北京交通大學

Beijing Jiaotong University

软件构思综合训练

影视舆情分析系统

方案建议书

组　　长：汤新宇 17301137

小组成员：王心蕊 17301048

陈嘉琪 17301060

唐　麒 17301138

张钰铎 17301145

贡献分配：同　等　贡　献

指导老师：李　宇

北京交通大学软件学院

2020 年 6 月 15 日

1 引言

本方案的制定，是基于对“影视舆情分析系统”现有问题的分析和市场上舆情分析系统存在的问题，以及对现代大数据、云计算等互联网技术发展的考虑。

方案的设计自始至终都遵循以下基本原则：

- 高起点，技术上先进可行且兼顾目前业务需求和今后较长时间（1~3 年）业务发展的可能；
- 系统安全可靠、可维护性、可扩展性好；
- 系统设计基于对原有系统的经验教训做出深入的总结的基础上

1.1 目的

本文档是“影视舆情分析系统”系统的系统方案建议书，在本文档中，我们通过对项目的背景（问 题）和范围进行重述，明确进行系统开发的必要性和可行性，同时对系统的方案提出建议，帮助客户进一步理解项目的关键问题和分析项目的可行性，帮助选择具体的项目方案。

1.2 背景

根据中国互联网络信息中心（CNNIC）第 44 次《中国互联网络发展状况统计报告》显示，截至 2019 年 6 月，中国网民规模达到 8.54 亿，互联网普及率为 61.2%，超过全球平均水平 7.6 个百分点。伴随着高歌猛进地互联网化，以微博、微信为代表的网络社区成为了新的最重要的舆论场。微信在 2019 年底共计 11.51 亿月活用户，公众号平台超过 1000 万个，微博在 2019 年底也实现了月活用户 5.16 亿人次，面对错综复杂的舆论阵地，舆情产业是信息服务行业在大数据时代的又一轮升级产业，通过研究用户及其相关舆情，将有价值的信息传递给用户，最终帮助用户解决实际问题。通过对目前市场上的舆情检测与分析系统的调查来看，现有的系统数据服务涵盖面广且主要面向政府事务和金融市场，而针对其他行业的舆情监测和分析，例如电商评论和游戏社区等舆论，相关系统相对较少。网络舆情是通过互联网传播的公众对现实生活中的某些热点、焦点问题所持有较强影响力、倾向性的言论和观点。影视舆情是社会舆情在影视市场的延伸。在互联网背景下,影视舆情特指影视受众在网络空间所表达的、针对一部或多部影视剧作品、人物、机构、产业、节展、相关政策法律法规等相关议题的意见、态度、倾向、情绪、行 为等信息的总和。对于影视作品

来说，舆论与口碑在很大程度上决定了影视作品的成 功与否。影视舆情信息在互联网上客观存在,形成了与影视剧内容产品相伴而生的衍生 信息产品,形成影视剧不同传播时期的先期舆情(影视剧首轮发行期前)、同期舆情(首轮发行期同步)、后期舆情(首轮播映后)、长尾舆情。先期舆情和同期舆情在一定程度上作用于影视剧作品的观影期待,影响观看意向、初步评价、后期评价、口碑评分等,进而影响收视率和票房。

因此，舆论导向的监测、分析与预警对于影视行业来说十分重要。并且，舆情分析预警系统的实现具有可行性，舆情导向的预测不是一种抽象的可能性，而是现实的可能性，这种现实的可能性并非凭空想象而是有其现实基础，是对舆情的历史信息和其 他因素经过判断、分析而得出的结论。舆情也同其他事物一样，是一种客观存在，有其产生、发展变化的规律。只要对其给予客观、全面、科学的考察，细致、认真、仔细的分析，我们是可以对舆情导向的有无、好坏、大小做出基本准确的评判和预测的。

