计算机科学与技术专业实习报告

姓名：王帆

学号：20152180

实习单位：天津理工大学

实习时间：2018 年 10 月-2018 年 12 月

1. 实习情况概述

天津理工大学是天津市属、以工为主，工理结合，工、理、管、文、艺等学科协调发展的多科性大学，入选“国家大学生创新性实验计划”，也是“卓越工程师教育培养计划”重点建设大学，是国家首批“卓越工程师教育培养计划 2.0”、“新工科研究与实践项目”入选高校。计算机科学与工程学院以本科教育为主，稳步发展研究生教育，同时承担全校计算机公共基础教学和学校校园网管理工作，积极开展中外合作办学。学院作为天津市地方高校中最早创办计算机本科专业的院系，经过 30 年的发展建设，现有计算机科学与技术、信息安全、信息与计算科学、物联网工程四个本科专业，与加拿大魁北克大学合作培养本科生；有计算机科学与技术一级学科博士点、博士后科研流动站，计算机科学与技术、软件工程、网络空间安全三个一级学科硕士点和计算机技术工程硕士授予权；计算机科学与技术、网络空间安全、软件工程三个学科为天津市“十三五”重点学科；有"计算机病毒防治技术"国家工程实验室、"计算机视觉与系统"省部共建教育部重点实验室、"智能计算及软件新技术"天津市重点实验室等研究平台；计算机科学与技术专业为教育部特色专业，信息安全专业入选教育部卓越工程师培养计划等。

在此次十二周实习的过程中，我进入天津理工大学计算机科学与工程学院 407 实验室，在李玉坤教授的指导下，完成有关数据抽取与集成系统设计方面的实习活动，为设计与实现高考志愿信息集成系统进行数据方面的准备工作。主要完成的内容为：了解数据集成与数据抽取方面的技术，学习相关工具与方法，实现对中国教育在线（eol.cn)与教育部高职院校信息系统相关数据的抽取与组织工作，以及基本的前端展示实现。

2. 所学知识在工程实践中的运用

本次专业实习中，我参与了高考志愿信息系统中关于高职高专院校信息系统非结构化数据抽取，组织，部署以及前端展示的部分内容。

在数据抽取方面，我主要使用 Java 程序设计语言，使用 jsoup HTML 解析器对静态页面进行信息筛选；在部分特殊动态页面数据处理过程中，主要使用 C# 进行预处理，将动态页面静态化，并使用 jsoup 对静态化后的数据源进行特征提取与数据抽取。

在数据部署方面，我运用数据库课程中学习到的数据库设计与实现方法，通过概念设计、逻辑设计、物理设计等数据库建模步骤，对数据库模型进行构建，使用 Java JDBC 进行数据库连接与部署，实现抽取后数据的组织过程。

最后，在前端展示与实现方面，我运用 ASP 技术，实现对前端页面的简要展示与基本实现，并使用 IIS 进行网站发布。

3. 职业道德和工作纪律执行情况

以实习单位为课堂，在具有丰富实践经验的导师以及学长的指导下，认真完成专业实习。在实习过程中，学生应努力做到：

1)自觉遵守国家法律法规和学校的实习纪律,严格按照规定时间进行实习。

2）不得提前结束实习，也不得未经批准随意延长实习时间。

3）遵守实习单位的劳动纪律和各项规章制度，树立良好的职业道德和组织纪律观念，与实习指导老师以及实习单位建立良好的关系。

4)对在实习中悉知的商业秘密要严格保密。借阅实习单位提供的各类文件、数据等资料，必须严格按照有关规定妥善保管，用后完整归还。

5）虚心学习，勤奋探索，认真求教；树立良好的精神风貌，维护学校班级名誉，遵守各项职业道德规范，提高职业素养，不得向实习单位提出不恰当的待遇要求。

在实习期间，我严格遵守各职业道德规范，尊师重道、勤恳虚心学习。在实习期间所用的电脑、设备等皆用心爱护保管，并在实习结束后及时完整的归还实验室。在实习期间，树立了良好的精神风貌，尊重实习单位各位老师，认真完成实习任务，并生成规范的实习报告。善于总结并吸取实践工作经验，与学校指导老师保持联系，及时汇报实习进程并向老师寻求实习指导与建议。

