附件

编号：_____

“国家级大学生创新创业训练计划”

创新训练项目申请书

项目名称：基于网络爬虫和数据分析的高校信息整合系统

所属行业：计算机科学技术

起止时间：2017 年 5 月至 2018 年 9 月

团队名称：Cyber-Tech 小组

负责人姓名：李锦瑢

所在学院专业班级：计算机学院计算机科学与技术 10011501

联系电话：18392758151

指导教师：史豪斌

联系电话：13891937822

填 写 说 明

1、本申请书所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要。

2、申请人可以是个人，也可为创新团队，首页只填负责人。“项目编号”一栏不填。

3、本申请书为大16开本（A4），左侧装订成册。可网上下载、自行复印或加页，但格式、内容、大小均须与原件一致。

4、负责人所在学院认真审核，经初评和答辩，签署意见后，将申请书（一式两份）报送西北工业大学教务处。

一、 基本情况

<table>
<tr><td>项目名称</td><td colspan="7">基于网络爬虫和数据分析的高校信息整合系统</td></tr>
<tr><td>所属学科</td><td colspan="7">计算机科学技术</td></tr>
<tr><td>申请金额</td><td colspan="2">20,000 元</td><td>起止年月</td><td colspan="4">2017 年 5 月至 2018 年 9 月</td></tr>
<tr><td>负责人姓名</td><td>李锦瑢</td><td>性别</td><td>女</td><td>民族</td><td>汉</td><td>出生年月</td><td>1997 年 3 月</td></tr>
<tr><td>学号</td><td>2015302314</td><td>联系电话</td><td colspan="4">宅：　　手机:18392758151</td></tr>
<tr><td>指导教师</td><td>史豪斌</td><td>联系电话</td><td colspan="4">宅：　　手机:13891937822</td></tr>
<tr><td colspan="2">负责人曾经参与科研的情况</td><td colspan="6">在足球机器人创新基地工作
2016 年 FIRA 足球机器人仿真世界杯比赛</td></tr>
<tr><td colspan="2">指导教师承担科研课题情况</td><td colspan="6">史豪斌老师主持航空科技创新基金项目 1 项；参与 4 项国家部委决策支持课题项目、1 项陕西省自然科学基金项目、2 项航空基金项目、1 项航空支撑基金项目；参与国家 863 项目“黑客监控系统”、“网络安全协同防卫”、“操作系统逆向分析”等科研课题 10 余项。</td></tr>
<tr><td colspan="2">指导教师对本项目的支持情况</td><td colspan="6">指导教师将为项目组提供理论学习方向，并给与工程实践中的指导。此外，还将监督项目组及时完成项目进度。</td></tr>
<tr><td rowspan="6">项目组主要成员</td><td>姓　名</td><td>学　号</td><td colspan="2">专业/班级</td><td>所在学院</td><td>项目中的分工</td></tr>
<tr><td>李锦瑢</td><td>2015302314</td><td colspan="2">计算机科学与技术/10011501</td><td>计算机学院</td><td>网络爬虫和 Android 开发</td></tr>
<tr><td>白向龙</td><td>2015302020</td><td colspan="2">自动化 /09011501</td><td>自动化学院</td><td>数据挖掘和数据分析</td></tr>
<tr><td>殷康龙</td><td>2015303119</td><td colspan="2">软件工程 /14011502</td><td>软件与微电子学院</td><td>服务器建设、创建数据库、Android 开发</td></tr>
<tr><td>薛建业</td><td>2015300067</td><td colspan="2">机械电子专业 /HC001505</td><td>教育实验学院</td><td>服务器建设</td></tr>
<tr><td>董金国</td><td>2016303242</td><td colspan="2">软件工程 /14011606</td><td>软件与微电子学院</td><td>网络爬虫</td></tr>
</table>

二、 立项依据（可加页）

（一） 研究目的

计算机和互联网的出现与发展推动了信息的传播，互联网具有信息量大、搜索快速、方便交流和传播广泛等优势,深刻地影响着人们的学习、工作、生活方式和认知方式，人们不再依赖于纸质的信息传播方式，通过互联网多媒体获得时下最新的消息成为了人们的首选，面对海量信息，人们更加注重获取信息的体验——通过便捷、智能的方式来获取自己感兴趣的内容。

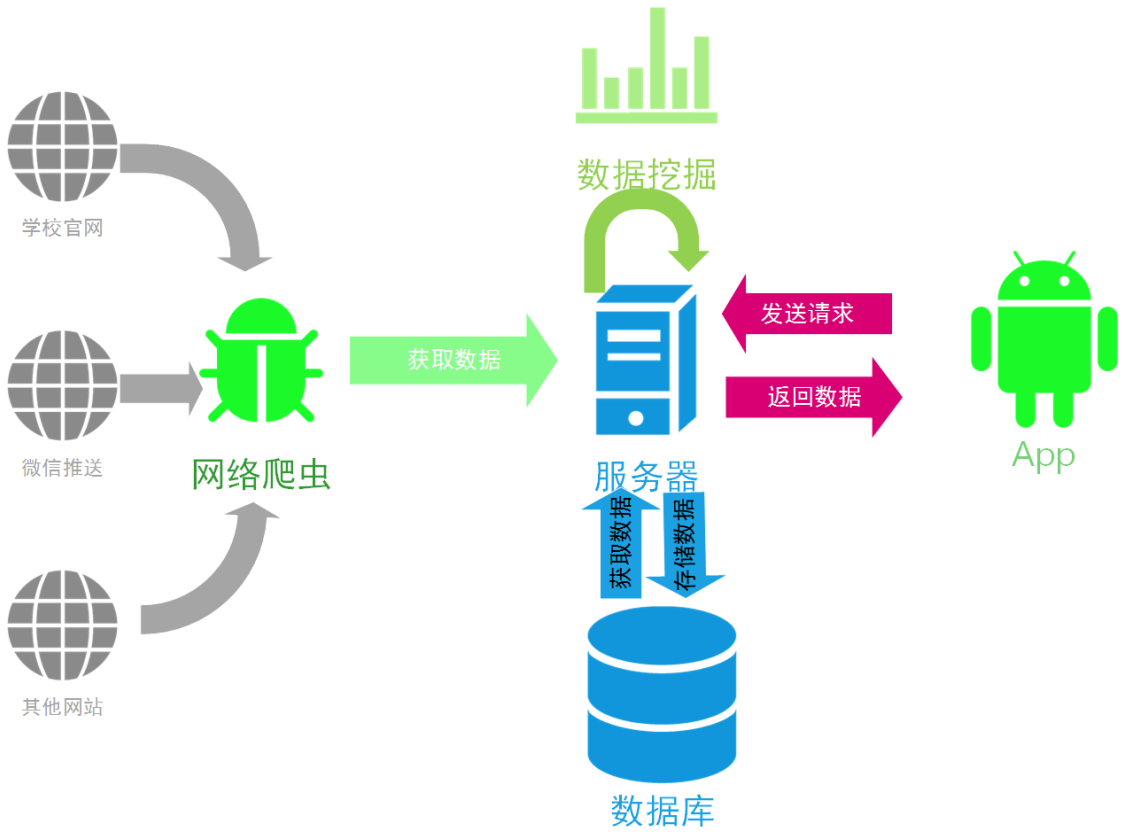
在学校，学生受众关注的校园信息发布途径多，主要有校园官方网站、微信公众号、校园论坛等，这些发布途径相对独立，限制了平台信息的关联，不利于信息的分类整合。例如，与校园有关的微信公众号繁多且各公众号平台之间相互独立，微信无法将各公众号的信息按照校园活动、学科竞赛、考研、就业等版块进行分类整理，学生要想获取信息就必须依次打开每一个公众号，查看内容，这降低了信息获取效率。

