舆论分析平台 产品构思

# 问题描述

微博用户总量及其使用频度的急速增长在一定程度上反映了当前人们对于信息获取及输出方式的转变。相较于以往被报纸、广播、电视等传统媒体单方面的灌输信息，人们变得更加倾向于在微博平台上主动发现自己感兴趣的内容，并在获取信息的同时，也能通过发布微博来对外输出自己的见解、态度和意见。正是由于民众这种对于信息获取及输出方式的转变,使得微博等社交网络平台逐渐成为了民众对外发声的重要渠道之一。众多用户在微博平台上围绕国家政策、民计民生、娱乐八卦等话题不断产出各类体现着个人意志的舆论数据，这些舆论数据随着时间不断增长，逐渐演变成为网络舆情，而微博也变成了各类网络舆情的主要载体。网络舆情是现实舆情在互联网上的投影，它是指在互联网环境下，针对各类社会问题、现象，多数群众所持有的意见、态度和情绪。如果能正确把握网络舆情走向，就相当于掌握了网络舆情背后的人(发起和传播者)的兴趣喜好、价值取向，从而为迎合用户需求或疏导舆论走向提供了实践基础;但相反，如果不能在网络舆情发生的萌芽阶段对其进行控制，带有主观性的意见和情绪就会通过互联网不断扩散，进而演变为群体事件，引发难以预估的损失l4。因此能够合理引导舆论导向，监督和约束网络行为，并具备一定的危机舆情预警机制就变得十分重要，而网络舆情分析及系统构建也就变成一种重要且势在必行的研究和应用。

网络舆情分析相较于传统的舆情分析研究,在数据采集层面可借助计算机进行自动化信息采集，不用再耗时耗力经过走访、民意调查等方式来获取舆论数据。但目前面向微博的网络舆情分析工作中，原始舆论数据的自动化采集还面临诸多问题。一方面官方提供的数据接口每分钟允许调用次数上限仅为150次，完全不能满足舆情分析基本需求;另一方面传统的面向微博的网络爬虫爬取到的数据大多格式混乱，干扰数据多，导致后期数据处理难度大，进而延长了网络舆情分析工作的工作周期。

除数据采集过程中存在的相关问题外，在舆情情感分析策略层面，现阶段的网络舆情分析系统也存在不足之处。在现有的网络舆情分析系统中，大多采用一种固定的情感分析算法对文本情感倾向进行分析、判断，用户只能被动的接收分析结果信息。这就导致当系统分析结果出现较大误差时，很难通过其他方式对该误差进行校准。一旦错误的分析结果呈送到用户面前，就会严重损害用户对于系统的信赖程度。

针对上述问题，本产品以微博为切入点，面向网络舆情分析领域，设计了一种高效的数据爬虫，并结合多种基于文本的情感分析方法，研发出一套面向微博平台的网络舆情分析系统。用户能够在该系统平台上获取更加精准的舆情分析结果，方便用户根据舆情分析报告及时制定相关措施进行正确的舆情引导。

# 产品愿景和商业机会

**定位：**

通过情感倾向分析技术，可以快速关注到事件、言论、评论的舆论方向，及时处理负面新闻，了解竞品动态。

通过新闻摘要技术，可以准确的为针对待发布的新闻进行摘要，从而提高传统人工提炼摘要的效率。

通过评论观点抽取技术，将能够将特定行业的海量评论内容进行观点抽取和情感分析，从而对产品的口碑、舆情进行有效监控和分析

**商业机会：**

* + 与各大产品公司合作，快速了解对于该公司产品的用户评价，建议和意见。
  + 社交平台可快速了解现如今社会热点话题，网民对事件的看法，即使发现处理负面信息
  + 娱乐公司可以快速发现旗下艺人的各种信息的舆论走向，即使发现处理负面新闻

**商业模式**

* 为客户提供数据，与分析结果

# 用户分析

本电子商务网站主要服务两类用户：

* 政企部门
  + 快速了解国民对新政策的看法
  + 发布更加便民、亲民的政策
  + 优势：更加亲民
* 娱乐公司
  + 痛处：各种绯闻、负面新闻，对旗下艺人造成不良影响
  + 优势：更加快速，提前了解信息，做到提前预知舆情走向，快速做出处理
* 产品公司
  + 快速发现民众对于公司产品的看法、意见、建议
  + 针对民众的想法，对产品做出改进优化，或声明设计用意等

