同的代码来进行不同浏览器的适配，这些问题都会干扰爬虫的正常解析。随着 Web 技术的不断发展，被应用到前端的新技术也不断涌现，如 AJAX、SVG、Canvas 绘图、XML 等都需要爬虫具有更高的标准。图 4.2 是爬虫运行主流程。

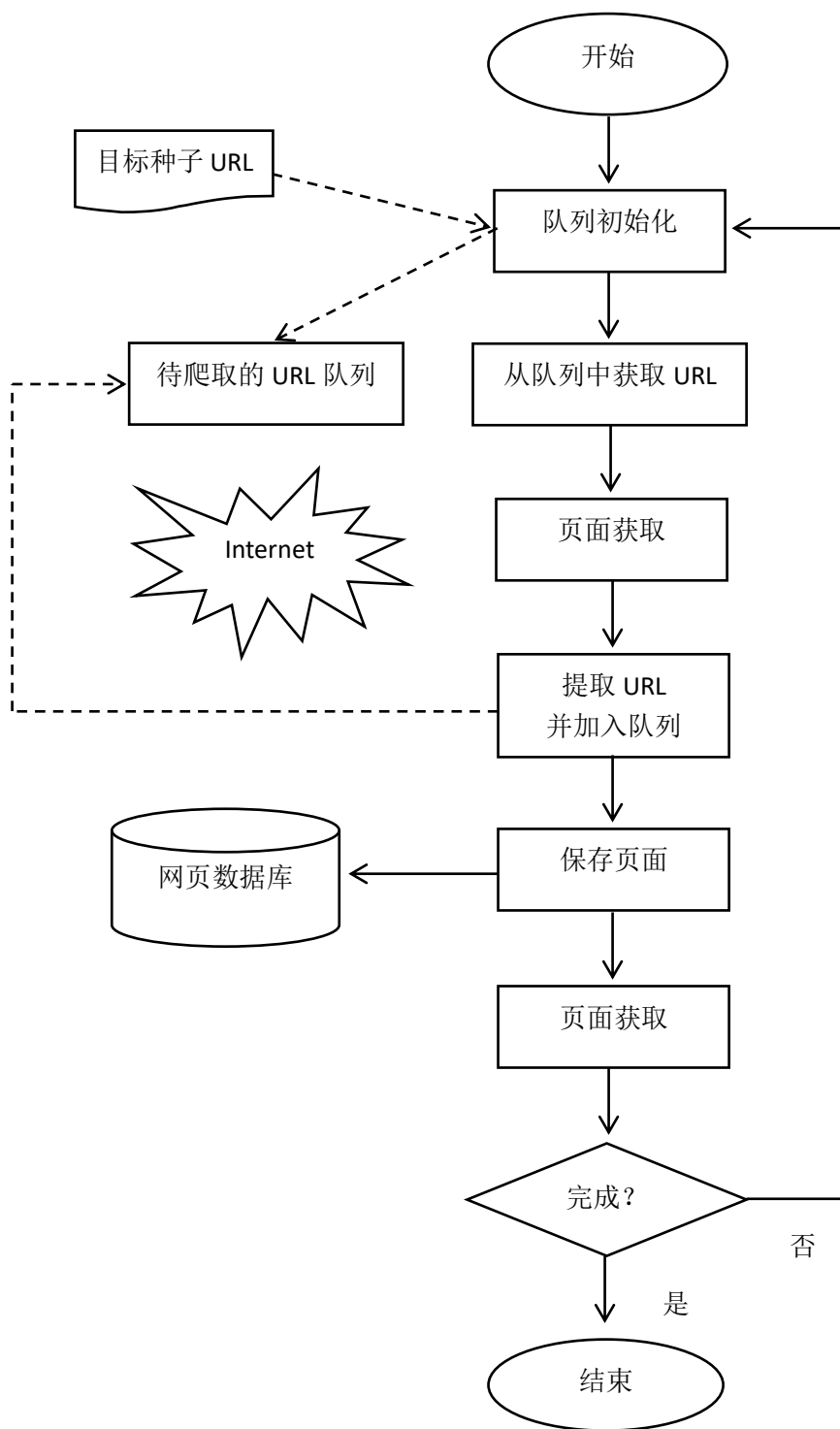


图 4.2 爬虫工作流程

#### 4.1.2 预处理与分词

在利用网络爬虫锁定种子 URL 之后，因为此时原始页面中存在着大量无用、不规范的标签与属性，所以并不能直接作为文本聚类的输入数据。因此必须要进行数据预处理，以



剔除 html 页面中无用的标签，从而只保留对我们有用的信息。经过这一步骤之后，我们还应该根据需要去除对聚类结果的主题反映没有帮助的停用词，例如“的”、“啊”、“是”、“在”等等虚词，这些词语尽管可能有比较高的权重，但是对我们的分析并无帮助。在一般的 Web 数据挖掘过程中，数据预处理的主要步骤如图 4.3 所示。

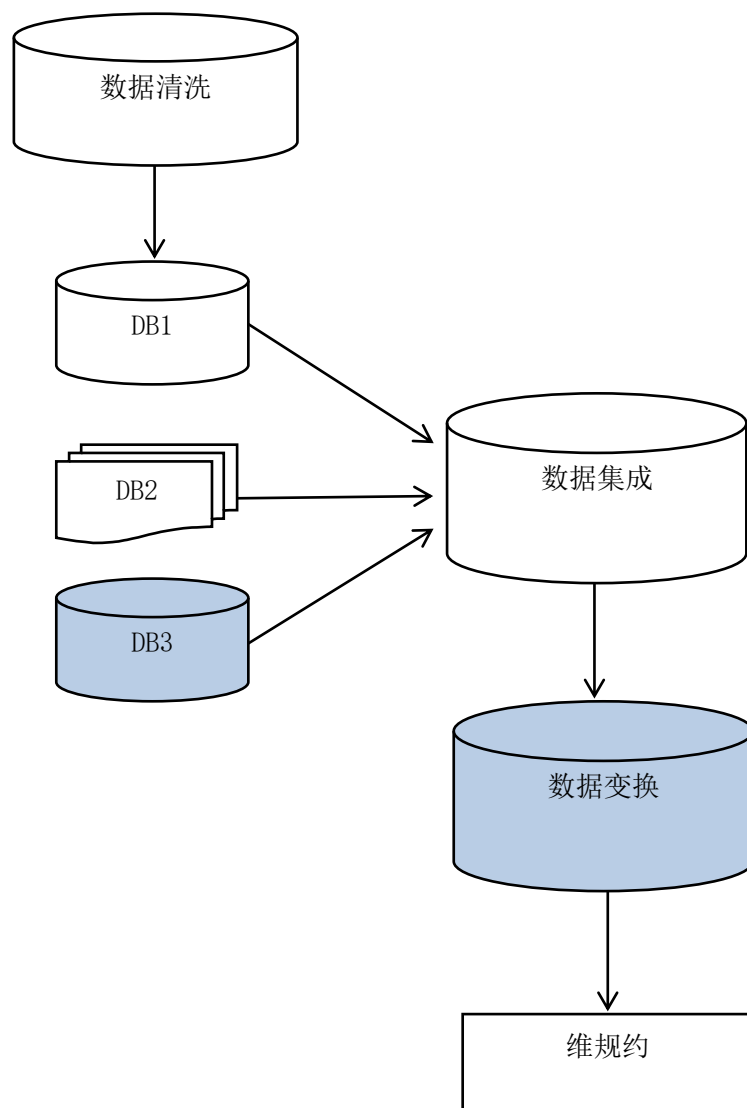


图 4.3 数据预处理常规步骤

从互联网上利用网络爬虫获取原始文本集之后，文本本身并不能直接用来分析，必须通过某种机制进一步处理来表征每篇文档的特征，汉语表述的互联网文本信息与英文有较大差异，英文文档可以直接以空格和标点符号作为分割来获取词语集，但是汉语的信息表达习惯是以句子为单位，因此需要对每篇文档进行分词处理。经过此步骤之后就可以获得每篇文档进行语义划分之后词语集合。

本系统的设计与实现采用的是基于 lucene 的分词系统进行分词，lucene 是一套优秀的面向对象架构设计的开源软件，它提供了一系列易于使用的 API，极大减小了开发门槛，

在该步骤设计分词功能模块的时候仅需导入相关的包然后调用其 API 即可，并不需要重新设计分词算法与编写相关的程序。同时，由于整套网络舆情预警信息系统设计与实现的后台业务逻辑模块与算法部分均是基于 java 语言编写的，这里采用 lucene 来进行分词也兼顾了整套系统的开发环境平台一致性，便于后期维护以及功能扩展。