因此，我们希望通过对大量信息采集并进行数据挖掘和分析的方式搭建一个智能友好的信息整合平台，该信息整合平台不再注重功能的多而全，而是针对信息发布这项功能，提供全面的校园信息整合分类与数据统计分析，并且保持用户界面的友好和简洁。信息发布者，例如微信公众号主体、招聘网站等，也可以通过该平台反馈的数据统计等内容，进一步调整优化自己的信息资源，更好地经营自己的平台。用户可以通过该平台的分类版块、个性推荐等功能，有针对性地获取校园活动、学科竞赛、考研、就业等方面的信息，既节约搜索信息的时间，提高信息获取效率，又增强用户获取信息的体验。

（二） 研究内容

本项目旨在利用网络爬虫获取校内不同来源的信息，通过数据分析和数据挖掘，对信息数据进行分析处理，结合智能分析的相关算法实现信息整合和智能呈现，提供给用户一个友好简洁的平台。同时利用统计图形展示数据统计结果，提供给数据发布者更好的决策方案。由于当下最流行的网络终端是 Android 移动终端，所以我们最终将在 Android 手机上搭建平台 。

项目中的校园信息平台的对信息的处理主要有四个步骤，依次是：信息采集、信息存储、信息分析、信息呈现。我们研究的核心是如何高效地采集数据，如何通过数据挖掘和数据分析寻找校园信息中有价值的信息，以及最终实现与用户的智能交互和数据反馈。



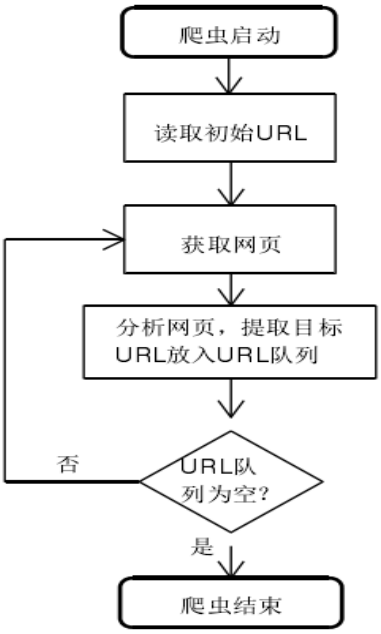
图片 1 校园信息系统工作图

本项目计划研究以下五个方面，从下面的五个方面中，逐渐实现校园信息系统。

1. 网络爬虫技术

网络爬虫又称为网络蜘蛛或网络机器人，是一个自动提取网页的计算机程序，它是构建搜索引擎不可或缺的重要组成部分。

网络爬虫是以若干个 URL(统一资源定位符)为爬取起点，在下载网页的同时又不断地从中提取出新的 URL，如此循环，直至没有 URL 需要下载为止。以爬取学校官网为例，在编写好爬虫代码后，启动爬虫程序，爬虫会先读取初始 URL 集，并将读取的结果放入“待爬取 URL 队列”中。每次爬虫会从 URL 队列中获取队头 URL，然后发出 HTTP 请求下载该 URL 所对应的校园网页。页面下载成功后，一方面我们将利用文本检索和文本分析等方法对网页内容进行过滤和提取，生成类型、标题、简介、发布日期等数据，分类存储在数据库中并建立索引，以便之后的查询和检索；另一方面，我们将从网页中提取出若干个新 URL 放入“待爬取的 URL 队列”，然后，继续从 URL 队列中获取下一个 URL，如此循环，直至“待爬取 URL 队列”为空时，爬虫程序停止。



