目 录

1. 毕业设计（顶岗实习）概况 1

1.1. 实习公司和岗位概况 1

1.2. 课题背景和价值 1

1.3. 国内外研究状况 3

1.3.1 数据采集技术 3

1.3.2 针对微博的数据分析 3

1.4. 课题整体完成情况概况 4

2. 复杂工程问题归纳与实施方案可行性研究 5

2.1. 需求分析 5

2.2. 工程问题归纳 5

2.3. 解决方案与可行性研究 5

3. 针对复杂工程问题的方案设计和实现 7

3.1. 方案设计 7

4. 知识技能学习情况 8

4.1. 数据分析基础知识 8

4.2. 爬虫相关技术 8

4.3. 搜索引擎基础知识 8

4.3. Elastucsearch学习总结 8

4.4. Kinaba学习总结 8

5. 工程计划管控与执行情况 9

5.1. 项目开发计划和周期管理 9

5.2. 开发总结 9

6. 职业素养与工程伦理的学习与培养 10

6.1. 个人职业素养培养 10

6.2. 软件工程对社会的影响 10

7. 对软件工程实践以及软件工程领域发展的认识 11

7.1. 软件工程和可持续发展 11

7.2. 软件工程发展状况 11

8. 结束语 12

参考文献 13

# 1. 毕业设计（顶岗实习）概况

## 1.1. 实习公司和岗位概况

**公司简介。**SAP公司成立于1972年，总部位于德国沃尔多夫市，在全球拥有6万多名员工，遍布全球130个国家，并拥有覆盖全球11,500家企业的合作伙伴网络。作为全球领先的企业管理软件解决方案提供商，SAP帮助各行业不同规模的企业实现卓越运营。从企业后台到公司决策层、从工厂仓库到商铺店面、从电脑桌面到移动终端—SAP助力用户和企业高效协作，获取商业洞见，并从竞争中脱颖而出。SAP的软件和服务能够帮助客户实现盈利性的运营，不断提升应变能力，实现可持续的增长。全球188个国家的232,000家客户正在从SAP解决方案中获益, 其中包括财富500强80%的企业及85%最有价值的品牌。

**岗位职责。**参与思爱普PCM（客户渠道管理）系统的开发，负责前期接口的设计和数据库设计，负责前端和后端的实现，负责PCM集成微信和集成内部的C4C（客户云服务）系统。项目使用odata数据交换协议，后端开发框架使用springboot。

## 1.2. 课题背景和价值

**课题背景。**随着互联网技术的飞快发展。网络在人们生活中的普及率越来越高，成为人们生活中不可以或缺的一部分，特别是微博，QQ，微信朋友圈社交网络的出现标志着互联网新格局的到来，改变了我们获取信息的方式和途径，使得我们可以更快更及时的获取信息，可以足不出户的购买我们想要获得的商品，通过各种社交平台分享和交流自己的经验和感受，实时状态和位置信息。同时越来越多的人开始享受互联网带来的便捷和乐趣。根据中国互联网信息中心（China Internet Network Information Center，简称CNNIC）于2016年1月发布的第37次《中国互联网络发展状况统计报告》显示，截至2015年12月，中国网民的规模己达到6.88亿，全年共新增网民3951万人。互联网普及率为50.3%，较2014年底提升了2.4个百分点。其中使用手机上网的网民达到6.20亿，比2014年底增加了6303万人，占总体网民人数的90.1%，比2014年増长了4.3%。由此可以看出，以手机为代表的移动设备终端已经成为了我国网民上网的主要方式。

在综合社交领域，最具代表性的朋友圈、QQ空间、微博，网民使用率分别为65.1%、80.6%、33.5%。其中QQ空间主要满足用户对个人关系链信息的需求,在产品形态和商业营销方面一直坚持变革，凭借良好的用户基础,在基于大数据的关系营销方面做了诸多有益的探索，回报显著；朋友圈则通过其优秀的个人隐私保护获得了很多青少年的青睐，并且正慢慢地获得更多较高年龄阶段的用户，称为中国目前现在占有率最高的熟人社交内容分享平台；微博则主要满足用户对兴趣信息的需求，是用户获取和分享"新闻热点"、"兴趣内容"、"专业知识"、"舆论导向"的重要平台。同时，微博在帮助用户基于共同兴趣拓展社交关系方面也起到了积极的作用。微博通过坚持去中也化战略,扶植各垂直行业自媒体，刺激原创内容产生，优质内容吸引和维持用户的活跃，用户规模稳步增长，内容平台价值得到进一步提升。