图 4.4 演示了分词。

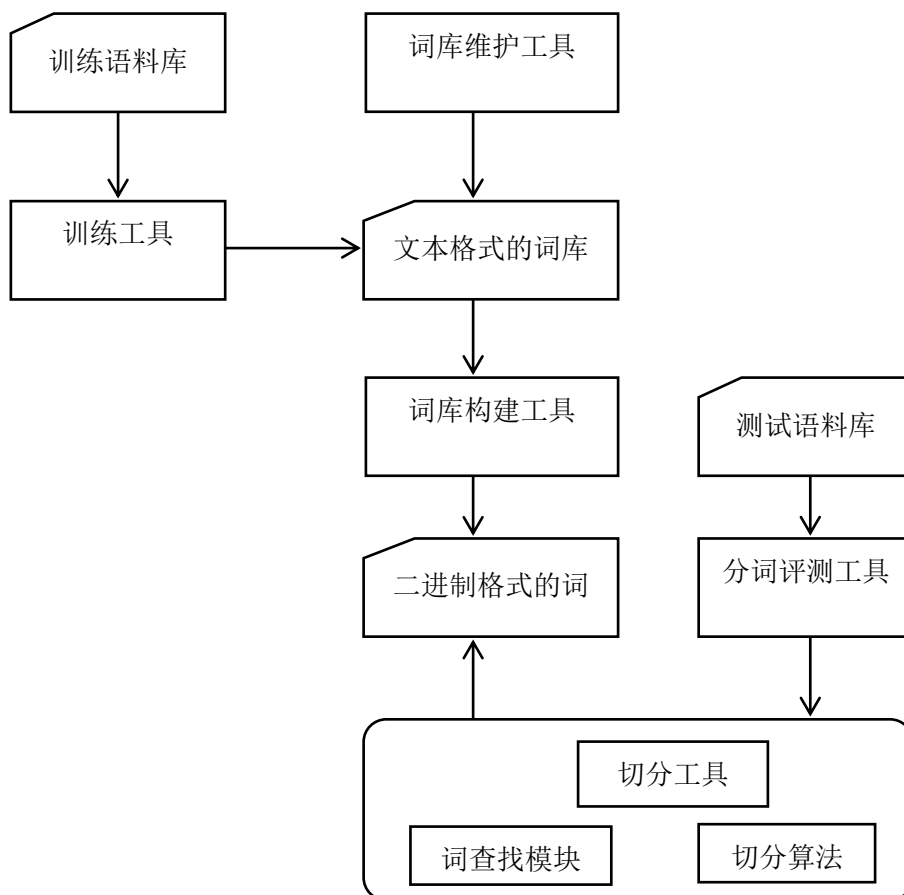


图 4.4 分词处理常规步骤

我们常见的有很多不同种类的数据集，但是这些数据集大都有着很相近的特性，在此先引入三个数据集特征的描述名词：第一个是维度，它代表着一个数据集之中数据对象的属性多少。通常属性数量比较多的我们称之为“高维数据”，它可以更详尽的表述一个数据对象，但是在数据挖掘中如果属性数量太多而之前不加以处理的话就很容易陷入“维灾难”。这样一来，无论算法设计的再精巧、优化的再好，在进行特征值计算的时候所需要的时间和内存开销是非常巨大的。而且数据将会越来越稀疏，在下一步的聚类过程中，每个点之间定义的距离和密度就失去了其原有的含义，聚类结果的质量将会大幅度降低。因此，我们在数据挖掘中对于高维数据往往会采取“维规约”的方式来对其进行降维处理，使得数据能够达到允许的范围之内。第二个是“数据稀疏性”，该术语的意思是说，对于在一个数据集的对象中，如果该对象有很多属性，但是大多数属性值是 0 的现象。这样虽

然对于对象的表述信息会有所缺失，但是另一方面由于在计算过程中非 0 属性才需要进行比较，所以会大幅度降低时间和空间开销。第三个是“分辨率”，对于相同的数据，它在不同的分辨率环境下它的性质也会有所差别，若太高可能无法分析出相应的数据模式，若过低则很容易出现检验不到数据模式的情形。因此，无论是做分类还是聚类，第一步的数据预处理步骤都在整个过程中都很重要，它直接影响接下来的聚类结果。

## 4.2 增量聚类

聚类又被称作是“群分析”，它是将数据集中的数据对象进行分组归类，经过这样的一个处理步骤之后，使得在同一个类的对象相似度最大，而不同的类的相似度最小。它在很多领域内都有着非常广泛的应用，例如在生物科学领域，从最早物种系统分类学的诞生，一直到今天以分析不同生物遗传信息的 DNA 序列大数据挖掘，聚类都很重要。对于搜索引擎，用户的每一次搜索请求提交，其背后都涉及到聚类算法的支撑，以保证所返回的结果相关度是最大的。相似度是分析的依据，归属于同一簇之内的相似度比不属于同一簇的具有更多的相似性。

以下几种是在数据挖掘领域内常常使用的聚类算法：

（1）基于划分的聚类主要思路如下：预先设立要创建的划分区域的值  $m$ ，第一步建立一个初始化的划分区域。接下来它运用一类迭代方式的重定位技术，通过把要聚类的对象从某一个簇转移到另一个簇来进行区域划分。划分好坏的一般评判标准是：位于相同簇中的聚类对象理论上来说应该尽可能具有比较高的相似度，位于不同的簇中的聚类对象应该具有尽可能大的相异性。一般而言，常见的区域划分方法可以扩展到某特定空间的子空间来进行聚类，而不是在原始的整个数据空间上来进行搜索。当属性数量巨大而且数据特征稀疏的时候，这将会带来极大的好处。一般而言，若想要达到全局最优化的标准，基于划分的聚类这种方式可能需要利用枚举法来列举出所有可能产生的划分，这样一来所造成的计算量是非常大的。然而实际上，数据挖掘领域常常采用普遍的启发式聚类方法来设计聚类算法与相关的应用，其中最熟悉的如 K-means 和 K-NN 算法均采用了这一方式，算法为了达到局部的最优解状态，往往会在聚类的过程中以某种机制来渐渐提高每一步聚类的质量。对于中小型数据库中的球形数据，这类启发式算法很适合用来解决这些问题。如果想要在复杂簇以及超大型数据集的基础上进行聚类，则往往有必要对当前方法进行相应的扩展。

（2）基于层次聚类算法：这种聚类方法的设计思路主要是对所聚类的原始数据集按层次进行分解，直到某种条件被满足的时候停止。它是各个嵌套簇的集合，它们以某种方式形成一棵树。在这颗树中除了叶子节点之外，该树中的每一个节点均是其子节点的合并，如果某个节点包含了该数据集中的所有对象，那么它被称作“树根”。它通过产生一个个树状图的方式来完成整个聚类过程，“合并”与“分裂”是两种常见的基于层次的聚类方法。合并方式的聚类主要设计思路就是从树的最底部开始聚类过程，为了形成上一层次的

聚类结果，则必须要以合并子节点中距离度量最近的节点。然后依次类推，一直到所有的数据全部都归到根节点中才停止整个聚类过程。然而，分裂方式的聚类设计思路与合并方式的聚类思路正好相反，它是从根节点开始整个聚类过程，然后接下来依次将该根节点分成一系列子聚类结果，每个子聚类结果再将自己分解，并且一直重复此步骤，直到每个最终的聚类中仅仅包含一个数据节点时就停止整个流程。在 Web 文本挖掘的有关聚类算法中，基于合并的方式比分裂应用场合更多。下面展示了合并方式的一般流程，如图 4.5 是合并方式与分裂方式的比较。

**算法:**基于合并方式的层次聚类算法

1. 使数据集中的每个点作为一个簇
2. 比较数据集中所有成对点的距离
3. 寻找距离最近的两个簇
4. 计算新合并得到的簇到其他簇的距离
5. 重复此步骤一直到合并到只剩下一个簇为止

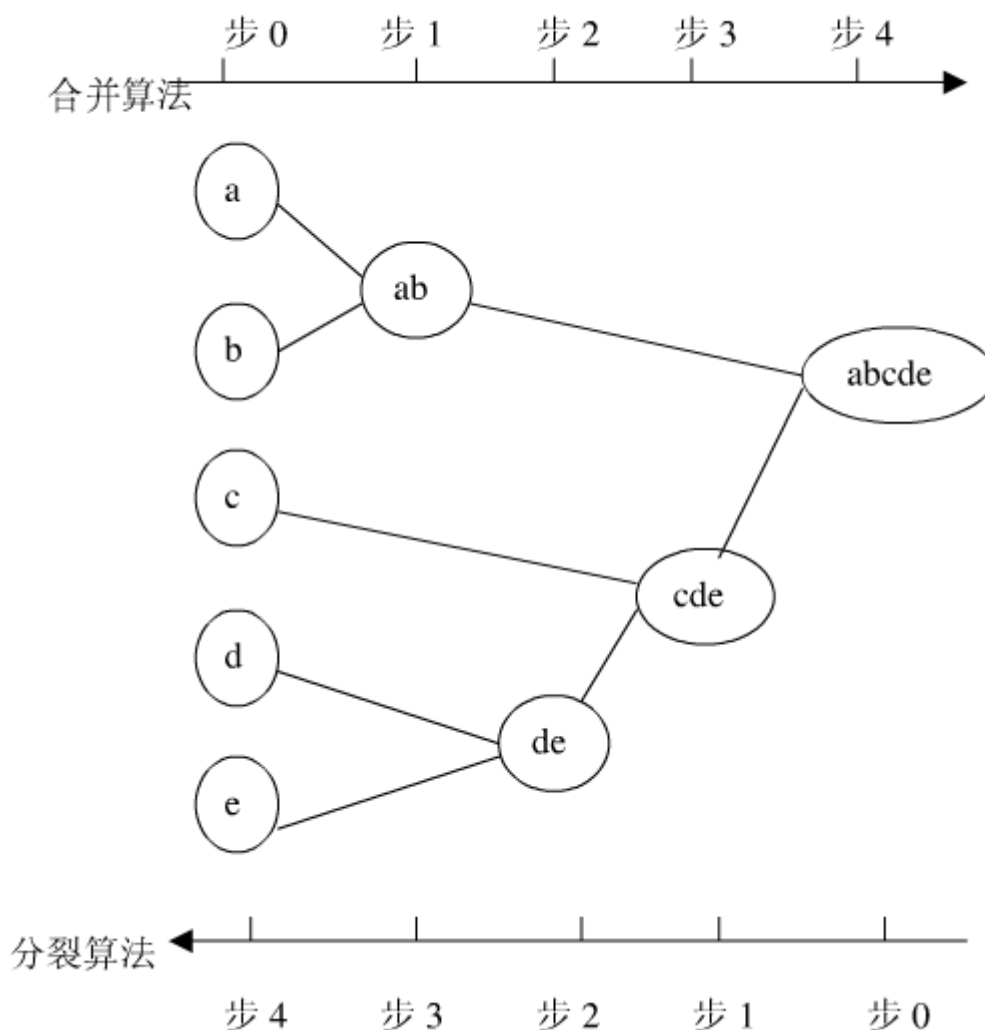


图 4.5 合并与分裂聚类

(3) 基于密度的聚类算法：这种聚类算法的主要思路也比较清晰：在一个特定的区域中，如果某个点的密度比预先设定的阈值数要大，那么就认为它们属于同一类并把它归纳到与之相近的聚类中去。