1.3 项目简介

本系统旨爬取互联网上有关影视作品方面的舆论，经过分析处理后为目标客户提供检索服务、专业化的舆情分析报告、敏感信息监测与预警服务以及为重大决策提供信息参考。

项目包含以下几点功能：

(1)影视舆情热点监测

对某一影视舆情事件在互联网上的整体传播情况，收集全网数据进行分析，自动 生成涵盖事件简介、事件走势、网站统计、关键词云、热门信息、热点网民、传 播路径、舆情总结等多个维度的全网事件分析报告，并进行可视化展示；

(2)口碑分析与情感趋势

揭示影视作品在过去一年内的讨论量趋势及观众态度占比，全方面分析作品口 碑，给予多维度（剧情、演员表现、服化、特效等）评分，构建影视画像，从影视舆论当中发现潜在的影视市场空白点与机会；

(3)营销效果分析

了解社会化营销事件的关注度，粉丝构成，传播途径，意见领袖参与度与传播声 量等，量化社会化营销效果，分析舆情影响带来的潜在价值；

(4)舆情预警

对海量网络舆论进行采集分析，并识别其中的关键舆情信息，及时通知到相关 人员，实现第一时间应急响应，提供邮件、短信、微信、APP 等多种舆情预警方 式；

(5)多行业舆情数据支持

对海量网络舆论进行采集分析，并识别其中的指向性的舆情信息，向政府、金融、 传媒、教育等多个领域提供数据支持和服务。

1.4 结构安排

本文以项目章程、问题陈述、需求分析等为基础，通过定义项目的范围，收集项目的功能性需求和非功能性需求，对项目进行概要设计，并对提出方案进行描述和可行性分析。主要内容如下：

第一部分是引言，主要介绍项目的背景、功能和本文档的内容与结构安排；

第二部分是对系统方案的建议，包括可行性分析；

第三部分是对系统需求的描述和分析；

第四部分是对业务人员的选择方案的建议；

第五部分是与本文档有关联的文档附录。

本文档的结构安排如下：

1 引言

1.1 目的

1.2 背景

1.3 项目简介

1.4 结构安排

2 使用的工具和技术

2.1 方案建议

2.2 可行性分析

3 信息系统需求

4 建议

5 附录

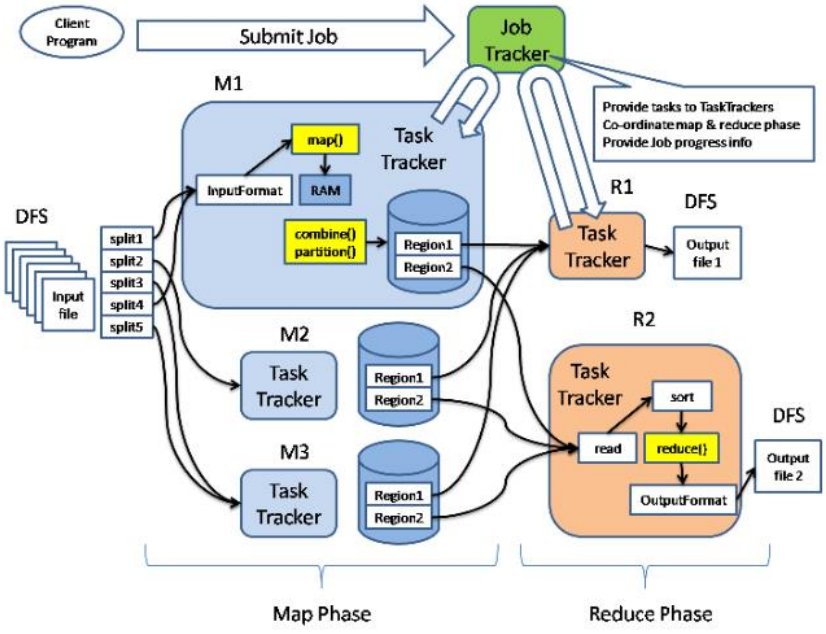
2 使用的工具和技术

2.1 方案建议

（1）Hadoop

针对网络上爬取的大量数据，使用 Hadoop 框架对大量数据进行分布式处理。

Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力进行高速运算和存储。Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称 HDFS。HDFS 有高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求，可以以流的形式访问（streaming access）文件系统中的数据。Hadoop 的框架最核心的设计就是：HDFS 和 MapReduce。HDFS 为海量的数据提供了存储，而 MapReduce 则为海量的数据提供了计算。下面是 Hadoop 的体系结构：