# 技术分析

采用的技术架构

分布式网络爬虫设计及实现:从总体架构设计、分布式设计优化、解析器与下载器交互优化、解析器设计优化等方面，详细的设计并实现了一款分布式网络爬虫

舆情分析各功能模块的算法应用实现:采用TF-IDF算法、基于情感词典的情感分析方法、基于朴素贝叶斯的情感分析算法、k-means算法、深度学习等文本分析方法对用户发布的舆论文本进行了关键词抽取、观点挖掘、情感倾向性分析、文本聚类，并在此基础上实现了用户观点挖掘，文本情感分析，热点话题发现和目标舆情预警等功能;。

平台

初步计划采用阿里的云服务平台支撑应用软件，早期可以使用一年的免费体验，业务成熟后转向收费（价格不贵）；

软硬件、网络支持

由于所选支撑平台均是强大的服务商，能满足早期的需求，无需额外的支持；

技术难点

数据爬取的困难性及全面性，许多网站的反爬技术很难攻克，导致数据不全面。

新词热词的出现，对于文本处理有一定的影响。

# 资源需求估计

人员

产品经理：依据本产品的背景和定位，吸取已有舆论分析网站的成熟经验，结合特点和用户特征，设计符合舆论走向预测及分析的产品

IT技术专家：快速架构和实现产品，同时确保对未来新事件的发现等支持。

人员代表：有相关信息需求的人员代表，提出合理的意见和建议

资金

产品验证阶段前暂无需要。完成产品后，需要资金进行推广。

设备

一台本地PC服务器；

设施

10平米以内的固定工作场地；

# 风险分析

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **事件描述** | **根本原因** | **类型** |
| R1 | 网络信息混杂 | 信息时代，信息传播速度较快，缺少一定的实时性 | 商业风险 |
| R2 | 网络信息庞大 | 对于信息处理，不免有‘垃圾信息’，造成误差 | 用户风险 |
| R3 | 无法实现低于2小时的快速送货 | 2小时从该市的小商品市场到该市的任何一个地方都足够了，真正时间的消耗主要在响应订单、准备货物和到达目的地后快速联系用户 | 流程风险 |
| R4 | 人员不能及时到位 | 无法快速组建技术团队 | 人员风险 |
| R5 | 无法获得足够的推广费用 | 产品快速推广时，需要大量的资金，目前团队不具备，需要寻找投资 | 资金风险 |

# 收益分析

财务分析的估算结果如下，几项重要参数说明：

1. 折现率假设为10%，这是比较通用的一个值；
2. 项目长周期设为5年；
3. 首年成本为上面资源分析中的成本加10万元推广成本，以后四年假设升级维护费和推广为每年20万；
4. 收益假设第一年为10万，第2年为30万，第3年为60万，第4年为100万，第5年为150万；

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 折现率 | 10% |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 | 汇总 |
| 成本 | 396000 | 200000 | 200000 | 200000 | 200000 |  |
| 折现因子 | 0.91 | 0.83 | 0.75 | 0.68 | 0.62 |  |
| 折现成本 | 360360 | 166000 | 150000 | 136000 | 124000 | 936360 |
| 累计成本 | 360360 | 526360 | 676360 | 812360 | 936360 |  |
|  |  |  |  |  |  |  |
| 收益 | 100000 | 300000 | 600000 | 1000000 | 1500000 |  |
| 折现因子 | 0.91 | 0.83 | 0.75 | 0.68 | 0.62 |  |
| 折现收益 | 91000 | 249000 | 450000 | 680000 | 930000 | 2400000 |
| 累计收益 | 91000 | 340000 | 790000 | 1470000 | 2400000 |  |
|  |  |  |  |  |  |  |
| 折现收益-折现成本 | -269360 | 83000 | 300000 | 544000 | 806000 | 1463640 |
| 累计收益-累计成本 | -269360 | -186360 | 113640 | 657640 | 1463640 |  |
|  |  |  |  |  |  |  |
| 净现值 | 1463640 |  |  |  |  |  |
| 投资收益率 | 156% |  |  |  |  |  |
| 投资回收期 | 第3年 |  |  |  |  |  |