(4) 基于网格的聚类方法主要设计思路就是使用网格中的单元来保存数据，从而实现不同分辨率的聚类方式。

(5) 基于模型聚类的设计基本理念是：给各个簇假定一种特定的模型，然后寻找符合该对应模型的对象，不断重复一直到所有的数据都被归类到与其最相似的模型中。

为了利用先前的分析结果，从而实现对每一次聚类结果的动态更新，本文所论述的网络舆情信息系统采用了基于时间升序的 T-Single Pass 增量聚类算法。T-Single Pass 是一种数据流式的算法，这种方法对文档集合中所有的文档按照时间升序进行排列，系统读取第一篇文档所对应的特征向量进内存，以此作为第一个簇的簇中心，接下来依次将剩余文档所对应的特征向量按时间升序读取进内存，并与第一篇文档的特征向量进行余弦相似

度计算，如果计算得到的余弦相似度的值大于或者等于我们预先设置的阈值（THRESHOLD），那么就认为该新闻文本与其他新闻文本具有很高的相似性，然后把该文本与第一篇文本归为同一个类并重新计算中心；如果计算的余弦相似度的值小于在聚类之前设置的阈值（THRESHOLD），则继续与其他已有簇的簇中心依次进行余弦相似度计算，若计算得到的余弦相似度都小于阈值要求，则说明该文档与已有的各个类均不属于同一类，所以系统新建一个类并将该文档直接作为新类的聚类基准，剩余文档均依次遵循此步骤直到聚类过程结束。

下面将 T-Single Pass 增量聚类算法与被大众了解最多的静态聚类算法 K-means 做一些简单的比较：K-means 是一种基于原型来产生对数据单层次划分的聚类方式，该算法使用“质心”作为数据原型的定义标准，所谓“质心”，实际上就是该组里面所有数据点的平均值。该算法主要设计理念如下。

<b>Algorithm:</b> Primary K-Means Clustering
Step 1: select $m$ points as original centroid of this group
Step 2: <b>repeat</b> following steps
Step 3:        assign every point to its nearest point, and form $M$ cluster
Step 4:        recalculate centroids of every existing cluster
Step 5: <b>until</b> its centroid won't change anymore

K-means 算法的第一步就是随机选取  $m$  个点当作最开始的质心，然后将数据集中的所有点来和这  $m$  个初始质心分别进行距离计算，在欧式空间里面，这里的点之间常常采用欧氏距离作为衡量基准，而对于文本结构的内容，由于其较多属性维的存在，因此常常采用 Cosine Similarity 作为准则。第二步是在前一个步骤所计算的结果中将距离最近的点归并为同一个簇，并再次对该簇的数据点进行平均值计算，得到校准后的中心。然后不断循环此步骤，一直到所有的点都不再发生变化为止，此时聚类计算的结果收敛。尽管 K-means 设计想法很简单易懂，并且用途广泛，可以应用于大多数不同的数据类型的聚类中。只是它也有很多不好的地方：对于非球状簇的数据类型，或者对于不同密度簇的数据类型，它并不能提供很好地支持。而且 K-means 只能用于具有数据中心这一概念的数据模型当中。除此之外，由于它是在每次聚类结束之后才重新更新簇的质心，所以这种聚类方式并没有利用每一步的聚类结果。

T-Single Pass 增量聚类的算法思路清晰，易于理解，算法时间复杂度是  $O(nc)$ ，其中  $c$  为类的数量，远远低于传统聚类算法，因为算法是在每次余弦相似度比较完成之后就动态地更新已有的簇中心，而不是与 K-means 一样等到一次聚类迭代完成以后才更新簇中心。传统的增量聚类比如 Single Pass 算法有比较显著的缺点：首先，Single Pass 算法对文本对应的特征向量读取进内存的顺序有很高的要求，对于同一篇新闻文本如果按照不同的次序进行聚类，则很可能最终会出现不同的聚类结果。为了避免这一影响，基于 T-Single Pass 增量聚类算法在设计时就以时间升序为基准对获取的新闻文本进行排序，



这样一来就避免了不同聚类时序对实验结果造成的影响。增量聚类一般流程如图 4.6 所示。

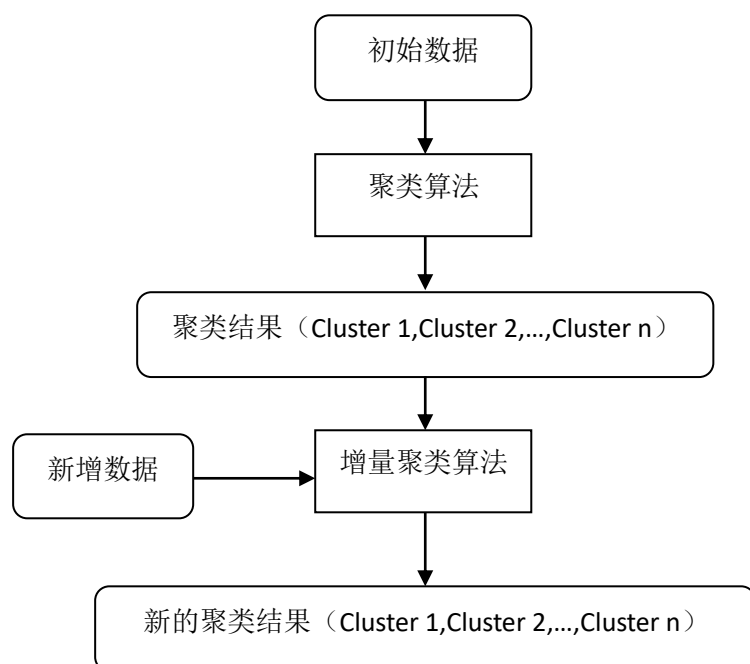


图 4.6 Procedure of Incremental Clustering

### 4.3 舆情预警

经过上一步骤的聚类过程之后，需要将聚类产生的结果展示出来。该模块为整套网络舆情信息系统的数据显示区域，可以查看聚类结果。将最终的信息以简洁清晰的图形为载体产生结果输出的方式被称作“数据可视化”，用户通过查阅图表，就可以很清晰地知道所要分析数据的特征以及每个数据与其对应的属性之间的关系。为了更明晰的结果显示，本论文所述的网络热点发现系统中采用了标签云和基于 d3.js 驱动的交互式图表设计，以及对每次分析得到的关键热度词都以标签云的形式展示出来，使得每次的分析结果一目了然。

在对最终的数据结果产生可视化的领域，对于高维数据的处理往往有以下几种技术：

第一，利用像素矩阵进行排列的可视化展示，相对应元素的属性值反映到图像上就是像素矩阵中像素的明亮程度和显示颜色，如图 4.7 所示。

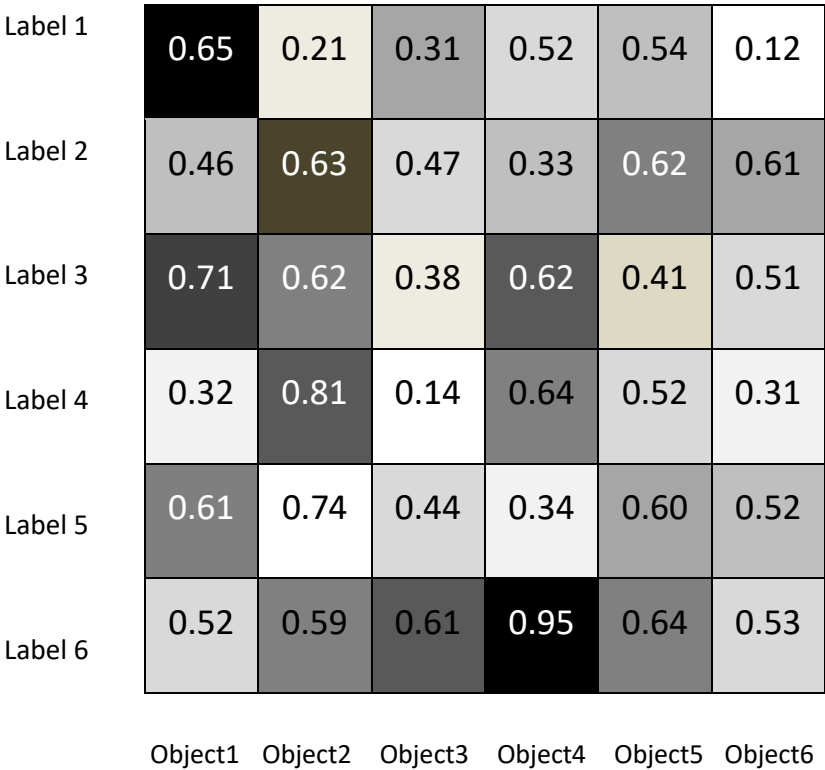


图 4.7 数据图形绝阵

第二，如果将每个属性看成是一条坐标轴，并且该坐标轴是互相平行的，而不是数学中两两垂直的几何概念。对象中每个属性值关联到对应坐标系上的点，将同一对象的每个点连接起来形成一条条线段，这样就得到了该数据对象的图形化展示。