它具有以下几个优点：

1.高可靠性。Hadoop 按位存储和处理数据的能力值得人们信赖。

2.高扩展性。Hadoop 是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

3.高效性。Hadoop 能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

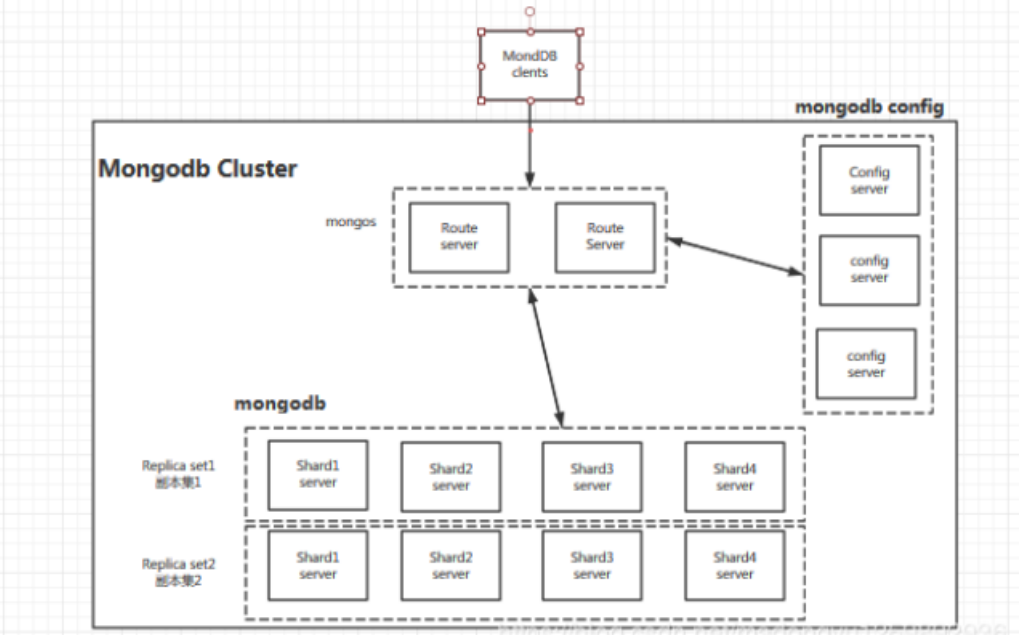
4.高容错性。Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

5.低成本。与一体机、商用数据仓库以及 QlikView、Yonghong Z-Suite 等数据集市相比，hadoop 是开源的，项目的软件成本因此会大大降低。

（2）Mongodb

针对网络上爬取的大量数据，使用 Mongodb 数据库进行存储。

MongoDB 是一个基于分布式文件存储的数据库。由 C++语言编写。旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。MongoDB 是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。它支持的数据结构非常松散，是类似 json 的 bson 格式，因此可以存储比较复杂的数据类型。Mongo 最大的特点是它支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。下图为混合部署方式的底层原理：



MongoDB 的设计目标是高性能、可扩展、易部署、易使用，存储数据非常方便。其主要功能特性如下。

1.面向集合存储，容易存储对象类型的数据。在 MongoDB 中数据被分组存储在集合中，集合类似 RDBMS 中的表，一个集合中可以存储无限多的文档。

2.模式自由，采用无模式结构存储。在 MongoDB 中集合中存储的数据是无模式的文档，采用无模式存储数据是集合区别于 RDBMS 中的表的一个重要特征。

3.支持完全索引，可以在任意属性上建立索引，包含内部对象。MongoDB 的索引和 RDBMS 的索引基本一样，可以在指定属性、内部对象上创建索引以提高查询的速度。除此之外，MongoDB 还提供创建基于地理空间的索引的能力。