4. 实习情况与取得的成果

一、数据源整理与获取

数据源是数据的本源。互联网中繁杂的数据使人眼花缭乱，而科学地获取高信度数据的过程是一个值得研究的课题。这个部分，我的思路是：先总后分，先模糊后具体。

首先，通过搜索引擎，我们可以找到现有的较为完整且准确的数据源，如院校信息源、开设专业信息源等；

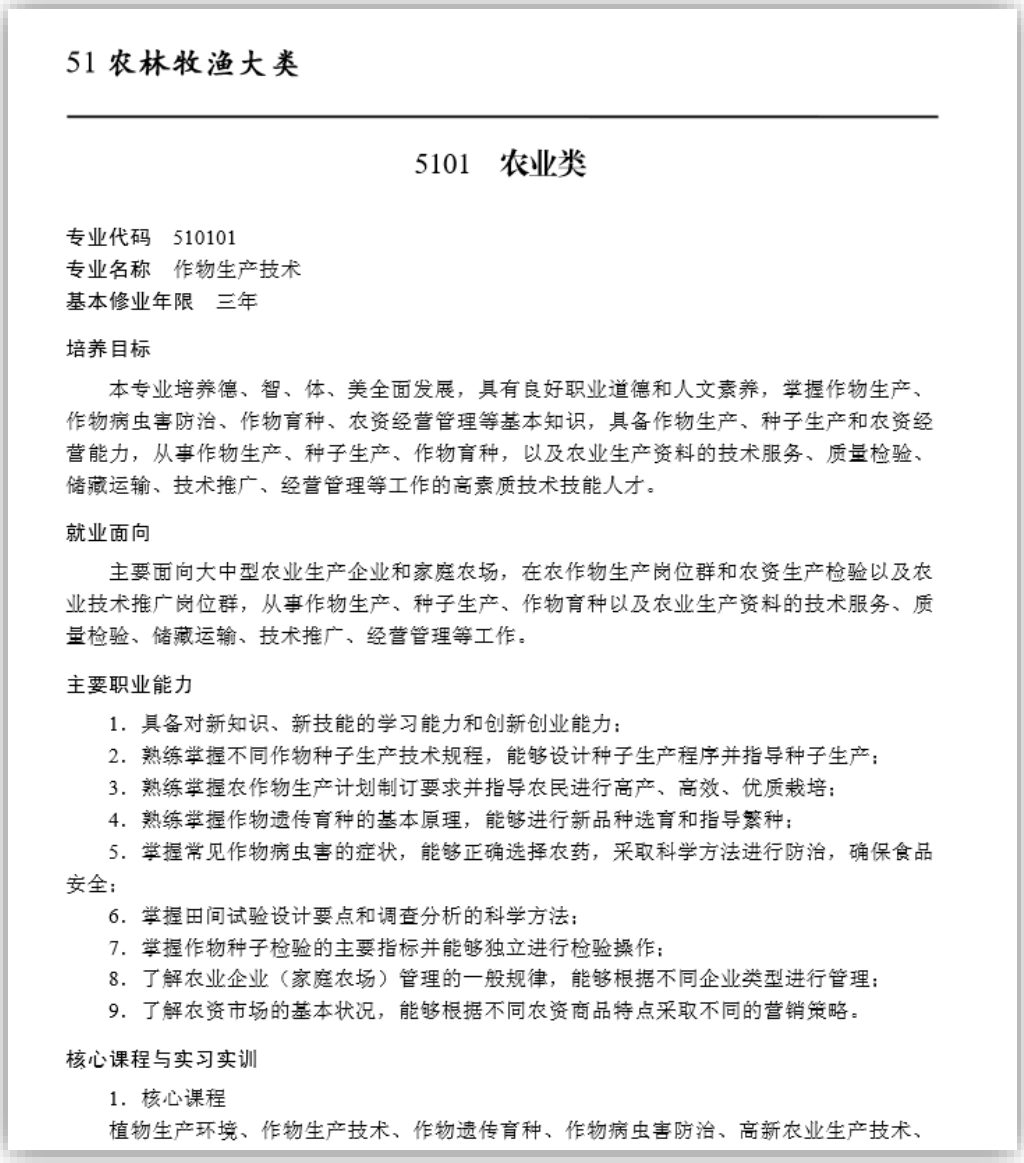
其次，对信息源进行分析与组织，确定待抽取信息源集合Σ，从而完成信息源初步筛选组织过程。

为获取较为准确的专业信息名录，根据教育部下发的相关文件，我们可以获取到另一种数据源。这类数据源属于格式较松散的非结构化数据。为此，我们需要进行数据预处理，一方面将松散的非结构化数据进行初步粗格式化，另一方面