第三，利用坐标轴标注属性，与平行坐标表示方法不同的是，这些坐标轴并不是互相平行的，而是选取某个中心点，然后依次将各个属性值以星形发散出去，就得到了数据对象的星形坐标展示。此外，还可以将数据对象的每个属性与类似于人脸形状的特征进行关联，数据可视化领域使用的是 Chernoff Face，例如对于一个人的数据可视化可以得到如下方案。

表 4.1 人的属性列表

数据属性	切尔诺夫脸谱图
身高	脸部的大小
体重	额头弧长
年龄	额头形状
肤色	颞的形状



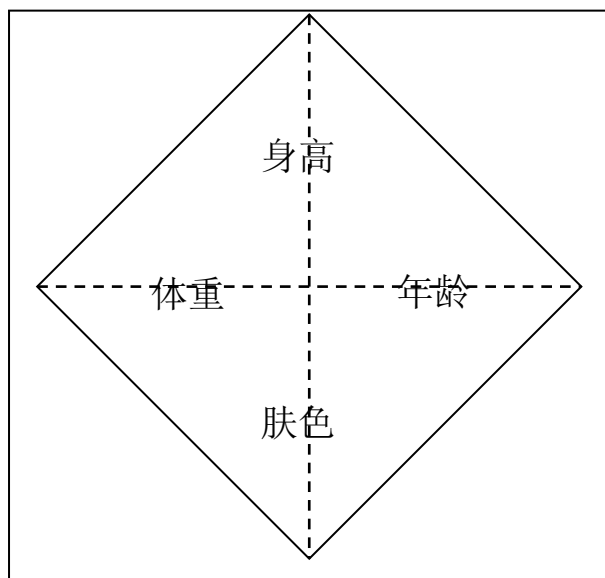


图 4.8 人类属性星形图

星形图映射属性规则如图 4.8 所示：为了便于分析，这里只选取 4 个属性作为观测值，图中以两条虚线为中心，其 4 个维的属性以中心向四周发散出去，就构成了其星形图表示方式。

对于数据可视化，通常有这样几个注意事项：图形表述就是要以最清晰直观的方式来传达尽可能多的信息，所以数据可视化应该尽可能的与先前已有的图形解释保持一致性，以及对各个不同变量之间的可理解性，图形化展示的界面是否合理等等。总体而言，在对数据分析的结果展示中，图形化无疑是最直观的表述方法。但是在实际操作中也应该考虑到，如果与其他表述方式（例如直接以文本展现）进行对比，是否有必要一定进行图形化的操作。

在本套网络舆情预警信息系统的设计中，由于从互联网获取的新闻文本数量比较多，所以对于最后的分析结果而言，如何能快速获取相关的网络热点话题无非是我们关注的重点。所以此处采用图形化和标签云的形式作为“舆情预警”功能模块的设计是很有必要的。但是考虑到可能会存在这方面的需求，所以本系统仍然保留了传统的文本形式聚类结果查看方式。

## 5 系统设计与实施

### 5.1 详细设计

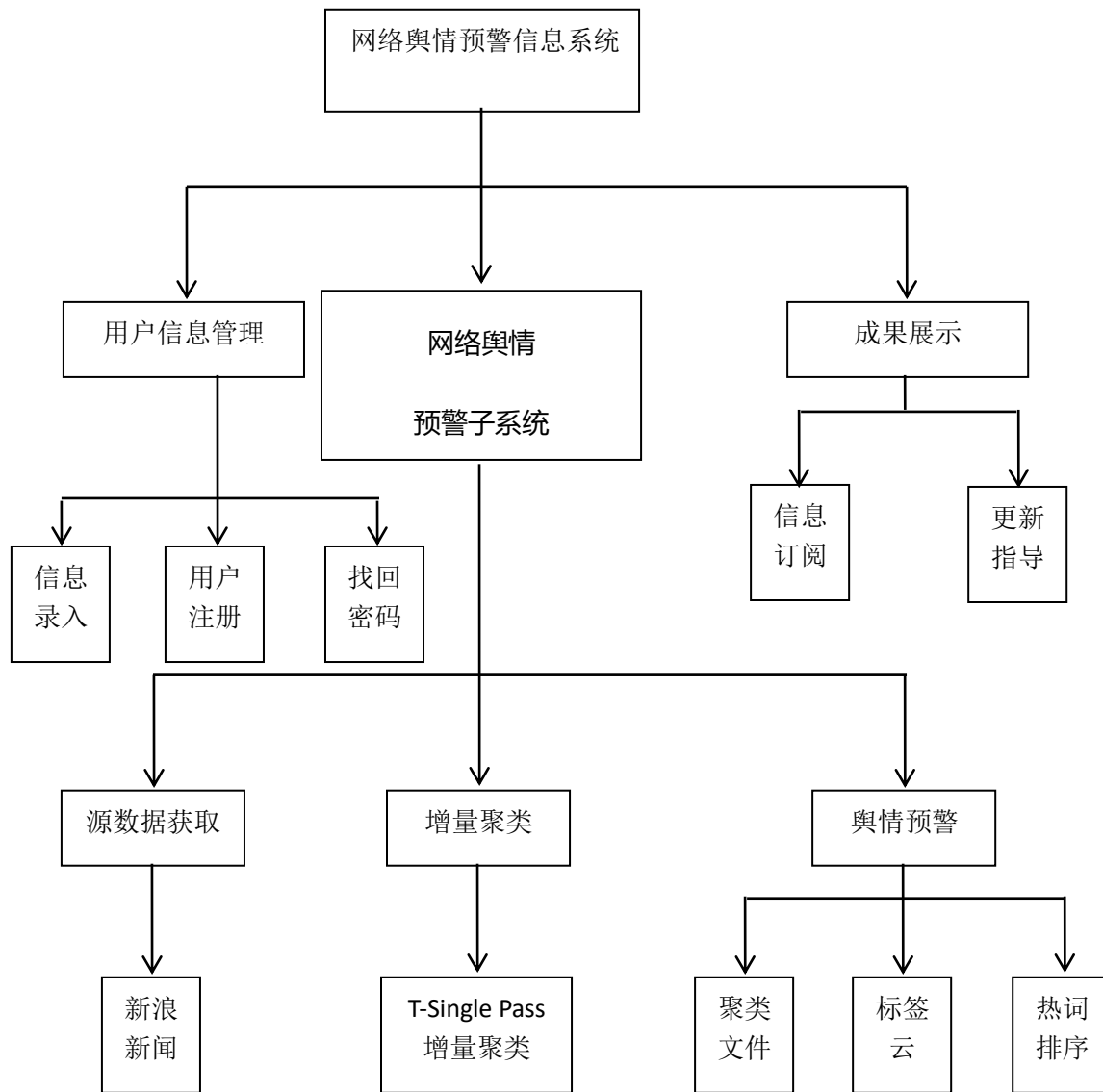


图 5.1 系统总体架构

系统总体架构如图 5.1 所示。

#### 5.1.1 数据获取模块

本网络舆情预警信息系统的获取模块是基于 JSoup 的网络爬虫，JSoup 是基于 Java 语言编写的网页 HTML 标签解析软件，可以直接通过调用其强大的 API 接口通过 DOM，CSS 以及类似于 jQuery 的操作方法来取出和操作数据、解析 URL 地址、HTML 内容。JSoup 的工作流程如下：

①从一个指定的 URL 中依次解析出 HTML 标签。

②使用文档对象模型来对 HTML 结构树来进行操作，使用 CSS 选择器来进行数据查找与属性抽取。

③依次解析 html 标签、属性和文本内容。

为了满足本文所论述中实验的需要，我们需要提取网页中提取有关标签来作为后续聚类步骤所需要的时间维。其中，本实验所利用的新浪新闻数据获取界面 HTML 结构如图 5.2 所示：

<!DOCTYPE html>
<html>
<head>
<title>新闻标题</title>
</head>
<body>
<h1 id="artibodyTitle" pid="1" tid="1" did="31856840" fid="1666"> 我们所需要的新闻标题 </h1>
<span class="time-source"> 2015 年 05 月 21 日 02:19 ----- </span>
<p> 所需要的新闻内容片段 1 ----- </p>
<p> 所需要的新闻内容片段 2 </p>
<p> 所需要的新闻内容片段 3 </p>
.....
</body>
</html>

图 5.2 新闻 HTML 页面结构图

至此互联网新闻文本获取阶段结束。

### 5.1.2 文本聚类模块

该模块采用基于 T-Single Pass 的增量聚类算法，它首先根据各新闻文本的发表时间进行排序，然后依次以数据流的方式读取进内存空间，进行聚类。此处需要我们人为的设定一个阈值（THRESHOLD），当后面到来的文本特征向量依次与各个类中心进行余弦相似度计算，当最大余弦相似度大于该阈值时，表明此条新闻属于该类，并动态的更新此聚类中

心，从而利用了先前的聚类结果。当最大余弦相似度小于该阈值，则认为此条新闻不属于该类，并依次和其他的簇进行计算，重复上述步骤。若  $\text{MAX}(\text{Cosine Similarity})$  全部都小于该阈值，则新建一个类，并将该新闻归并入该类中心。

