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. 课题整体完成情况概况 3

2. 国内外研究状况 4

2.1 国内研究现状 4

2.2 国外研究现状 5

1.3.1 数据采集技术 6

1.3.2 针对微博的数据分析 6

3. 复杂工程问题归纳与实施方案可行性研究 8

3.1. 需求分析 8

3.1.1 功能需求 8

3.1.2 非功能需求 10

3.1.3 数据需求 11

3.2. 工程问题归纳 11

3.3. 解决方案与可行性研究 11

3.4. 方案设计与实现 12

4. 知识技能学习情况 13

4.1. 数据分析基础知识 13

4.2. 爬虫相关技术 13

4.3. 搜索引擎基础知识 13

4.3. Elastucsearch学习总结 13

4.4. Kinaba学习总结 13

5. 工程计划管控与执行情况 14

5.1. 项目开发计划和周期管理 14

5.2. 开发总结 14

6. 职业素养与工程伦理的学习与培养 15

6.1. 个人职业素养培养 15

6.2. 软件工程对社会的影响 15

7. 对软件工程实践以及软件工程领域发展的认识 16

7.1. 软件工程和可持续发展 16

7.2. 软件工程发展状况 16

8. 结束语 17

参考文献 18

# 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。

微博一词译自英文单词“micro-blog”，是博客的微缩版或者变体。这一理念最早由博客先驱埃文，威廉姆斯提出并于2006年创办了世界上最早提供微博服务的网站Twitter。我国微博应用始于2007年。“无处不在的沟通”是它的理念，“What are you doing”是它的宣传口号。目前，网络时代已经来临，上网购物#查阅信息等已经代替了传统的报纸商店，随着微博的兴起，国内学术界对微博的研究日益增多，迄今为止已有许多学者从教育学，传播学，经济学等多种学科视角，运用定量分析或者定性分析以及定量与定性相结合的研究方法对微博进行分析。时隔两年，2009年国内互联网商家掀起建设微博平台热潮。2009年10月，新浪微博正式向公众开放，中国网民的微博时代到来，2010 年被称为微博元年。根据中国互联网络信息中心（CNNIC）的第31次“中国互联网络发展状况统计报告”，截至2012年12月底，我国网民规模达到5.64亿，其中微博用户规模为3.09亿，比2011年底增长了54.7％，同时手机微博用户规模达到2.02亿，占微博用户的65.6％。在这样的背景之下，微博及其相关话题成为政府、商务机构和寻常民众的热门话题，也是信息科学和社会科学共同关心的焦点之一。

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. 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。**通过这个项目，让我对使用软件工程开发模式管理软件开发周期有了更深层次的理解。能够使用软件工程知识来完成需求分析、项目周期管理、测试管理和持续集成。

# 2. 国内外研究状况

随着互联网技术的发展和收集数据手段的提高，微博积累的数据越来越大，形成了大数据的问题。现在常用的分析手段，如概率统计等，已经很难满足分析的需要。很多时候，人们并不知道如何来分析这些数据，因为人们看不到数据，形成了“数据盲人”。现在，可视化正作为一个新的领域蓬勃发展起来。可视化是解决这一问题的良好方法。

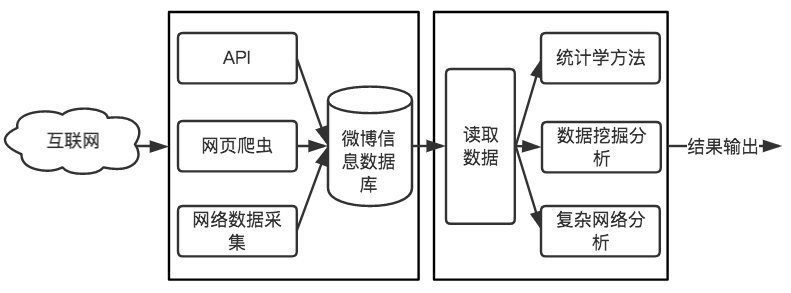
可视化可以粗略地被定义为通过图形的表现形式，进行信息传递、表达的过程。虽然我们今天谈到可视化一般是指利用计算机图形学和图像处理分析技术，将各种数据依据其特点转换为相应的图形图像，并提供界面实现人机交互的工作，但早在计算机发明之前，可视化就已为人类广泛应用。化技术最早运用于计算机科学中，并形成了可视化技术的一个重要分支——科学计算可视化(Visualization in Scientific Computing)。科学计算可视化能够把科学数据，包括测量获得的数值、图像或是计算中涉及、产生的数字信息变为直观的、以图形图像信息表示的、随时间和空间变化的物理现象或物理量呈现在研究者面前，使他们能够观察、模拟等。可视化的三个方向——科学可视化、信息可视化、可视分析密切相关，同时又各有特点，有其研究内涵与外延。科学可视化处理的对象包括医学、气象环境学、化学工程、生命科学、考古学、机械等领域的具有空间几何特征数据的时空现象，对测量、实验、模拟等获得的数据进行绘制，并提供交互分析手段。研究重点包括对表面、体绘制和对复杂数据中信息的选取、表达等。