为后序精处理做准备。

图 1.1 数据源-全国职业院校专业设置管理与公共信息服务平台

图 1.2 数据源-中国教育在线高职院校信息库



51 农林牧渔大类

5101 农业类

专业代码 510101
专业名称 作物生产技术
基本修业年限 三年

培养目标

本专业培养德、智、体、美全面发展，具有良好职业道德和人文素养，掌握作物生产、作物病虫害防治、作物育种、农资经营管理等基本知识，具备作物生产、种子生产和农资经营能力，从事作物生产、种子生产、作物育种，以及农业生产资料的技术服务、质量检验、储藏运输、技术推广、经营管理等工作的高素质技术技能人才。

就业面向

主要面向大中型农业生产企业和家庭农场，在农作物生产岗位群和农资生产检验以及农业技术推广岗位群，从事作物生产、种子生产、作物育种以及农业生产资料的技术服务、质量检验、储藏运输、技术推广、经营管理等工作。

主要职业能力

1．具备对新知识、新技能的学习能力和创新创业能力；
2．熟练掌握不同作物种子生产技术规程，能够设计种子生产程序并指导种子生产；
3．熟练掌握农作物生产计划制订要求并指导农民进行高产、高效、优质栽培；
4．熟练掌握作物遗传育种的基本原理，能够进行新品种选育和指导繁种；
5．掌握常见作物病虫害的症状，能够正确选择农药，采取科学方法进行防治，确保食品安全；
6．掌握田间试验设计要点和调查分析的科学方法；
7．掌握作物种子检验的主要指标并能够独立进行检验操作；
8．了解农业企业（家庭农场）管理的一般规律，能够根据不同企业类型进行管理；
9．了解农资市场的基本状况，能够根据不同农资商品特点采取不同的营销策略。

核心课程与实习实训

1．核心课程
植物生产环境、作物生产技术、作物遗传育种、作物病虫害防治、高新农业生产技术、

图 1.2 数据源-普通高等学校高等职业教育（专科）专业目录及专业简介

二、数据库设计

对于待存储的数据，我们设计如下图所示的数据库模式进行存储，其中：

1 院校信息与专业信息是多对多关系,因此设计院校-专业关系表进行关系存储。
2 对于高职专业与本科专业，存在一个高职专业，对应多个可能的本科专业，属于多对多关系。在本系统中，当前我们不需要关注本科专业信息表，因此只设计一个高职-本科专业关系表进行存储即可。
3 每个高职专业都具有专业大类代码，通常大类代码不会变更，因此对专业大类进行抽象，形成独立的专业大类实体。