聚类功能模块如图 5.3 所示。

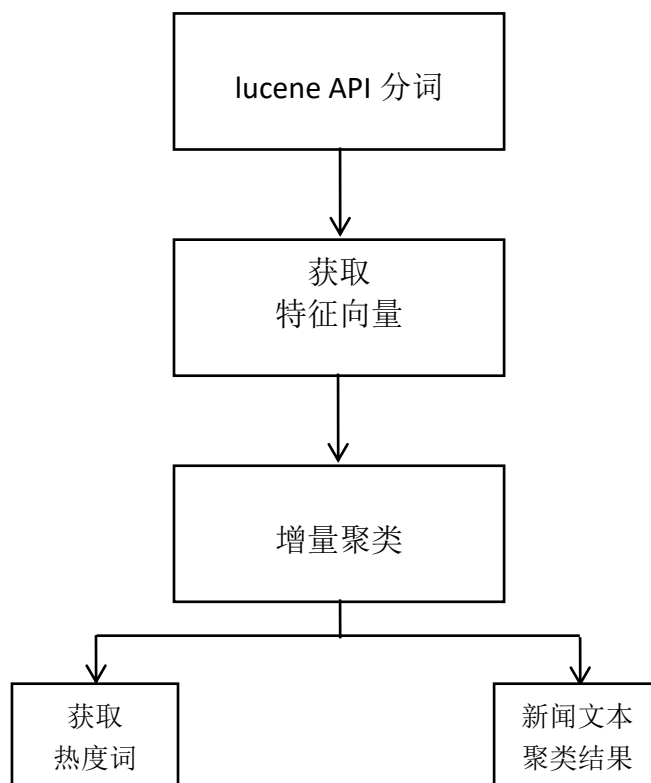


图 5.3 聚类模块功能结构图

### 5.1.3 舆情预警模块

聚类结果完成后，接下来的舆情预警模块就是对计算结果进行展示，产生预警。在本文所论述的网络舆情预警信息系统的设计中使用的是基于 B/S 架构的设计框架。其中 Web 前端采用开源的 bootstrap 的 JavaScript 开发框架，数据可视化部分采用基于标签云(Tag Cloud) + 基于 d3.js 的 JavaScript 交互式图表技术来对聚类得到的数据结果进行较高美观度的展示。bootstrap 是 Twitter 推出的基于 HTML5 的开源 Web 前端开发框架，能够很好地被各大主流浏览器所支持，bootstrap 的开发门槛极低，且有着很好地维护，通过阅读官方文档说明了解各个 API 的功能及使用方法，并根据需要使用就能够很容易的创建网页端的响应式 UI，并且和移动设备也做到了比较好的兼容。

D3 (Data-Driven Documents) 即数据驱动文档，是一个专门用于创建数据可视化图形的 JavaScript 库，使用 D3 的步骤也比较简单，仅需要以下步骤即可创建：①浏览器将数据加载到内存空间；②把相关的数据以某种机制附加到特定文档格式中的属性元素当中去；③对每一个元素的数据维度进行解析，与此同时为该元素创建相应的数据可视化属性，

从而实现数据元素的正常转换；④响应用户在前端的输入动作并实现元素属性的状态过渡。

其中，舆情预警的设计如图 5.4 所示。

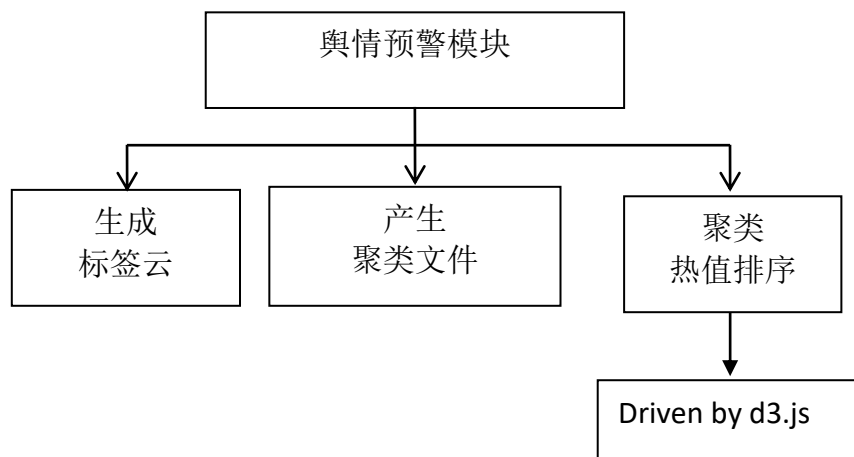


图 5.4 舆情预警功能模块图

## 5.2 系统运行

### 5.2.1 系统简介

遵循网络舆情预警系统功能模块的设计规范，本套系统在功能上也主要划分为爬虫、聚类、预警这三个方面。但是考虑到对于网络热点舆情发现方面的使用者角色而言，对于用户的不同背景可能对系统的功能偏好会有些许差别，例如用户身份为高校教师、政府人员、研发人员等不同职业背景时，他们的要求都不尽相同，虽然系统原则上在功能的使用方面对所有用户开放，但是在每次用户信息录入的同时后台会记录使用者身份，以便更好的改进后续功能，所以希望所有使用者都能辅助我们完成该步骤的实施。

为了保证舆情预警数据可视化部分的易识别性与美观度，本系统采用基于 B/S 模式（即在用户浏览器端仅作为数据返回的显示层，在服务器后台调用相应的业务逻辑层和算法）搭建的舆情信息系统框架，前端部分采用 bootstrap 框架技术作为表示层，使用 Java 语言书写其中的算法部分，本系统用来保存使用者身份信息数据库采用的是 MySQL。Web 前端采用 HBuilder 作为开发工具书写网页表示层，后台使用 Eclipse EE 编写业务逻辑模块。实现了如下几项功能：调用相关的接口实现网络爬虫爬取互联网新闻、基于 T-Single Pass 的增量聚类、聚类结果的图形化展示、输入邮箱订阅最新研究进展、系统更新说明、用户信息录入等。

配备好相应的运行环境之后，图 5.5 显示了系统运行的主页 index。

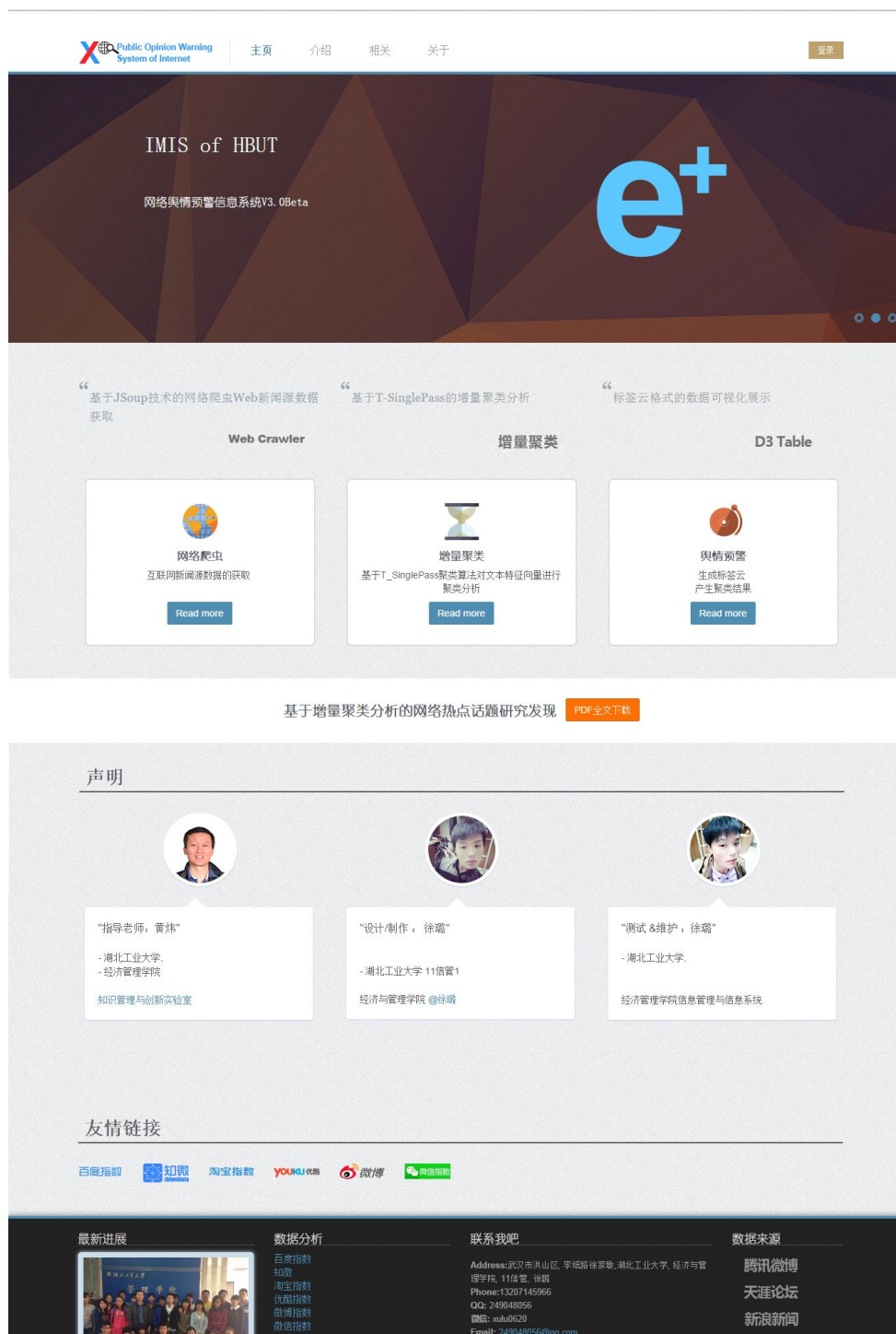


图 5.5 系统运行主界面

## 5.2.2 运行环境

为了能使本网络舆情预警信息系统能够正常运行，还需要一定的硬件与软件环境的支持，具体如表 5.1 所示。

表 5.1 系统运行环境

硬件环境	CPU	Intel Core i3 Dual 2.1GHz			
	RAM	4GB			
	ROM	640GB			
软件环境	OS	Windows 7 Ultimate 64bit			
	Run Environment	Apache Tomcat Server	JDK 1.8	MySQL	Web Browser (Chrome、Safari、Firefox Recommended)
开发环境	IDE	Eclipse EE 后台开发		HBuilder Web 前端设计	
其他要求	由于爬虫部分需要从网上获取内容，标签云的动态匹配也需要网络服务， 因此本系统的使用需要连接网络 如果条件不允许，也可以直接使用系统内自带的样本数据				