## 2.1 国内研究现状

**主要研究的问题。**在计算机网络和通信等学科领域，微博的信息数据特征是近年来关注的主要问题，很多微博研究都基于网络信息数据测量分析展开，微博信息数据是指微博数据中心中存储的各类数据，主要包括：微博用户档案、微博用户关系、微博消息热点话题等，它是此类研究方法的基础。在这类研究中，大多以微博消息传播三大构件为研究对象，以微博消息传播和微博成员组织为主要研究内容，目的在于发现微博中的用户、消息传播、热点话题、用户关系网络等的规律。主要研究的问题如下。

1. 基于微博用户的研究，主要研究用户的行为特征及用户的影响力。
2. 基于微博用户关系的研究，主要研究用户关系网络的基本属性、关系网络生成和演进、微博人员关系挖掘、微博用户人际关系特点。
3. 基于微博内容的研究，主要研究微博消息内容特点、消息活跃时间特点、微博热点话题特点。
4. 基于微博消息传播的研究主要研究微博消息传播的特点，微博消息传播影响力。

**微博信息数据分析的研究方法。**基于信息数据分析的微博研究方法，以分析微博信息数据为基础，致力于发现微博中的各种规律和特点它可以分为两个阶段: 信息数据获取和信息数据分析。在信息获取阶段，主要任务是获取大量微博信息数据，主要采用三种方法: 基于微博第三方应用程序接口(Application Programing Interface， API)编程实现微博信息爬取，利用网络爬虫在微博网页上爬取关键字信息，利用网络数据采集设备直接获取微博服务网络传输数据。在数据分析阶段，主要任务是对微博信息数据进行特征提取和分析，挖掘出微博中的关键特征，采用的主要方法包括统计学数据分析方法、复杂网络分析方法、数据分类及挖掘方法等。微博信息数据分析方法基本流程如下图1.1所示。

图1.1 微博数据分析方法

基于微博信息数据分析的研究是微博研究中非常重要的一个方向，它是开展其他微博研究的基础，能够积累大量的微博特征，对微博的理论和实践研究都至关重要。

## 2.2 国外研究现状

首先介绍Twitter的概念：Twitter是一个在线社交网络服务，允许用户发送和阅读简短的在140个字符内的消息，而这条消息称为"推文(tweet)"。

Twitter平台包含了大量的用户信息W及丰富的内容，为预测研巧提供了重要的数据基础，因此也就出现了利用微博信息进行各种问题的预测。Jurgens等人利用地理位置推理算法对Twitter用户的地理位置进行了预测并取得了较好的预测结果。Bollen等人针对股票预测问题，用每天包含正面情感词和负面情感词的推文数比值来计算情感时间序列，通过自组织模糊神经网络(Self-Organizing Fuzzy Neural Network，简称SOFNN)预测道琼斯工业平均指数(Dow Jones Industrial Average，简称DJIA)。Wolfram等人使用SVM模型，从推文文本直接提取了特征属性来进行模型训练，预测纳斯达克指数。Zhang等人分析了微博情感和DJIA变化走势之间的相关性。欧美的学者还利用Twitter数据对大选进行了预测，而且逐渐成为了研究热点，取得了较好的预测效果。