高速发展的互联网产生了大量的数据，大量的数据必然会有大量有用的信息，因此，“大数据”由此成为时下最为火热的词汇之一，数据仓库，数据分析和数据挖掘以及数据安全等围绕数据商业价值的利用被各行各业的人相互追捧。大量数据的出现，正引发全球范围内的商业技术和商业变革。从海量的大数据中，我们可以快速的获得更价有意义和有价值的信息，由此衍生出了数据挖掘方面的研究。数据挖掘是一项借助工具探测海量数据中有意义的信息的一种行为，借用该过程获得的信息，我们可以很方便的支撑其它的商业和科研活动，通过充分得使用数据，挖掘数据，我们可以从中发掘出潜在问题，作出更加合理的决策。

**课题价值。**新浪微博（http://weibo.com）是新浪旗下网站，是中国最具影响力的社会化媒体平台，是一个基于用户关系的信息分享、传播以及获取的平台。用户通过WEB、WAP以及各种客户端，以不超过140个字的文字更新信息，实现即时分享。微博作为用户占有率排行靠前的社交平台，拥有大量的用户产生的数据，这些数据隐藏了很多有价值的信息，因此，针对微博的数据挖掘和分析工作已经引起了越来越多的人的关注和重视。微博具有良好的互动性，覆盖领域多样化，用户人数多，热点信息传播快，用户范围广的特点，因此针对微博的数据分析能够更加全面，分析微博数据所获得的信息对于实际需求也更佳有价值和意义。

通过爬取微博树洞数据，可以获取关注树洞的活跃用户的明显特征，从多个数据维度对人群进行分类，分析用户行为。例如通过分析树洞动态评论数据，可以获得关于该树洞人群的微博使用习惯数据，通过分析树洞人群的个人信息，可以获取树洞人群特征信息。除此之外，企业也可以对这些数据进行深度挖掘，获得更加具有商业价值的信息，如通过热点和用户特征进行活动和精准用户广告的投放。

目前，还没有一个系统能够直接对这样的数据进行相关的分析，本课题的系统以新浪微博为基础，通过爬虫的方式，采用多线程来快速获取微博上面的数据并且将微博数据存储在Elasticsearch中，按照微博动态内容，用户关系，用户评论数据，个人信息进行多维度的数据分析，通过用户界面能够直观的展示树洞人群特征。

同时通过这个课题，对我个人也有很大的提升，能够加深我对数据分析的学习，开扩了自己的技术眼界，从之前的开发应用程序产生数据，延伸到学习如何在数据中获取信息，利用信息产生更大的价值，让我明白数据对于提高生产力的作用。

## 1.3. 国内外研究状况

### 1.3.1 数据采集技术

**基于网站的信息抓取。**这种方法是从一些基础URL出发，采用深度优先或者广度优先策略进行爬取，逐步扩展到整个网站。这种方法的优点是信息全、覆盖面广，主要被各大搜索引擎所使用，其缺点是爬行耗时长，不利于最新信息的搜集，此外由于所需的硬件资源多，只有少数几个服务商能够采用这种技术。

**基于主题的信息爬取。**这种爬虫的特点是先进行网络内容的判断，如果与想要爬取的主题一致，再进行爬取。由于只对某一个主题进行爬取，所以此种爬虫的优点是采集的页面针对性很强，效率也较高，比较适用于专门对某一个问题进行研宄的研究人员。如果想要先采集数据，再进行数据分析，则此种爬虫不适用。

**增量式信息爬取。**此种爬虫是针对第一种网络的对新信息反应慢的缺点而提出的。它只对有变化的网页进行再爬取，而不对没有变化网页进行爬取。此种爬虫的效率高，在信息量大的今天有很强的实际意义，它的难点在于如何判断网页是否有变化，所以它的爬行效果完全取决于所使用的判断算法的有效性。

**分布式网络爬虫。**由于目前网络上的信息量巨大，如果只使用一台计算机进行采集，耗时巨大，信息更新也慢，所以研宄人员提出了使用很多台计算机同时对网络进行信息采集，这样一来可以大大提高爬行的速度和效率，但是如何进行爬行任务的分工以及多台计算机如何协调配合，使每台的工作压力相对均衡是此种爬虫的设计难点。

### 1.3.2 针对微博的数据分析

**微博短文本挖掘。**由于微博的长度较短，所以可以把微博上的内容看成是不同短文本的集合，这个集合有一些显而易见的特点:文本口语化，存在大量错别字、谐音字、符号等;信息碎片化，文本的特征不明显，意思表达不清楚;信息量巨大，又由于转发量大，原创数量小，微博内容的重复性很高。为弥补短文本特征值的不足，S.Bharath将用户所发的微博按所涉领域进行划分，然后对不同微博进行分类，这样提高了短文本的特征值。而针对微博内容的重复性，王永恒提出了用于聚合大量短文本的方法，通过聚合，可以使得重复性的文本大大减少，提高了剩余文本间的区分度