### 5.2.3 运行演示

导航栏“相关”折叠面板展开之后是本系统的核心功能区，如图 5.6 所示。



图 5.6 主功能面板



## （1）网络爬虫

根据需要选择相应的网站点击即可运行，该步骤对用户透明，但是可以在控制台看到程序运行，本次系统运行采用的是新浪新闻的数据，新浪新闻是目前国内最大的新闻类门户网站之一，因此从这里爬取的数据具有一定的说服力，如图 5.7 所示。

```
Tomcat v8.0 Server at localhost [Apache Tomcat] D:\Program Files\Java\bin\javaw.exe (2015年5月20日 上午9:09:24)
This is 8th article
201505201233
URL is : http://news.sina.com.cn/c/2015-05-20/121731854757.shtml Title is: 北京昌平幼儿园女教师虐待儿童被行拘
This is 9th article
201505201217
URL is : http://news.sina.com.cn/c/2015-05-20/120331854750.shtml Title is: 陕西榆林安监局官员上班玩游戏 自曝中午饮酒
This is 10th article
201505201203
URL is : http://slide.news.sina.com.cn/weather/slide_1_29155_84520.html#p=1 Title is: 高清图: 广西福建多地暴雨成灾引发洪涝
This is 11th article
URL is : http://news.sina.com.cn/c/2015-05-20/115531854738.shtml Title is: 李克强会见巴西众议长库尼亚

This is 12th article
201505201155
URL is : http://news.sina.com.cn/c/2015-05-20/114831854731.shtml Title is: 李克强会见巴西参议长卡列罗斯
This is 13th article
201505201148
URL is : http://news.sina.com.cn/c/2015-05-20/112731854739.shtml Title is: 安徽地震活跃态势为40余年最高水平
This is 14th article
201505201127
URL is : http://news.sina.com.cn/c/2015-05-20/112631854725.shtml Title is: 广西柳州76个乡镇遭暴雨袭击 部分路口交通瘫痪
201505201126
This is 15th article
URL is : http://news.sina.com.cn/c/2015-05-20/112631854664.shtml Title is: 南宁交警1个月查处30万起电动车违法行为
```

图 5.7 爬虫程序运行 Console

## （2）T-Single Pass 增量聚类

点击 T-Single Pass 增量聚类模块即可开始运行聚类算法，该过程也对用户透明，在控制台可以看到执行结果，如图 5.8 所示。

```
Tomcat v8.0 Server at localhost [Apache Tomcat] D:\Program Files\Java\bin\javaw.exe (2015年5月20日 上午9:09:24)
0.062494136 0.009760168 0.013887586 0.010868879 0.022764262 0.0023152002 0.0058999504 0.011563145 0.0037017674 0.00424
0.039470185 0.004669569 0.00707679 7.3225563E-4 0.001920147 0.01158541 0.04046077 0.004638788 0.0019792973 0.02601367
7.013262E-4 0.003562038 0.0040616733 0.008123347 0.0010656693 0.008123347 0.01218502 0.0023726614 0.0020308367 0.00406
0.0021908558 0.0012887353 0.006767218 5.273641E-4 0.0050754133 0.0028865822 0.014358834 0.010049931 0.0011867069 0.001
0.01600554 0.010249151 6.268152E-4 0.009366164 0.008597453 0.0016550984 0.047895562 0.026979728 0.08460647 0.010131340
0.0055743144 0.0046648583 0.011903545 0.02217968 0.008265721 0.01296754 7.7591033E-4 0.001896115 0.005042317 0.0147864
0.021926636 0.001733239 0.015151516 0.0014720076 0.0019456539 9.153196E-4 0.029849043 0.04854727 0.011205877 0.0050682
0.002224882 0.0023084437 6.088602E-4 0.0023084437 0.0100760115 0.001187747 8.327468E-4 4.7058222E-4 0.0024049089 0.012
0.010665306 0.028921096 0.0075204046 0.012405364 0.005524834 0.011387495 0.0034801157 0.012405364 0.009260462 0.008629
0.0086197825 0.012594814 0.014650711 0.009053943 0.0028222671 0.023583824 0.009168947 0.0013277164 0.007723986 0.00174
0.028541844 0.0034493795 0.1098724 0.028541844 0.02724909 0.013451571 0.003144812 0.01166962 0.0065608597 0.36790764
0.0014071554 0.0025763488 0.003004491 0.002418707 0.0026349023 0.0011087661 0.00379145 7.882944E-4 8.775497E-4 0.00241
0.039470185 0.004669569 0.00707679 7.3225563E-4 0.001920147 0.01158541 0.04046077 0.004638788 0.0019792973 0.02601367
7.013262E-4 0.003562038 0.0040616733 0.008123347 0.0010656693 0.008123347 0.01218502 0.0023726614 0.0020308367 0.00406
0.0021908558 0.0012887353 0.006767218 5.273641E-4 0.0050754133 0.0028865822 0.014358834 0.010049931 0.0011867069 0.001
0.01600554 0.010249151 6.268152E-4 0.009366164 0.008597453 0.0016550984 0.047895562 0.026979728 0.08460647 0.010131340
0.0055743144 0.0046648583 0.011903545 0.02217968 0.008265721 0.01296754 7.7591033E-4 0.001896115 0.005042317 0.0147864
0.021926636 0.001733239 0.015151516 0.0014720076 0.0019456539 9.153196E-4 0.029849043 0.04854727 0.011205877 0.0050682
0.002224882 0.0023084437 6.088602E-4 0.0023084437 0.0100760115 0.001187747 8.327468E-4 4.7058222E-4 0.0024049089 0.012
0.010665306 0.028921096 0.0075204046 0.012405364 0.005524834 0.011387495 0.0034801157 0.012405364 0.009260462 0.008629
0.0086197825 0.012594814 0.014650711 0.009053943 0.0028222671 0.023583824 0.009168947 0.0013277164 0.007723986 0.00174
0.028541844 0.0034493795 0.1098724 0.028541844 0.02724909 0.013451571 0.003144812 0.01166962 0.0065608597 0.36790764
0.0014071554 0.0025763488 0.003004491 0.002418707 0.0026349023 0.0011087661 0.00379145 7.882944E-4 8.775497E-4 0.00241
```

图 5.8 增量聚类运行 Console



### （3）数据可视化

聚类执行完毕之后，算法会根据聚类结果自动提取相关的热度词并随机抽取 20 个关键词（本实验中产生的热度词为：留言 青春 英烈 团中央 共青团中央 遗志 缅怀 铭记 弘扬 书记处 先烈 中华魂 复兴 继承 革命 第一书记 继往开来 爱国精神 舍生取义 中央网络 精神 负责同志 全国各地 奋斗 中华民族）生成标签云的可视化页面，点击相关热度词之后即可跳转到百度搜索引擎对该热词的搜索结果界面。与此同时，在指定目录下，会生成文本聚类结果，如图 5.9、5.10 所示。

名称	修改日期	类型	大小
Category0	2015/5/20 19:30	文件夹	
Category1	2015/5/20 19:30	文件夹	
Category2	2015/5/20 19:30	文件夹	
Category3	2015/5/20 19:30	文件夹	
Category4	2015/5/20 19:30	文件夹	

图 5.9 聚类结果文件



图 5.10 生成标签云

#### （4）系统更新介绍

点击导航栏的“介绍”选项卡，可查看网络舆情预警信息系统的更新摘要, 如图 5.11 所示。



图 5.11 更新摘要说明

#### （5）订阅最新

如图 5.12 为相应的页面，输入邮箱即可完成订阅。



图 5.12 订阅最新

#### （6）用户信息管理

该网络舆情预警信息系统的设计原则上对所有用户开放使用权限，但是为了能对下一个版本有针对性的更新，因此需要获取使用者的少许信息。点击主页的“登录”按钮，即可登录本系统，当每一个用户在前台登录的时候，后台会生成记录使用日志如图 5.13 和图 5.14 所示。设计模块如图 5.15 所示。


名称	修改日期	类型	大小
 2015-05-21-05-58-55.log	2015/5/21 17:58	文本文档	1 KB

图 5.13 Login 日志

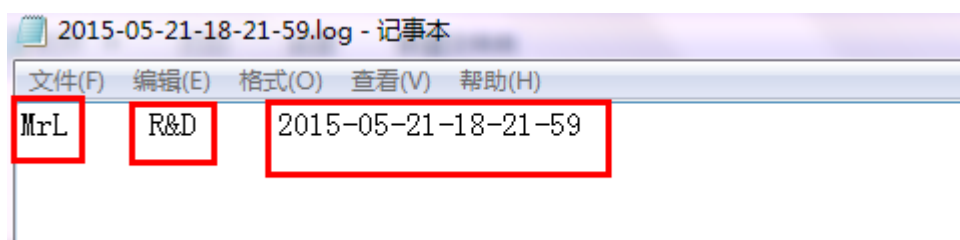


图 5.14 身份日志记录信息

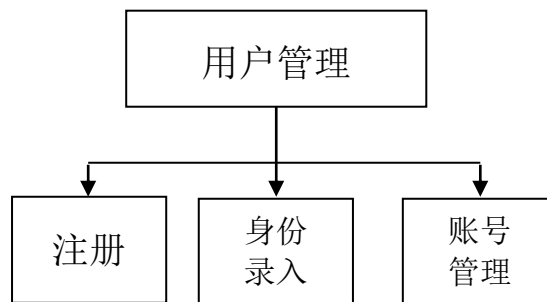


图 5.15 使用者信息管理

#### （7）用户信息录入

如果没有注册该系统，可以点击登陆页面的“点此注册”按钮，跳转到相应页面后，仅需输入用户名、邮箱、密码，以及使用者的身份即可注册，如图 5.16 和图 5.17 所示。