4.支持查询。MongoDB 支持丰富的查询操作，MongoDB 几乎支持 SQL 中的大部分查询。

5.强大的聚合工具。MongoDB 除了提供丰富的查询功能外，还提供强大的聚合工具，如 count、group 等，支持使用 MapReduce 完成复杂的聚合任务。

6.支持复制和数据恢复。MongoDB 支持主从复制机制，可以实现数据备份、故障恢复、读扩展等功能。而基于副本集的复制机制提供了自动故障恢复的功能，确保了集群数据不会丢失。

7.使用高效的二进制数据存储，包括大型对象（如视频）。使用二进制格式存储，可以保存任何类型的数据对象。

8.自动处理分片，以支持云计算层次的扩展。MongoDB 支持集群自动切分数据，对数据进行分片可以使集群存储更多的数据，实现更大的负载，也能保证存储的负载均衡。

9.支持 Perl、PHP、Java、C#、JavaScript、Ruby、C 和 C++语言的驱动程序，MongoDB 提供了当前所有主流开发语言的数据库驱动包，开发人员使用任何一种主流开发语言都可以轻松编程，实现访问 MongoDB 数据库。

10.文件存储格式为 BSON（JSON 的一种扩展）。BSON 是对二进制格式的 JSON 的简称，BSON 支持文档和数组的嵌套。

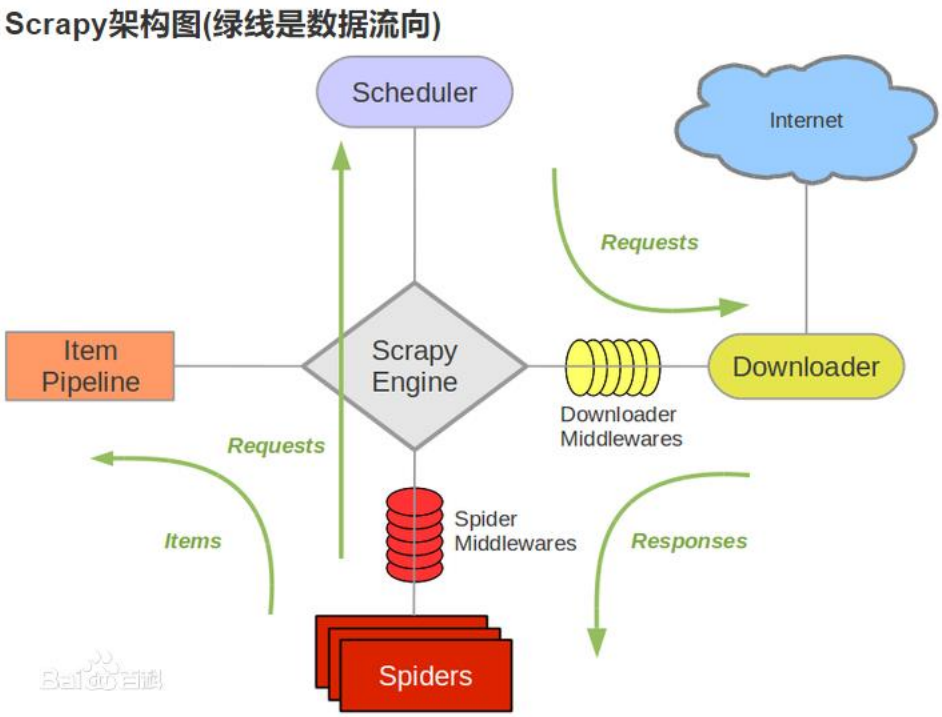
11.可以通过网络访问。可以通过网络远程访问 MongoDB 数据库。

（3）Scrapy

使用 Scrapy 在微博、头条等网站上爬取热点事件、热点人物等相关信息。

Scrapy 是适用于 Python 的一个快速、高层次的屏幕抓取和 web 抓取框架，用于抓取 web 站点并从页面中提取结构化的数据。Scrapy 用途广泛，可以用于数据挖掘、监测和自动化测试。 Scrapy 吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。它也提供了多种类型爬虫的基类，如 BaseSpider、sitemap 爬虫等，最新版本又提供了 web2.0 爬虫的支持。

其架构图如下：



Scrapy Engine(引擎)：负责 Spider、ItemPipeline、Downloader、Scheduler 中间的通讯，信号、数据传递等。

Scheduler(调度器)：它负责接受引擎发送过来的 Request 请求，并按照一定的方式进行整理排列，入队，当引擎需要时，交还给引擎。