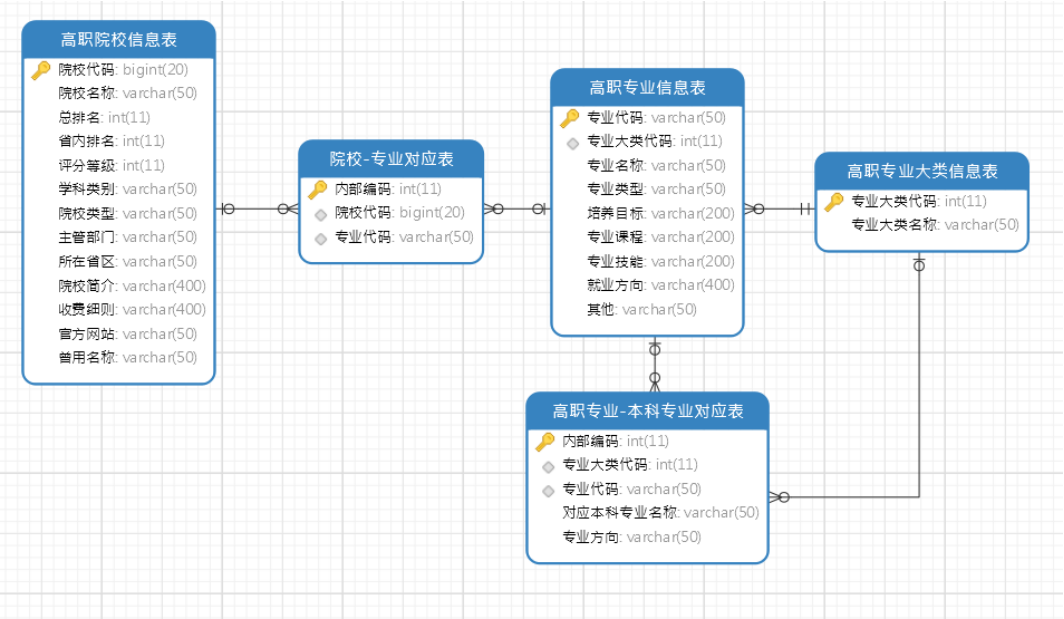


图 2　数据库 E-R 图

三、数据抽取、重组与结构化整理入库

对于上述两种不同类型数据源，我们分别采用不同的方式进行处理。

1　结构化数据抽取：

使用 jsoup 对 HTML DOM 进行解析与遍历，使用 getElementById()方法遍历一个 Document 对象，并使用 Element.select 抽取特定标签/类/id 的数据，最后在遍历过程中将数据进行组织，从而实现结构化数据的抽取。

2　非结构化数据抽取：

使用朴素的格式匹配方法，对此前预处理后的文件进行特征提取，将各条目信息注入 java bean 中，对 bean 进行组织，最后使用序列化方法对其进行格式化。

3　数据库部署：

对上述两类数据源处理后，我们需要对所得结构化数据进行入库操作。这里我们使用 JDBC 进行实现。对于 json 类型数据，使用 Fastjson 类库对其进行 json 转 bean，对每个 bean 进行数据库 SQL 注入。对于非结构化数据，我们已经将其实例化为各个 bean，因此直接进行 SQL 注入即可。

四、前端展示与测试

我们选用 ASP 对数据进行前端展示。选用 ASP 的原因大抵有二，一方面，ASP Script 使用 VB，属于解释型语言，易于上手与实现。另一方面，对于整个

系统的部署来说，使用 ASP 与原系统相符合，便于未来实现各部件的组合。

此外，我们使用当前较为流行的 Bootstrap 前端开发框架，便于对系统的进一步开发，以及响应式设计的实现。

图 4.1 系统主界面

前端展示方面，根据数据库，我们可以设计出四个子功能，分别为院校信息查询、专业信息查询、高职对应本科专业查询以及院校排名查询，从而基本实现对于高职高专相关招考信息的展示。

图 4.2 院校查询结果列表

图 4.3 院校详细信息

图 4.4 专业详细信息

图 4.5 高职对应本科专业信息查询结果

取得成果：软件著作权一份

5. 实习收获与能力提高情况

本次专业实习中，我遇到了一些问题。大抵来说分为数据抽取问题与组织部署问题。

数据抽取方面，首先是对目标地址的分析过程，即如何获取到全体待爬取信息源。通过对网页源代码逆向分析，我们可以了解到网页表单提交字段的意义，从而实现数据源集合的获取。在这个过程中，我惊喜地发现了中国教育在线的院校数据 json 文件，因此之后的工作就是对 json 进行解析与入库即可，这大大地提升了数据抽取的效率。

此外，对于动态页面如何使用 jsoup 进行遍历与数据提取也是一个问题。最终的解决方案是对动态页面静态化，即先对目标动态页面进行获取，此时完成了动态页面的静态化。之后，使用 jsoup 对本地 HTML DOM 进行遍历与提取，这样就实现了动态页面的抽取工作。

数据组织部署方面，由于粗处理后的数据源仍然存在一些小问题，这时，我们可以使用正则表达式对其进行精处理，如实现对空行等特征元素的批量处理，最终得到标准 json 文件，便于此后对其进行入库处理。

通过对以上问题的解决，我对数据抽取与组织有了进一步的认识与了解，也提升了我对问题解决的能力。

参考资料：

[1] 全国职业院校专业设置管理与公共信息服务平台-首页
https://www.zyyxzy.cn/index.shtml
[2] 普通高等学校高等职业教育（专科）专业目录及专业简介（截至 2018 年） - 中华人民共和国教育部政府门户网站
http://www.moe.gov.cn/s78/A07/zcs_ztzl/2017_zt06/17zt06_bznr/bznr_ptgxgdzjml/
[3] 中国教育在线-高职院校信息库
http://gaokao.eol.cn/gaozhi/
[4] jsoup 开发指南,jsoup 中文使用手册,jsoup 中文文档
http://www.open-open.com/jsoup/
[5] ASP 教程 | 菜鸟教程
http://www.runoob.com/asp/asp-tutorial.html
[6] 正则表达式 – 语法 | 菜鸟教程
http://www.runoob.com/regexp/regexp-syntax.html
[7] Fastjson 简明教程 | 菜鸟教程
http://www.runoob.com/w3cnote/fastjson-intro.html
[8] 柏玉. 面向网络数据的信息抽取研究与应用[D].西南交通大学,2015.