登录

邮箱

密码

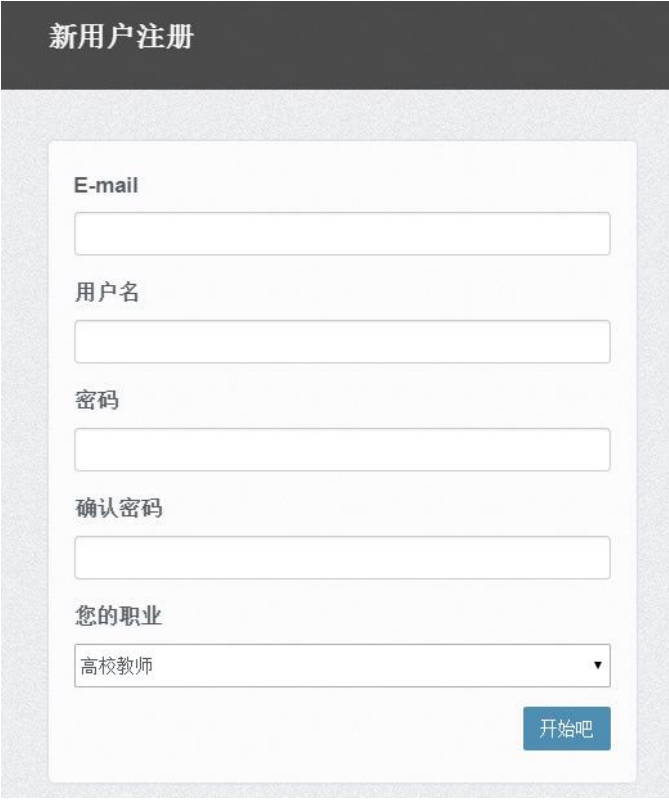
☐ 记住密码

[忘记密码?](#)

[登录](#)

还没有账号? [点此注册](#)

图 5.16 用户信息录入



The image shows a web form titled "新用户注册" (New User Registration) in a dark header. The form itself is a light gray box with a white border. It contains five input fields: "E-mail", "用户名" (Username), "密码" (Password), "确认密码" (Confirm Password), and "您的职业" (Your Profession). The "您的职业" field is a dropdown menu with "高校教师" (University Teacher) selected. A blue button labeled "开始吧" (Get Started) is at the bottom right of the form.

新用户注册	
E-mail	<input type="text"/>
用户名	<input type="text"/>
密码	<input type="password"/>
确认密码	<input type="password"/>
您的职业	<input type="text" value="高校教师"/>
<input type="button" value="开始吧"/>	

图 5.17 新用户身份注册

基于本文所设计的系统在实验过程中采用的是新浪新闻爬取的数据，其总的实验步骤如下图 5.18 所示。

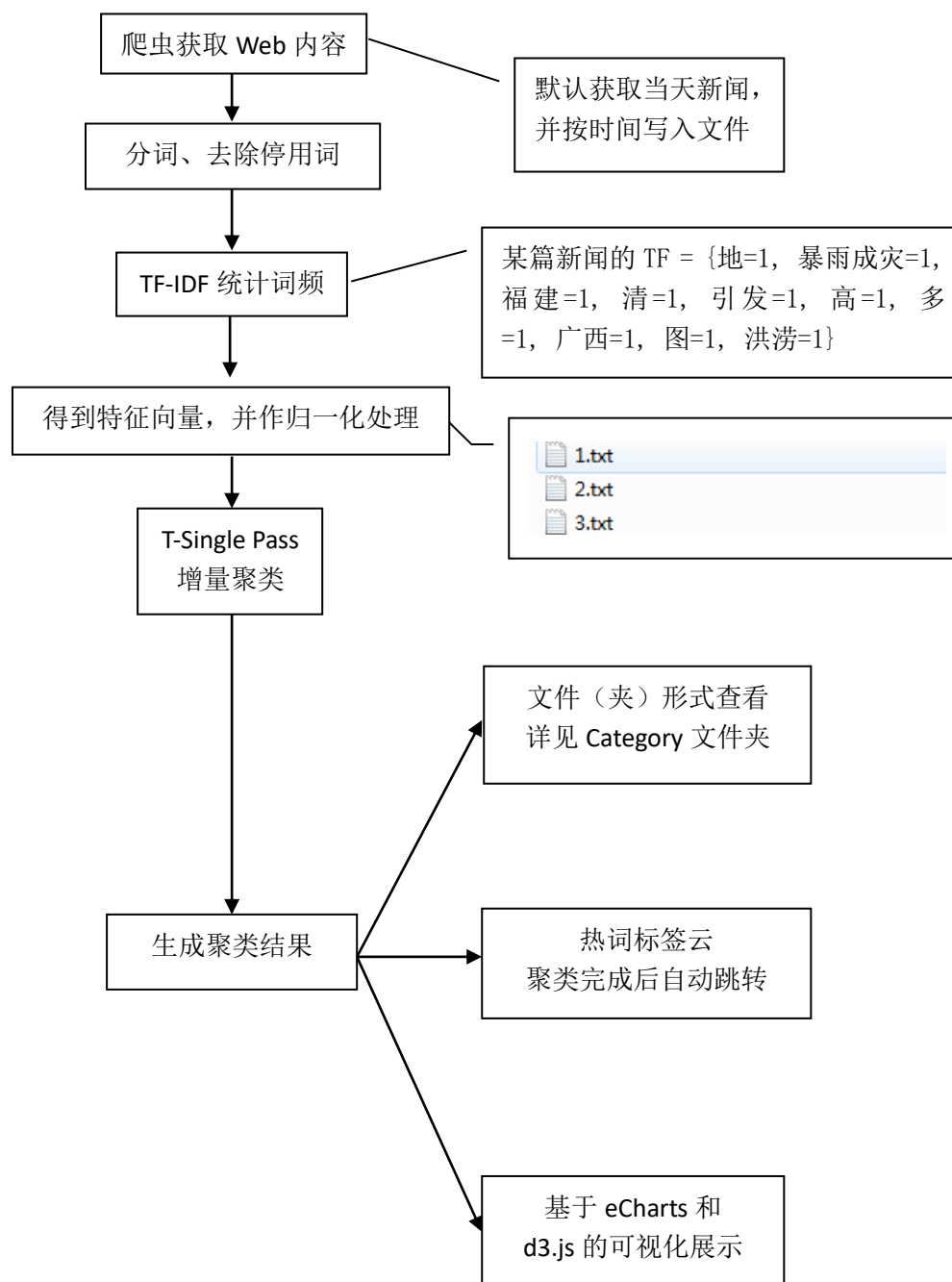


图 5.18 运行流程

### 5.3 实验结论

通过对新浪新闻进行爬取，对数据的增量聚类测试结果显示，相比于以往的非增量聚类算法，基于时间序列分析的 T-Single Pass 算法能够很好地利用先前的聚类结果、在对数据流式的文本聚类中具有明显的优势。其中，基于 d3.js 的数据可视化图表创建如图 5.18 所示。如图 5.19 和 5.20 展示了基于 eCharts 的可视化处理。

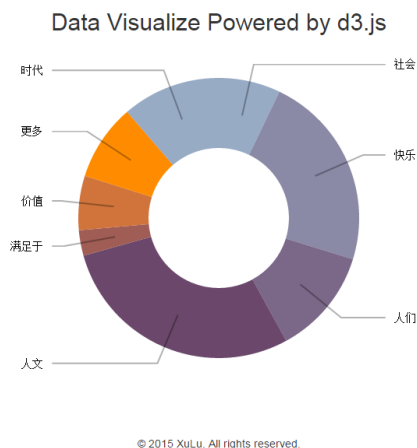


图 5.18 基于 d3. js 的图表展示

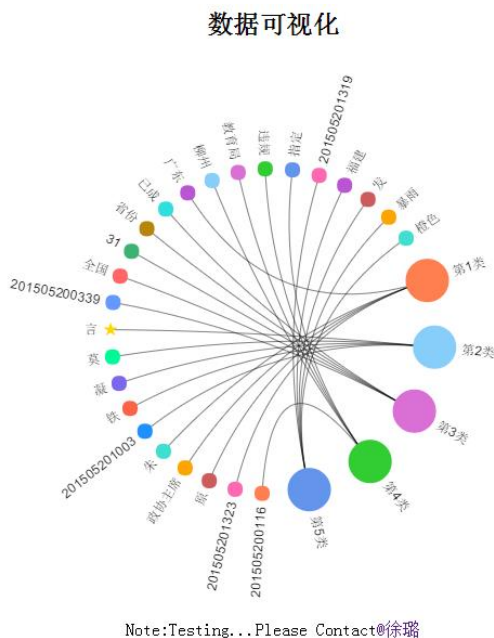


图 5.19 基于 eCharts 的聚类可视化

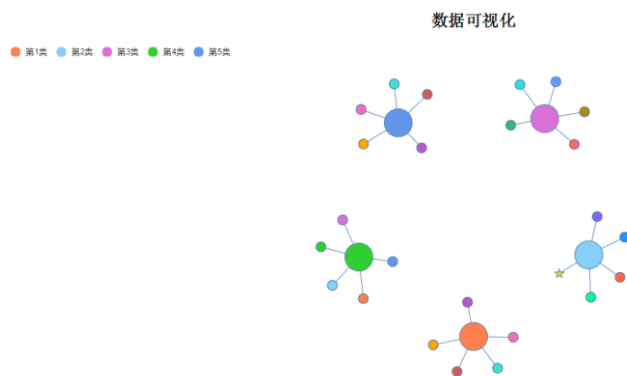


图 5.20 聚类结果图

## 6 总结与展望

### 6.1 全文总结

通过这一段时间毕业设计的制作与毕业论文的撰写中，一共涉及了 Web 前端设计、网络爬虫、基于 T-Single Pass 的增量聚类、以及数据可视化、JSP 动态网页等技术，使我对这些技术都有了比先前更深的理解。通过将该算法与经典的 K-means Clustering 进行对比后发现，对于结果而言 T-Single Pass 具有明显的优势。互联网时代，数据已经成为各大信息技术公司纷纷趋向的战略制高点，并且已经开始逐渐成为除人力、资金外的新型无形资产，我们正在由 IT（Information Technology）时代迈向 DT（Data Technology）时代。因此如何从众多数据中挖掘出具有代表性的信息就成了时下研究的热门领域，在这样一个新时期下的网络舆情相关的研究必然会发生一些新的改变，据相关报道分析，新浪微博部署的分布式神经网络分析系统将每个用户都归为整套网络中的神经元，这样对于新浪微博的舆情控制体系非常有利。然而，由于时间和个人水平等各方面的影响因素，本系统依然存在着一些不足之处，主要体现在系统源代码较繁杂，所以后期维护方面比较繁琐，核心算法也并没有调校到最优状态，关于阈值（THRESHOLD）的选择也没有一个确定的标准，总体而言现阶段只是实现了相应的功能，对于优化方面还有很大的提升环节。