图片 2 网络爬虫的流程

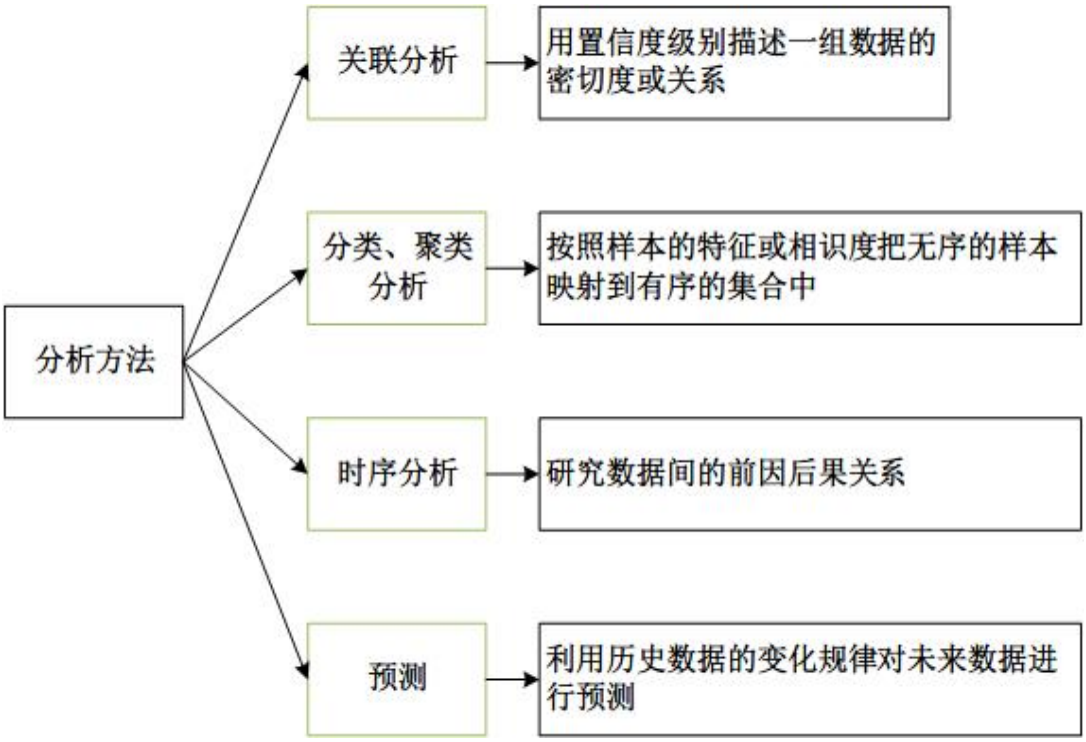
2. 数据分析和数据挖掘

大数据分析的热潮，为信息的发布和展示提供更好的决策方案。数据分析（Data Analysis）是一个检查、清理、转换和建模数据的过程，目的是发现有用的信息，得出结论和推动决策制定。数据挖掘（Data Mining）是指从大量的数据中通过算法搜索隐藏于信息中的价值的过程。

以往的校园信息平台无法根据用户的喜好和文章的特点实现智能化的信息推荐，用户需要逐条阅读文章才能找到自己需要的信息，这样不但会浪费用户的大量时间，而且容易遗漏重要信息。通过信息采集过程，我们会存储大量的信息数据，包括文本、数字、图形图像等。我们将结合统计分析中知识，对数据之间的关系进行建模分析，挖掘其中隐含的联系，寻找对用户和信息发布者有用的数据信息。最终在移动端应用中使用分析的结果，并通过图形化的方式展现分析的结果。

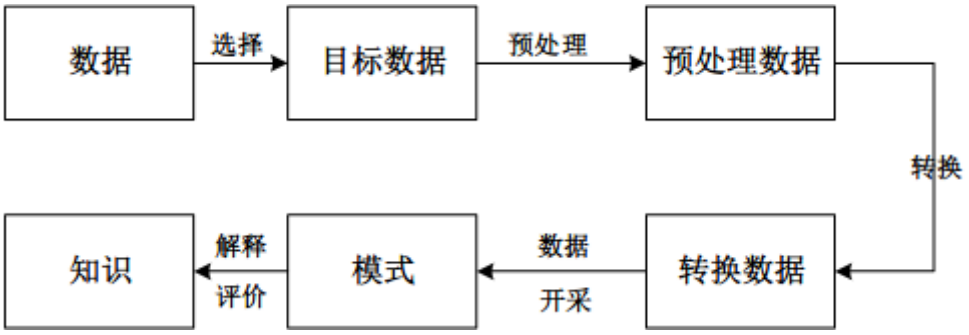
数据挖掘还可以通过对信息发布者创作的内容和读者阅读行为的记录和分析得出信息发布者发布的信息受欢迎程度。找出信息发布者的发文主题是否符合当下热度、内容是否吸引用户等方面优势或不足，将这些信息反馈给信息发布者帮助他们更好的创作用户感兴趣的文章。

数据挖掘分析方法可以分为四类：关联分析、时序分析、分类聚类分析和预测。常用的数据挖掘技术主要有统计分析和一些成功的人工智能技术。常使用的统计模型有线性分析和非线性分析、连续回归分析和逻辑回归分析等；人工智能技术包括神经网络、决策树等智能算法。



图片 3 数据挖掘和数据分析的方法

数据挖掘的过程如下：确定任务的目标数据，包括消除噪声、推导计算缺值数据、消除重复记录、数据类型转换等的数据预处理，通过消减数据维数转换数据、确定开采任务并进行数据开采，经过用户和机器的评价结果，剔除冗余和无关的内容，使用户更容易理解和应用，最后用可视化等方式展示数据挖掘的结果。



图片 4 数据挖掘和数据分析的流程

3. 搭建信息系统服务器

服务器用来向外界提供各种服务，如数据库服务、web 服务等。它能够响应来自其他计算机的请求，并进行处理，实现一定的业务逻辑。在信息管理系统中，服务器作为信息周转的中心，负责协调网络爬虫、数据库访问、服务器脚本运行、响应 Android 客户端的请求等任务，是系统核心功能的载体。

在信息管理系统的服务器上，建立相应的 web 服务，使其成为信息流中的核心节点。服务器能够满足和协调多个用户的使用，能够稳定流畅的运行爬虫程序、数据库和服务器脚本。

4. 关系数据库建模和实现

数据库（Database）是用来存储数据的仓库。它按照一定的数据模型组织和存储数据，有较小的冗余度、较高的数据独立性和易扩展性，使得数据能够得到统一的管理和控制。计算机技术技术越来越广泛地应用到社会生活的各个领域，越来越多的信息需要存储和管理，数据库是有效的管理信息的工具，能够快速的存储、查询信息，成为了计算机的核心技术。而关系数据库提供了非常优秀的数据管理能力。