Downloader（下载器）：负责下载 Scrapy Engine(引擎)发送的所有 Requests 请求，并将其获取到的 Responses 交还给 Scrapy Engine(引擎)，由引擎交给 Spider 来处理。

Spider（爬虫）：它负责处理所有 Responses,从中分析提取数据，获取 Item 字段需要的数据，并将需要跟进的 URL 提交给引擎，再次进入 Scheduler(调度器)。

Item Pipeline(管道)：它负责处理 Spider 中获取到的 Item，并进行进行后期处理（详细分析、过滤、存储等）的地方。

Downloader Middlewares（下载中间件）：一个可以自定义扩展下载功能的组件。

Spider Middlewares（Spider 中间件）：一个可以自定扩展和操作引擎和 Spider 中间通信的功能组件。

2.2 可行性分析

本节我们将主要对项目的经济效益进行分析，从而说明项目在经济上的可行性。基于系统开发的生命周期，对系统的预算从人员和设备两方面进行评估。

岗位	人数（人）	平均月薪（元/月）
系统架构师	1	29,600.00
前端工程师	1	17,680.00
后端工程师	2	23,580.00
测试工程师	1	10,721.00
合计	5	5 81,581.00

注：忽略开发人员工作年限、经验和开发岗位实际工作内容造成的薪资差异，以平均薪资作为评估标准

- 设备费用

  以阿里云服务器为参考预算评估标准：

设备	数量（个）	平均价格（元/月）
云服务器 ECS (32vCPU/128GiB)	6	39,310.80
合计	6	39,310.80

- 其他预算
  - 场地租赁费：北京办公租金平均总价 14.2 万元/月/套
  - 管理和其他费用

    如在项目需求分析阶段通过电话、会议等方式与项目利益关系人进行交谈从而收集需求并最终确认、需求分析书等纸质报告的制作费用等，约计2500元/月。

- 项目总经费预览

参数	预算（元）	含义	说明
C	——	项目总经费	——
L	210，322.00	开发人员预算	以开发周期为一年计算
E	471，729.60	设备费用	以租赁周期为一年计算
O	6，429.91	管理和其他费用	以整个开发周期计算

注：管理和其他费用整体评估以某企业2018 年年报数据[2]——其他关联交易：向北京市天元网络技术股份有限公司支付水电费、房租费等 771,588.64 元为参考，以5%作为项目预算评估。