### 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虫的速度和效率，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。

# 3. 复杂工程问题归纳与实施方案可行性研究

## 3.1. 需求分析

### 3.1.1 功能需求

功能需求按照系统需要实现的主要功能可以划分为：

1. 微博数据爬取模块。
2. 数据分析功能。
3. 数据可视化模块。

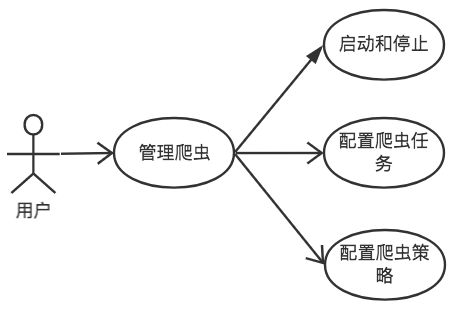
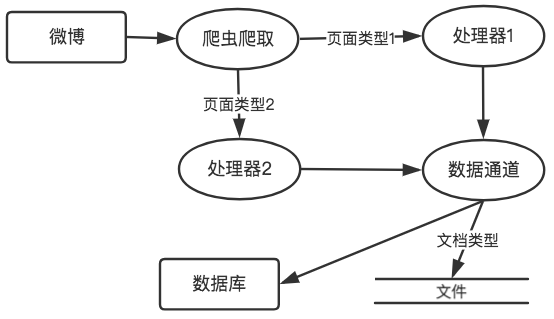
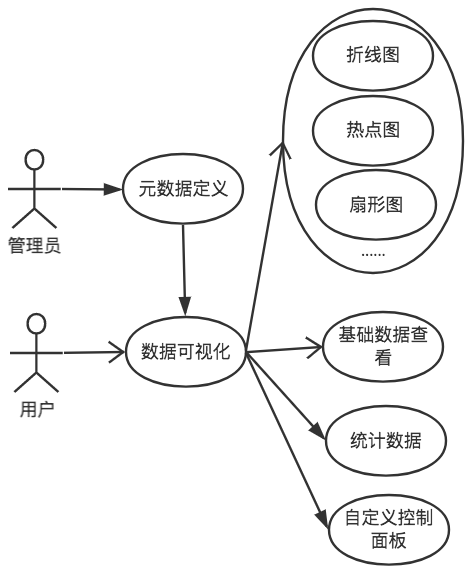
**微博数据爬取模块。**微博数据爬取模块需要实现对微博动态、评论以及用户信息的获取，支持爬虫的配置文件控制，能够通过配置文件设置和改变爬虫的行为，如爬虫启动任务的数量、爬虫的请求头、爬虫主机地址、文件缓存路径和代理池等。同时该模块还应该能够实现错误处理和非法功能，能够对爬虫运行时遇到的错误进行处理并且报告，避免因为错误异常导致爬虫程序的异常退出。该模块的用例图如下图2.1所示。

图2.1 爬虫模块用户用例图

**数据分析功能。**微博数据分析功能主要实现将爬取的微博数据进行一定的格式化处理后持久化存储在磁盘，以实现微博数据分析功能。该功能可以使用传统的数据库实现数据的存储，也可以使用基于文档的存储引擎进行数据的存储。数据流图如下图2.2所示。

图2.2 微博数据流图

**数据可视化模块。**数据可视化模块主要是对分析后的数据进行可视化，包括常用的折线图、柱状图、扇形图和热点图等，从而能够一目了然的从数据中获取有用的信息，该模块应该实现用户可以自定义数据可视化的方式，并且支持用户扩展对数据进行进一步的查询组合。可视化可以通过APP和网页两种方式提供，但是通过网页提供，应该能够支持通过媒体查询实现页面的自适应。该模块的用例图如下图2.3所示。

图2.3 数据可视化模块用例图

### 3.1.2 非功能需求

非功能性需求是指依一些条件判断系统运作情形或其特性，而不是针对系统特定行为的需求。包括安全性、可靠性、互操作性、健壮性、易使用性、可维护性、可移植性、可重用性、可扩充性。

本课题设计程序的非功能需求主要包括：

用户界面美观，具有针对性。所有页面应该支持主流的浏览器（IE9.0+，Chrome，Firefox等）下正常显示，而且不同浏览器看到的页面外观应该一致。

在操作上要易于使用。使用合理的存储方式来存储数据，保证数据的独立性，也保证了系统的可移植性。

数据存储量。使用合理的存储硬件，能够易于存储爬虫爬去的零散数据，确保不会因为硬件原因导致数据存储出错，进而影响数据分析的完成。

系统可靠性。在实现一定的吞吐量的基础上，需要满足系统运行期间不会出现严重错误导致异常退出，从而导致系统不可用、数据丢失等严重故障。

可扩展性。可扩展性主要指微博爬虫的配置应该足够灵活，且内部功能应该能够支持扩展和重用，如果有其他数据分析的需求，核心处理代码应该能够重用。另外一方面，系统在实现时应该考虑到日后若信息量较大，则系统可相应增加服务器实现扩展。

### 3.1.3 数据需求

功能需求和非功能需求都集中在系统需要做什么的层面上，对于该课题任务的主体数据却没有进行更多的定义，因此这一小节主要讨论系统的数据需求。

数据需求的基本定义：数据需求分析是从对数据进行组织与存储的角度，从用户视图出发，分析与辨别应用领域所管理的各类数据项（Data Items）和数据结构，形成数据字典的主要内容。

本课题中的三个主体数据对象为微博动态、微博评论、微博用户。数据分析对于这三个主题的数据对象应该包含足够的原始数据量。有效数据的比例应该大于90%，数据格式定义应该固定，类型使用合理。

## 3.2. 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。

## 3.3. 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。

## 3.4. 方案设计与实现

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