在实现网络爬虫过程和大数据分析中，涉及到的数据量很大，需要快速高效地访问、存储大量数据，必须建立一个稳定安全的数据库用来存储信息。数据模型必须能够较好的对应现实生活中信息的结构、信息的流动情况、信息之间相互制约的关系。并且在实践过程中能够跟随用户的需求变化进行不断地调整扩充。

5. 移动端界面开发

Android 是当前中国手机市场的主流系统，它基于 Linux 系统采用四层架构 ，代码开源，提供了丰富的接口，使得 Android 应用发展迅速，应用繁多，已经超越 Windows 成为世界上应用最广泛的系统。

我们最终的目的是在移动端实现信息的发布功能，并且能够提供优良的交互性能，所以首选 Android 系统作为应用开发的平台。利用 Android 系统已有的优秀框架和资源库，搭建一个具有智能化服务和分析的信息整合平台。

（三） 国、内外研究现状和发展动态

当下是自媒体多媒体盛行的时代，每个人都会通过各种渠道获取有价值的信息。随着网络规模和用户数量的迅速扩大，应用信息系统的不断丰富，高校网络环境已经出现了翻天覆地的转变。传统校园网络平台在进行设计的过程中，只是依照计算机技术要求及学生需求进行设计，没有对平台性能、平台质量进行全方位控制和提升，在很大程度上制约了主体的建设质量。因此，在现代平台建设的过程中，设计人员要选取高性能、高质量数据体系，要对新型 android 平台的校园信息系统 app 进行深入探究，从本质上提升校园信息交流的移动便捷性，安全性和可靠性。

由于互联网信息发展迅速，一个服务型软件想要做的大而全越来越难，这样的服务软件不仅不会给我们带来便利，反而会让用户觉得信息杂乱繁多，大大降低用户的体验。因为信息发布的平台逐渐增多，各个平台之间的资源相对独立，在生活中信息十分分散，不利于用户在最短的时间内获取到有价值的信息。

以微信公众号为代表的自媒体数据数总量巨大，数据变化非常快，数据库中的数据每天增加超过 500TB。信息挖掘技术致力于从海量数据中过滤对用户有用的信息，然后将这些过滤结果返回给用户，但是这些结果间存在的内在关联并没有被很好地挖掘和解析出来，缺乏从语义角度去挖掘深层次的规律和知识的能力，用户只能从结果中自己理解和筛选知识。所以目前引擎急需从基于关键字或者基

于文本内容检索这一浅层次的知识理解和挖掘工作，向表达、理解语义和关系这些深层次的知识挖掘方面发展。

（四） 创新点与项目特色

1. 通过网络爬虫获取数据

“我们不是信息的制造者，我们只是信息的搬运工。” 以前的信息发布平台经常会考虑如何让信息的创造者舍弃原来的平台而选择新的平台，最终因为信息发布的各方无法参与其中，最终宣告失败。我们的想法是在信息发布者允许的权限范围内进行信息的采集工作，并且对采集到的大量数据进行分析，挖掘其中有价值的信息，最终通过一种智能的方式，发布信息。

我们不需要信息发布者在不同的平台进行重复的消息发布，减少了信息的冗余情况，能够大大减少信息发布者的工作，也能有效的避免信息的遗漏情况。

当下，获取校园信息的途径和方式繁多，加大了用户在不同平台之间转换的时间成本。我们通过网络爬虫的方式，将信息整合在一起，大大节省了用户在不同平台之间转化的成本。利用爬虫技术爬取高校的官网、微信平台以及其他相关网站的信息，对热门话题，例如竞赛、考研、就业等进行整理，在该平台进行发布，具有信息全面、覆盖率广的特点。

2. 基于数据挖掘和数据分析的校园信息处理和分析

我们会从网络中获取大量信息，首先要使用信息筛选算法去掉其中重复冗余的内容，将存在潜在价值的信息进行存储。这些数据之间蕴含着丰富的复杂关联，有效利用数据分析的方法，去冗分类、去粗取精，从数据中挖掘知识，对数据网络后面的知识进行深入分析。

我们打算对用户的操作事件和信息的文本内容进行数据分析，最后将数据分析的结果结合统计学的相关知识，以一种形象直观的方式展现数据。这能让用户和信息发布者了解到当前的信息发布趋势和信息之间的关联，促进用户更好的浏览和帮助信息发布者更加高效便捷的发布信息。

3. 信息的智能分类和智能推荐

用户在浏览信息的过程中，希望能够更加方便快捷的实现信息的获取。我们通过分析用户浏览文章时的偏好，结合数据挖掘和分析的结果，对相关的文本信息进行有针对的推送。

我们会结合数据分析和数据挖掘相关的知识，挖掘数据之间的关联性，不仅仅提供一个信息的呈现界面，也要通过更加复杂和只能的方式对文本信息进行分类。例如：通过文章的文本分析，我们能够寻找当下信息发布的热门话题，并将其中的热门话题与相应的文章进行关联，最终形成一份热门推荐的标签云。

4. 移动端数据呈现

移动端不仅仅提供智能化的信息分类与推荐功能，还将把数据挖掘和数据分析

的结果通过图形化的方式展示，结合统计分析中的饼图、柱状图、折线图等数据呈现方式，使信息和数据更直观的呈现出来。用户和信息发布者都能够快速得知当前的热门话题、各个发布者之间的信息等。