由此，可得出初期项目开发的预算为

$$C = L + E + O = 688,481.51（元）$$

若对设备租赁进行会计确认，则初期项目的开发预算为

$$C = L + E + O = 216,830.53（元）$$

	第0年	第1年	第2年	第3年	第4年	第5年
开发成本	(216830.530)					
运行/维护成本		(173464.424)	(156117.982)	(140506.183)	(126455.565)	(113810.009)
折现系数	1.000	0.893	0.797	0.712	0.636	0.567
成本的现值	(216830.530)	(154903.731)	(124426.031)	(100040.403)	(80425.739)	(64530.275)
累积的成本现值	(216830.530)	(371734.261)	(496160.292)	(596200.695)	(676626.434)	(741156.709)
收益	120472.170	144566.604	173479.925	208175.910	249811.092	299773.310
收益的现值	120472.170	129097.977	138263.500	148221.248	158879.854	169971.467
累积的收益现值	120472.170	249570.147	387833.647	536054.895	694934.750	864906.216
累积的净现值	(96358.360)	(122164.113)	(108326.645)	(60145.799)	18308.316	123749.507

根据上表数据分析，系统的全生存期的投资回报率为57.1%，年平均投资回报率为11.4%。

3 信息系统需求

3.1 首页模块

本模块为进入本系统的第一个界面，主要展示本系统的简要功能，提供用户登录、注册和进入个人中心的入口。平台提供的入口根据用户类别和功能需求可分为个人用户和企业用户，前者根据自身对影视舆情的兴趣浏览或搜索有关舆情事件或影视参与者及其相关信息系统，后者除了可以获取平台提供的舆情数据，还可以根据需求定制预警服务和分析报告等。

3.1.1 用户登录

用户打开平台首页页面后，状态为未登录的用户通过输入账号和密码进行登录操作，系统将在登录过程中对用户账号及密码的正确性进行验证，并判断用户为个人用户还是企业用户，在登录后根据不同的账号类型在部分页面进行不同的页面显示。

3.1.2 实时展示热点舆情

用户打开平台首页页面后，即可浏览当前系统推荐的热点影视舆情事件，并附有热点事件的简要说明，想要深入了解的用户需要完成登录，再进入舆情事件详情界面了解事件舆情。

3.1.3 查看个人中心

用户完成登录后，根据用户的账号类型、个人信息和其他不同数据展示出不同的个人中心界面。

3.1.4 功能入口

用户进入首页后，查看系统功能，若需使用各项功能需要用户完成登录并获取相应权限。

3.2 个人中心模块

展示用户个人账号基本信息，也可以查看一定事件段内的历史浏览记录和近期主要关注事件、人物等，并可查看相关数据。此外，还可以允许用户查看充值记录、系统消息等。企业用户还要能查看预警数据及详细说明和历史分析报告。

3.2.1 查看账号信息

用户进入个人中心界面后，查看自己的账号信息，用户也可以在此界面对自身的资料进行修改。

3.2.2 查看历史浏览数据

用户进入个人中心界面，查看账号的历史浏览数据，方便用户查看之前没有保存的数据。

3.2.3 查看近期关注

用户进入个人中心界面，点击近期关注按钮，查看该用户近期关注的热点事件或其他内容，系统也会根据关注推荐相关热点事件。

3.2.4 查看系统消息

用户进入个人中心界面，点击系统消息按钮，查看该用户收到的系统消息。

3.2.5 查看历史分析报告

企业用户在进入个人中心界面后，查看历史分析报告。

3.2.6 查看预警详情

企业用户在进入个人中心界面后，查看预警详情。

3.3 个人用户模块

购买个人服务的用户，可以通过平台首页进入个人用户页面，而未获得此项服务的用户（或游客账号）在点击首页入口时，则会由于权限认证失败而收到邀请购买服务的询问。在进入个人用户页面后，用户可以通过平台提供的类别，如人物（导演、艺人、自媒体制作人等）和题材（电视剧、电影、综艺等）筛选热点舆情事件，也可以在搜索栏中通过关键词进行筛选。在点击相关词条后，可以进入舆情事件的详情界面。用户也可以根据喜好将关注的影视舆情事件参与者、影视剧作品等设为关注列表添加到该页面。