**情感倾向性分析。**人们在所发的微博中，通常都带有自己的情感倾向。这些情感倾向通过使用语气词、具有判断性的词汇表达出来，情感倾向对于研究诸如用户满意度、大众对社会热点的态度等尤为重要。分析人们的情感倾向，最常用的方法就是建立情感词库，然后进行词汇的匹配和表情、语气等的标记，以此来发现人们的倾向另外，也出现了一些情感分类的算法，如最大信息熵、向量机、贝叶斯算法等，在使用此类算法时，为了提高正确率，除了进行情感的分类外，还需要结合上下文进行判断。

**话题趋势检测。**由于微博具有快速更新的特点，一个热点话题可以在极短的时间内广泛传播，造成巨大的影响。所以快速的找出热门话题，判断话题传播的方向、发展的趋势等，就能更好的掌握舆论的走势，为下一步的行动争取时间。最常见的话题趋势检测就是进行词频的检测，通过检测某个词出现的频率，就能判断某个话题的热门程度。除此之外，可以结合PageRank算法，将转发量、作者的重要性、词频等作为加权因素，就可以计算出此话题的重要性，并判断其未来的走势。

## 1.4. 课题整体完成情况概况

**完成爬虫模块程序设计和实现。**爬虫模块支持多线程和代理池进行数据爬取，同时通过使用文件缓存的方式来记录爬虫运行时的URL爬取记录实现对爬虫的暂停，启动和终止（终止使用JMX接口实现）。能够批量爬取微博动态，评论和用户数据，能够将这些数据处理后发送到Elasticsearch中存储。

**完成分析数据对象建模。**本课题中的主体数据时微博动态，评论和用户，由于数据需要存储在Elasticsearch中，因此需要定义好数据存储的元数据，才能通过Elasticsearch实现对数据的分析。

**完成数据分析。**借助Kibana提供的操作Elasticsearch的用户界面和数据可视化能力，实现对获取的微博数据的分析。

**软件工程能力培养。**

# 2. 复杂工程问题归纳与实施方案可行性研究

## 2.1. 需求分析

业务需求。分析微博树洞数据，提取出有价值的信息

功能需求。

系统需求。

质量需求。

## 2.2. 工程问题归纳

系统模块划分。为什么划分系统模块

技术方案制定。怎样合理的制定技术方案

开发工作流。

程序兼容性设计。

## 2.3. 解决方案与可行性研究

数据采集解决方案。

数据存储解决方案。

数据分析解决方案。

持续集成。

项目管理。

数据分析解决方案。

持续集成。

项目管理。

数据分析解决方案。

持续集成。

项目管理。

数据分析解决方案。

持续集成。

项目管理。

数据分析解决方案。

持续集成。

项目管理。

# 3. 针对复杂工程问题的方案设计和实现

## 3.1. 方案设计

3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

3.1.2. Elasticsearch存储数据结3.1.1. 爬虫模块方案设计

# 4. 知识技能学习情况

## 4.1. 数据分析基础知识

## 4.2. 爬虫相关技术

## 4.3. 搜索引擎基础知识

## 4.3. Elastucsearch学习总结

## 4.4. Kinaba学习总结

# 5. 工程计划管控与执行情况

## 5.1. 项目开发计划和周期管理

## 5.2. 开发总结

# 6. 职业素养与工程伦理的学习与培养

## 6.1. 个人职业素养培养

## 6.2. 软件工程对社会的影响

# 7. 对软件工程实践以及软件工程领域发展的认识

## 7.1. 软件工程和可持续发展

## 7.2. 软件工程发展状况

# 8. 结束语

（前期任务完成度自我评价及后续工程环节的实施计划安排。除针对复杂工程问题的实施方案执行计划，还包括针对个人或职业发展需求的知识技能学习计划。）

# 参考文献

[1] 王浩刚，聂在平.三维矢量散射积分方程中奇异性分析[J]. 电子学报，1999, 27(12): 68-71

[2] X. F. Liu, B. Z. Wang, W. Shao. A marching-on-in-order scheme for exact attenuation constant extraction of lossy transmission lines[C]. China-Japan Joint Microwave Conference Proceedings, Chengdu, 2006, 527-529

[3] 竺可桢.物理学[M]. 北京：科学出版社，1973, 56-60