（五） 技术路线、拟解决的问题及预期成果

技术路线

1. 构建网络爬虫

- 基于 scrapy 框架构建网络爬虫

我们拟选用的爬虫构建框架为 Scrapy 框架。在 Scrapy 框架中，数据在各个模块间的流动是由 Scrapy 引擎负责控制的，其具体过程如下：

（1）Scrapy 引擎打开，向爬虫模块请求其需要爬取的起始 URLs

（2）Scrapy 引擎从爬虫模块那里获取到需要爬取的起始 URLs，将它们放入调度器模块，作为待爬取的 URL Requests；

（3）Scrapy 引擎向调度器模块发出请求，获取下一个待爬取页面的 URL；

（4）调度器模块返回下一个待爬取页面的 URL 给 Scrapy 引擎，Scrapy 引擎则将该 URL 经下载中间件模块（请求方向）转发给下载器；

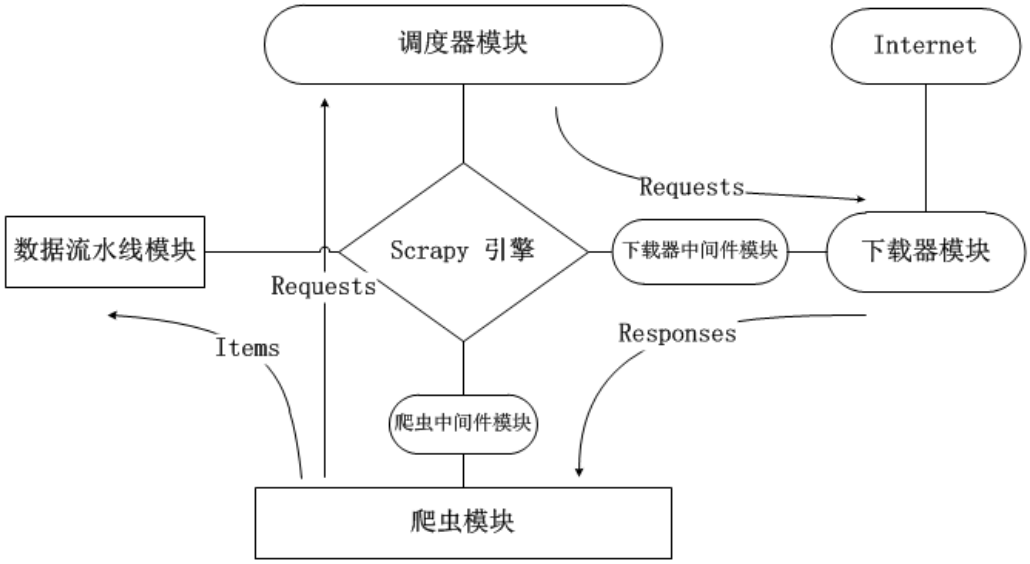
（5）当下载器模块完成页面下载后，会生成一个该页面所对应的 HTTP Response，并将其经下载中间件（响应方向）发送给 Scrapy 引擎；

（6）Scrapy 引擎从下载器模块处接收到 HTTP Response 后，会将其经爬虫中间件模块（输入方向）发送给爬虫模块处理；

（7）爬虫模块处理接收到的 HTTP Response，并返回从中爬取到的 Items 及需要跟进的新的 URLs 给 Scrapy 引擎；

（8）Scrapy 引擎将（爬虫模块返回的）Items 转发给数据流水线模块，同时将（爬虫模块返回的）URLs 转发给调度器；

（9）以上 2~8 的过程会一直重复执行，直到调度器模块中没有 URL Requests 时，Scrapy 引擎关闭，爬虫停止。



图片 5　Scrapy 框架架构图

- URL 去重

网络爬虫在爬行时，每秒下载几十个乃至上百个页面，解析出的 URL 大多都是重复的，实际有效的新 URL 才是下一步要爬取的目标，如果不将重复的 URL 去掉，就会导致网络爬虫在爬取网页时会将相同的页面重复地下载，再加上 Python 本身是脚本语言，其对象占用内存往往比 C/C++等编译型语言要大得多，而且 Python 垃圾收集器的释放内存算法并不会在对象不再被引用时立即释放内存，这将会大大地消耗系统资源甚至造成系统崩溃。

为有效地实现 URL 的查重操作，我们需要先将所有的 URL 保存下来，然后通过比较知道它是不是在集合内。我们团队通过查阅相关资料和讨论，制定了三种方案。

方案一：为了尽快把整个爬虫搭建起来，最直观的方法就是建立一个内存中的 HashSet，在 HashSet 中放置 URL 字符串。任何一个新的 URL 首先在 HashSet 中进行查找，如果 HashSet 中没有，就将新的 URL 插入 HashSet，并将 URL 放入待抓取队列。这个方案的好处是它的去重效果精确，不会漏过一个重复的 URL。但它的缺点是随着抓取网页的增加，HashSet 会一直无限制的增长，最终会超出内存而导致程序无法运行。

方案二：在方案一的基础上增加一个小功能，即 HashSet 中不存储原始的 URL，而是将 URL 压缩后再放进去，采用消息摘要算法对 URL 做编码，消息摘要算法的结果是 128 bit，也就是 16 byte 的长度。相比于估计的 URL 平均长度 100byte 已经缩小了好几倍，但随着 URL 的增多，最终还是出现内存不足的问题。所以，这个方案也不能解决本质问题。

方案三：采用布隆过滤器进行去重。布隆过滤器采用的是哈希函数的方法，

是一种空间效率很高的随机数据结构，它利用位数组表示一个集合，并能判断一个元素是否属于这个集合。它的空间复杂度是固定的常数 O(m)，而检索时间复杂度是固定的常数 O(k)。相比而言，有 1%误报率和最优值 k 的布隆过滤器，无论元素的大小，每个元素只需要 9.6 个比特，是相对于前两个方案的进一步优化。