3.3.1 筛选影视舆情事件

本功能主要为用户提供快速筛选出想要了解的事件，方便用户获得想要了解的最新事件以及当前事件的最新走向的信息。

3.3.2 添加关注

用户可对最近比较关心的事件和任务等添加关注，添加后，系统可以为用户进行实时跟踪，关注事件的未来发展走向。

3.4 企业用户模块

购买企业服务的用户，可以通过平台首页进入企业用户页面，而未获得此项服务的用户（如游客账号或个人用户账号）在点击首页入口时，则会由于权限认证失败而收到邀请进行企业认证并购买服务的询问。在进入企业用户页面后，用户可以通过平台提供的类别，如人物（导演、艺人、自媒体制作人等）和题材（电视剧、电影、综艺等）筛选热点舆情事件，也可以在搜索栏中通过

关键词进行筛选。在点击相关词条后，可以进入舆情事件的详情界面。用户也可以定制某一舆情事件或影视及相关人员的舆情预警、期末查看本月（或年）的舆情数据分析报告，并可以定制营销效果分析。

3.5 舆情事件模块

购买平台服务的用户，在通过系统认证后，可以通过不同的功能入口进行影视舆情事筛选，并经由筛选出的词条进入详情界面查看影视参与者或影视作品的有关信息，例如事件的主要人物、人物关系图谱、事件时间轴、事件情感分析与关注度等，通过关系图谱和时间轴的结点可以查看相关信息。

3.5.1 查看舆情事件详情

用户点击具体影视舆情事件（包括首页热点事件、自主筛选事件和相关事件引用）等，获得授权认证的用户可以查看具体事件中的参与人物及其关系图谱、该事件的时间轴和情感分析走向与事件关注度。

3.6 后台管理模块

系统后台的用户管理模块分为个人用户和企业用户。管理员需要能查询用户并管理与用户相关的信息。

3.6.1 后台管理

管理员通过管理员账号登入后台管理页面，对既有授权用户信息和数据的查看、维护，保持平台安全、平稳的运行。

4 建议

本系统的开发在当前的社会环境是很有必要的，几乎所有人都使用网络，在网络上对热点事件发表自己的意见，尤其是在娱乐方面，人们的关注往往会更多。对此，我们对项目提出以下建议。

4.1 对项目领导的建议

项目领导决定开发出这样的软件一定是从中看到了市场的需要，所以领导要时刻关注系统的开发进度以及系统的需求变化，因为领导是统筹整个系统的关键，只有项目领导者对项目有足够的了解，才会使项目的利益最大化。

在了解整个项目的基础上，既了解项目的进度有了解市场，才能将需要的功能完善到最好，达到项目开始时定下的项目目标。

4.2 对开发人员的建议

由于当前也有很多类似的系统，所以想要在其中脱颖而出，就需要系统具有更加精准的功能。首先，就需要庞大的数据量，本系统应该可以广泛的从网络上收集相关的信息，尤其是一些人们主流的社交媒体，例如微博等等。

还有就是在系统的算法上，系统应该可以先根据事件的整体舆情分析出事件的初步走向，然后就需要系统可以随着事件的发展，实时的对舆情进行监控，对事件未来的发展等进行预测。

由于本系统主要面向于企业，那就需要系统具有良好的安全性，可以保护好用户的信息，防止用户使用本项目后，发生数据信息泄露的情况。界面的设计也应该以简洁为主，方便用户的操作，使用户具有良好的体验。

4.3 对宣传方面的建议

系统的宣传应该建立在用户的体验上，良好的用户体验是项目宣传最有利的工具。由于本项目为分析热点舆情的系统，主要的收益来源就是需要这些分析的企业，所以我认为第一步推广到一些需要的企业，这样，企业在体验过本系统的功能后，就可以成为本系统的长期用户，有利于本系统的发展。

5 附录

《大数据舆情分析系统需求分析报告》

《大数据舆情分析系统业务分析报告》

《大数据舆情分析系统概要设计说明书》

《大数据舆情分析系统详细设计说明书》

《大数据舆情分析系统技术方案说明书》