本网络舆情预警信息系统实现了 Web 数据爬取、数据预处理、基于 T-Single Pass 的增量聚类以及数据可视化等功能，考虑到界面美观度及易用性和可移植性等因素，采用了基于 B/S 的架构设计，前端使用基于 bootstrap 的 Web 前端开发框架的用户界面，以及基于 d3.js、eCharts 驱动的数据图形化展示，保证了系统整体界面的简洁和美观性，后台使用 Java 语言编写的业务逻辑，易于部署到不同的平台和新功能模块的扩展。

### 6.2 研究展望

我国有关网络舆情方面在学术界已有一些历史，并且也收获频出，以新浪微博为代表的新型媒体平台也都纷纷建立了属于自己的舆情预警机制，时下舆情主要以门户网站、新闻客户端、微博、微信为平台产生与传播。随着移动互联网时代的到来，网络舆情事件的爆发更倾向于出现在微博、新闻媒体客户端等新型平台，因此对于以微博为代表的短文本数据挖掘，以关注、转发、评论、点赞、收藏为规则的分析，从中挖掘出用户的情感偏好，对于每一次热点事件的传播进行一系列的网络传播拓扑结构分析，由电脑程序控制的僵尸粉或者人工操作的水军活跃数据。对于热点话题的发现、网络舆情等相关研究领域，要想与时代接轨，务必要将相关研究方向与社会化网络的挖掘与分析进行有机结合起来。此外，截止到本文完成之时，国内最大的社交媒体平台——新浪微博 iOS 和 Android 客户端已经上线了图片评论功能，以图片为载体的传播方式可能会绕过新浪相关的程序监测机制，除此之外，尽管基于微博短文本的数据具有很高的分析价值，但是微博数据由于其本身的特点，所以产生的特征稀疏问题会对聚类结果产生非常大的影响。而对于涉及某一类特定主



题的热点事件相关微博内容本身而言，用户的每一次互动参与，包括隐式部分（如阅读相关微博的次数、在特定话题或者微博下的停留时间）和显示部分（收藏、点赞、转发、评论、艾特其他用户、搜索行为等等），可以针对这些来对算法进行设计，从而实现基于用户情感分析的增量聚类。这些问题都使得网络舆情热点话题相关的研究依然还有很大的提升空间。

## 参考文献

- [1] 潘敏, 王明文. 基于簇特征的文本增量问题研究[J]. 江西师范大学学报, 2014.
- [2] 周刚, 邹鸿程. MB-SinglePass: 基于组合相似度的微博话题检测[J]. 计算机科学, 2012, (10).
- [3] 吴绍忠, 李淑华. 互联网络舆情预警机制研究[J]. 中国人民公安大学学报, 2008.
- [4] 刘玉新. Web2.0 互联网在线话题发现和热度评估[D]. 华南理工大学硕士学位论文, 2013, (6).
- [5] 孟海东, 王淑玲, 郝永宽. 动态增量聚类的设计与实现[J]. 计算机工程与应用, 2009, (2).
- [6] 朱恒民, 朱卫未. 基于 Single-Pass 的网络话题在线聚类研究[J]. 现代图书情报技术, 2011(12): 52-57
- [7] 许峰. 基于 Web 的实验室互联网舆情分析处理系统的研究与实现 [J]. 科技情报开发与经济, 2011.
- [8] 陶舒怡. 基于簇相合性的文本增量聚类算法研究[D]. 江西: 江西师范大学, 2013, (5).
- [9] 张东晋. 基于单事件新闻多文档聚类及自动文摘的设计与实现[J]. 厦门大学, 2014
- [10] 王丹, 张兆心, 宋颖慧. 基于高权重词集的增量聚类算法研究[J]. 微计算机信息, 2011.
- [11] 王伟, 许鑫. 基于聚类的网络舆情热点发现及分析[J]. 情报分析与研究, 2009.
- [12] 于翔. 基于网格的数据流聚类方法研究[D]. 黑龙江: 哈尔滨工程大学, 2010, (4).
- [13] 张小明, 李舟军, 巢文涵. 基于增量聚类的自动话题研究[J]. 软件学报, 2009.
- [14] 杨震, 段立娟, 赖英旭. 基于字符串相似性的网络短文本舆情热点发现技术[J]. 北京工业大学学报, 2010, (5).
- [15] 黄晓斌, 赵超. 文本挖掘在网络舆情信息分析中的应用[J]. 情报科学, 2009, (1).
- [16] 戴维民, 刘轶. 我国网络舆情研究现状及对策思考[J]. 图书情报工作, 2014, (1).
- [17] Agostino Forestiero. A Single Pass Algorithm For Clustering Evolving DataStream Based On Swarm Intelligence[J]. Data Min Knowl Disc, 2011(10)
- [18] 税仪冬, 瞿有利, 黄厚宽. 浅周期分类和 Single Pass [J]. 北京交通大学学报, 2009, (10).
- [19] 李桃迎, 陈燕, 秦胜君, 李楠. 增量聚类算法综述[J]. 科学技术与工程, 2010, (12).
- [20] 高燕飞, 陈俊杰, 乔冰琴. 增量聚类算法的研究与设计[J]. 太原科技大学学报, 2012, (8)
- [21] Issei Sato. Deterministic Single-Pass Algorithm for LDA [J]. University of Tokyo, Japan 2010.
- [22] Chengyue. Natural Language Processing in Web Data Mining [J]. School of Computer Science and Technology, 2010.

## 致 谢

大学生涯即将走到终点，毕业论文与毕业设计是对这四年大学生活的一个完美收梢。在指导老师的热心帮助和自己的努力下，终于圆满完成了论文终稿与系统设计。

在这里首先要感谢我的指导老师黄炜老师，黄炜老师从任教以来一直秉承严谨治学的作风，并且在网络舆情热点话题发现研究领域有着非常丰富的经验，其为人方面也备受学生尊敬。在毕业设计的制作与毕业论文的撰写期间，我十分感谢黄炜老师所提供的各种帮助，不管是专业方面还是待人接物方面我都有了很大的收获。刚开始接触 Web 数据挖掘方面，自己确实感觉到难以接受，不仅需要足够熟练的编程能力，也对数学能力有很高的要求。遇到弄不明白的地方时，在黄老师的指导下、在和学长的沟通过程中、在自己这接近 5 个月的不断学习不断提升中，使得自己不管是在系统的设计与实现方面还是在理论方面都得到了很大提高，从技术层面来讲，Web 数据挖掘是个很奇妙的领域，它的目标是从一大堆无规则、有噪声的数据中挖掘出潜在的、对我们分析有用的信息，从零开始编写算法、优化算法、调整参数，一直到最后的系统设计，以及数据装载与系统测试阶段，都是一件非常值得钻研的事情，但是与此同时学习起来也是比较困难的，并不同于管理信息系统，算法方面则是一件非常枯燥乏味的事情，因为要涉及比较广的技术层面和理论深度，尤其是算法的编制与后期优化方面，需要参考很多书籍与论文资料，而且理解上对于笔者现阶段的水平而言也面临着不少挑战。但是通过黄炜老师的教导和自己的努力，最终还是完成了本系统的设计，再次向黄炜老师的指导和所提供的平台致谢！

其次也要感谢我的同学们，有时候碰到疑问，大家都会聚集在一起想办法解决，有个良好的氛围学习起来也比较愉快。系统设计完成之后，同学们帮我测试了 bug，对于系统中有些不合理的功能模块也都一一指出，并提出了一些很有参考价值的意见。

此外，在这期间也要感谢我的父母，虽然他们对于我所学习的专业并不是太了解，也没办法为我提供一些设计上的建议。在之前的那段时间里一直处于低谷状态，但正是因为有了他们的鼓励才使得我重拾信心，为自己的大学生活画上一个最圆满的句号。

最后，衷心感谢自己学生时代以来所有传授过我知识和做人道理的老师。师者，传道授业解惑也。作为学生，从小到大从老师这里接收到的知识最多，并且在我迷茫的时候都能够提出相关的建议助我走出困惑。与此同时，我也深知自己还有太多的地方需要不断学习、不断强化、不断提高。

每个人的人生道路都是个不断学习、不断成长的过程，对于生活中别人的帮助，理应怀有一颗感恩的心。借此毕业设计完成之际，再次向一路走来所有给予过帮助的老师、亲友、同学们表达我发自内心最真诚的谢意与最美好的祝愿！