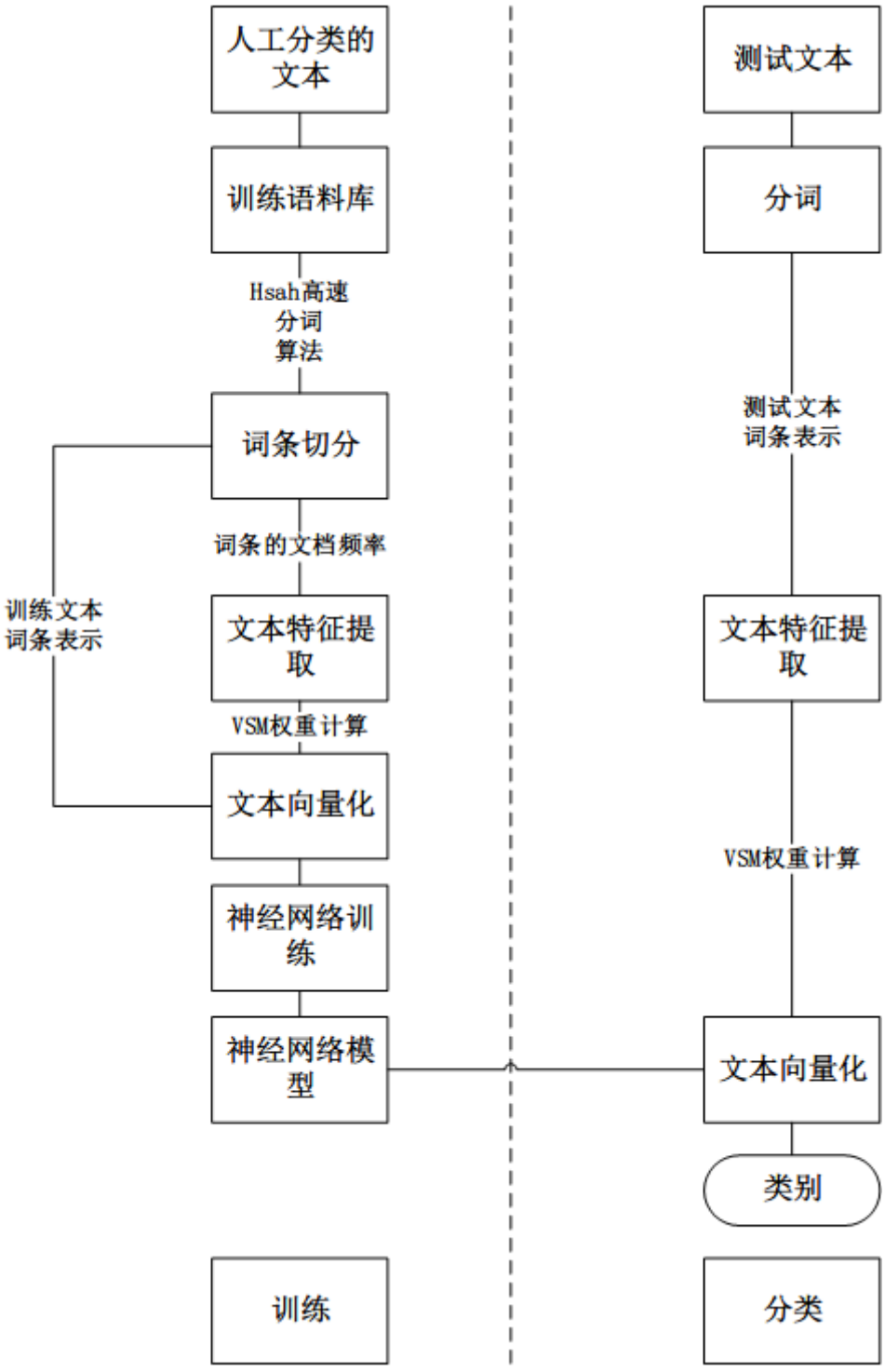
2. 数据分析和数据挖掘

- 基于神经网络的文本分类

文本分类是根据文本内容自动确定文本类别的过程，使用户更方便的浏览文档，同时通过限制搜索范围提高搜索效率。利用 BP 神经网络对数据的并行处理和自学习能力，构建基于 BP 神经网络的文本分类器。

BP 神经网络文本分类器分为训练和分类两部分。首先建立专用的分词词典；然后对经过人工分类的文本进行整理，形成训练语料库；利用分词词典对训练语料库和训练样本进行词条切分、文本特征提取、词频统计、文本向量化；使用训练样本得出的矩阵作为输入，用 BP 网络进行训练，达到满意效果后，得到固定的权值，作为分类知识存储在网络中。

分类器训练完毕后，就可以对测试样本进行分类了，分类过程和训练过程大体相似，首先利用分词词典对样本进行词条切分、词频统计、文本特征提取、文本向量化，生成待分类文档的特征向量，由于训练过的网络权值已经固定，可以直接运用它得出分类结果。



图片 6 文本分类与神经网络

实现文本分类后，就可以把获得的数据按照不同的归类方式分到各个类别中，例如通过主题分类将数据分到校务新闻、实习就业、竞赛信息等主题中，通过来源分类将数据分到不同的来源中。文本话题分类后用户就可以浏览自己需要的某一类话题精确快捷地得到需要的信息，文本来源分类后用户就可以持续关注某个信息发布者的信息。在文档分析过程中获得热门词条标签云，例如中国航天日、翱翔系列卫星等，并将文章与这些标签关联起来，使用户通过热门标签云准确把

握当下最热门的话题和文章。

- 个性化推荐

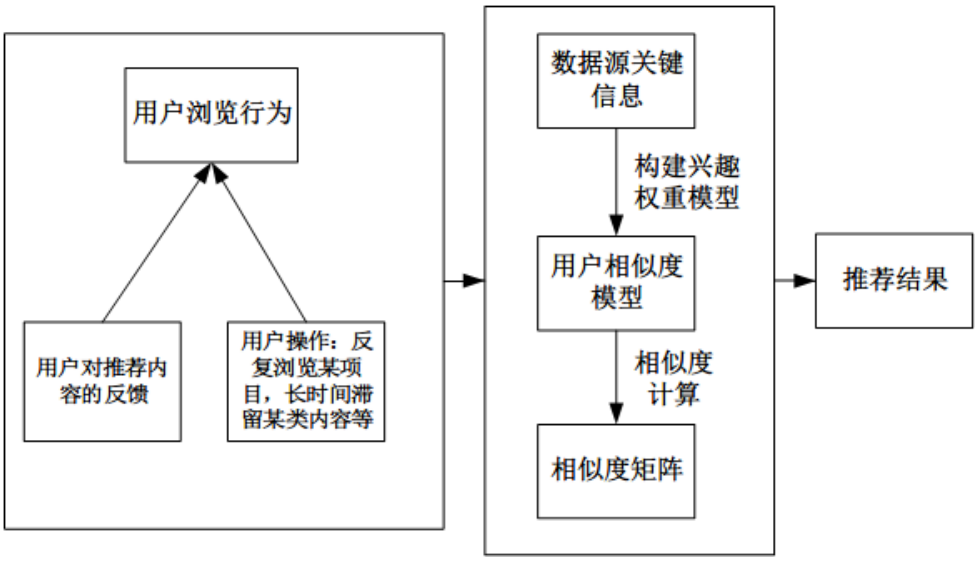
个性化是通过挖掘用户的历史记录和浏览记录，将这些记录与用户信息相结合，推荐出用户可能感兴趣的事物。这样既给用户节约时间搜索信息，又让用户体验到智慧推荐方便快捷的优点。

目前主要的推荐技术分为，协同过滤推荐，对用户做出相似比较，为相似度高的用户推荐相同的内容；基于内容推荐，在用户浏览操作记录的基础上推荐用户可能感兴趣的内容。本项目将协同过滤和基于内容过滤两种算法取长补短进行组合，设计出混合推荐算法。混合推荐算法的主要过程如下：

（1） 用户浏览行为分析设计：用户特征、浏览记录、历史操作等作为智能引擎推荐的基础。

（2） 分析数据，构建用户与内容的关联矩阵。用 KNN（K 最近邻）算法匹配用户最可能感兴趣的 K 个内容：用户的特征作为 KNN 算法的特征属性；设计相关度计算方法，构建相似度模型，本项目通过科学合理地设置个特征值的权重建立用户兴趣的权重模型；生成形似度矩阵。

（3） 过相似度矩阵得到推荐结果。



图片 7 用户推荐与相似矩阵

个性化推荐将自动为用户筛选感兴趣的话题，例如为热衷科研竞赛的用户推送竞赛的详细信息，为毕业生提供实习招聘信息。个性化推荐是综合推荐用户可能感兴趣的各类信息，并将这些信息发布到同一个界面中，用户不需要跳转界面就可以浏览自己需要和感兴趣的各类信息，可以是用户短时间内获得自己感兴趣的所有类型文章。

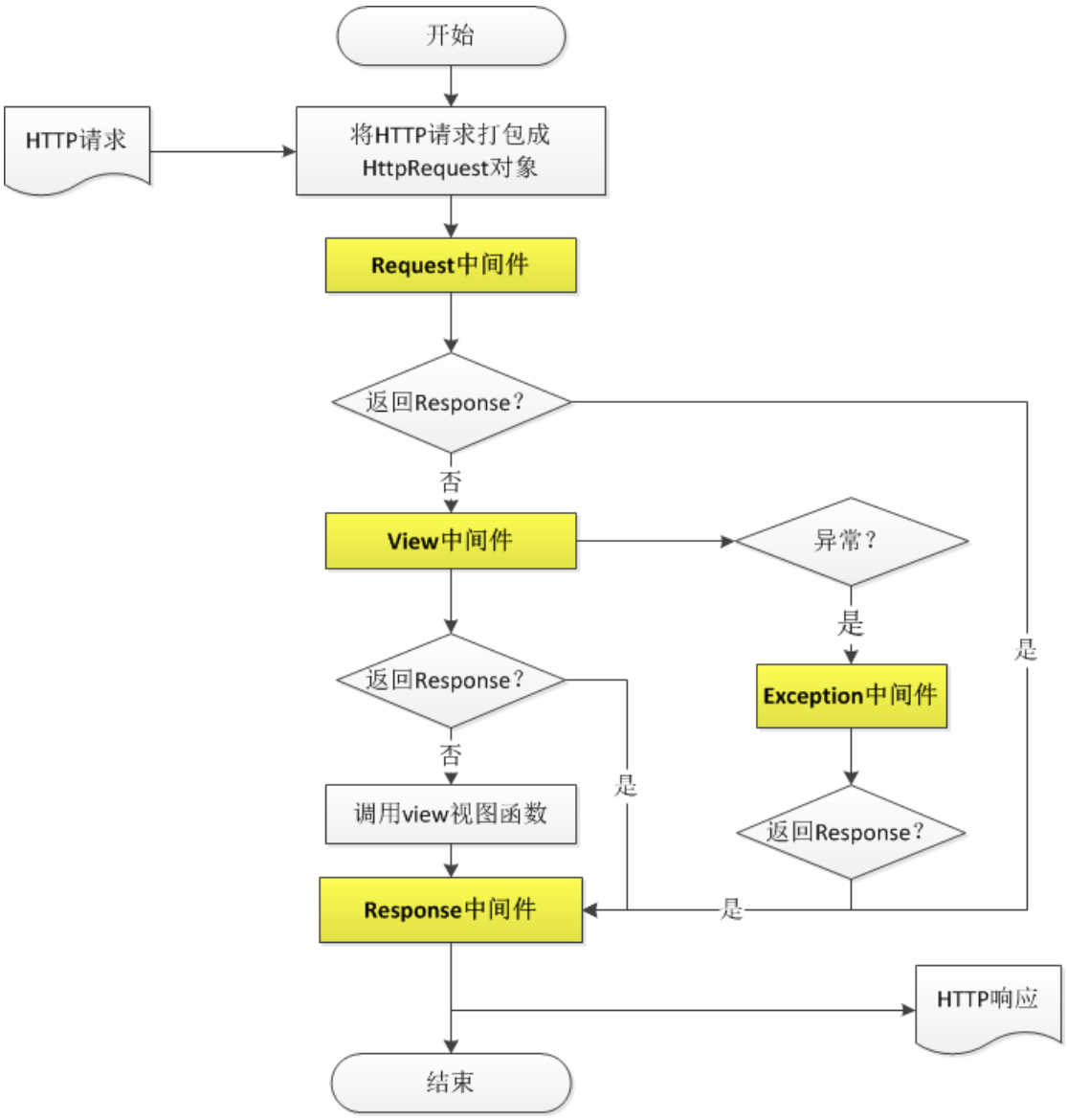
3. 搭建网络服务器

为了实现需要的功能，我们将在 Linux 系统下搭建 Apache 服务器。Apache 服

务器是一种通用的 web 服务器，能够响应 HTTP 请求，并且发送给指定的服务端程序进行处理，而且 Apache 服务器稳定高效，代码开源，开发成本较低。

我们使用 Python 作为服务器端的脚本语言，Python 语言精炼、开发效率高，并且可以通过简单的 API 扩展，将 Python 嵌入到 Apache 服务器当中。Python 有十分强大的函数库供我们使用，有利于开发的效率提升。Python 有许多性能较好的开发框架，我们将使用 Django，Django 是一个 Python 语言编写的开源 web 框架，遵循 MVC（model-view-control）的代码架构模式，对开发者友好，有详细的说明文档和完整的网络教程。

当我们做好 Android 客户端向服务器发送一个 http 请求时，服务器上的 Django 捕获请求内容，并进行分析，交给指定的中间件进行处理。然后从数据库中获取数据，封装成 json 格式，最后将响应的数据传递给 android 客户端。



图片 8 Django 框架响应 HTTP 请求过程

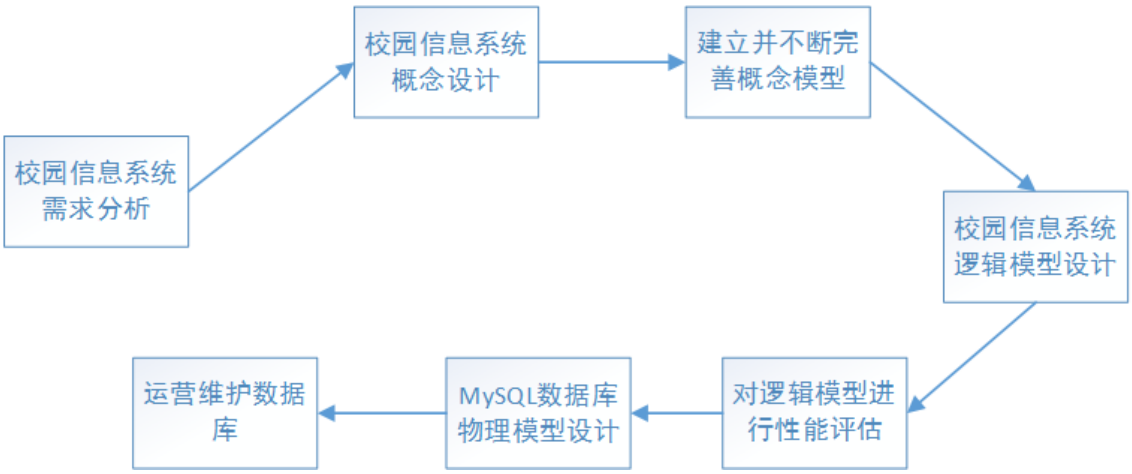
4. 进行需求分析创建数据库

我们拟采用 MySQL 数据库进行数据的存储和管理。MySQL 是 Oracle 公司旗下的关系型数据库管理系统，它体积小、速度快、成本低，而且代码开源，同 Apache 服务器具有良好的兼容性。

1）数据库设计的原则

在设计数据库时，遵循数据库规范化原则，使建立的数据库符合第三范式规范，便于数据库的管理和使用。根据校园信息系统的实现目标进行分析讨论，对关系数据库的实体和实体之间的关系进行抽象分析，得出一个满足数据需求的数据库，并且能够在实际开发过程中不断的修改和扩展。

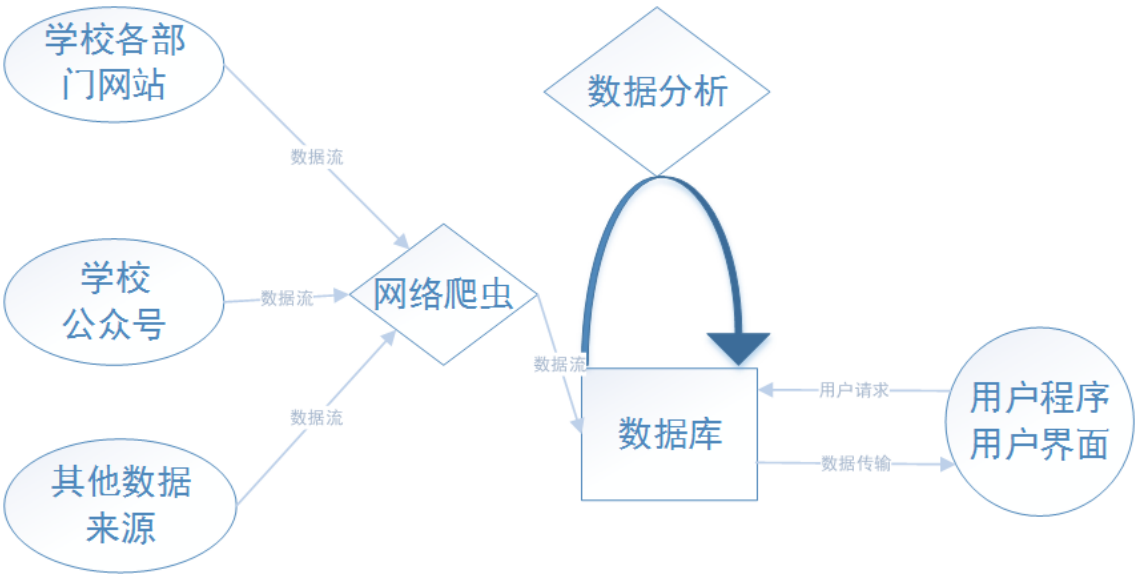
2）校园信息系统的数据库设计流程



图片 9 校园信息系统创建流程

3）用户需求分析阶段

- 数据流图



图片 10 校园信息系统数据流图

- 数据字典

序	相关的实体	实体的属性	唯一标识	补充

号				
1	文章	文章 id、发布者、发布时间、存储更新时间、图片封面、文章链接、文章简介、类别、标签	文章 id	存储特点：文章不断增加，在制定时间内刷新
2	用户	用户 id、用户名、头像、性别、电话、邮箱、密码、隶属学校、用户权限	用户 id	关联属性：隶属的学校对应学校的 id，数据获取来源：微信的授权登录
3	发布者	发布者 id、公众号或网站名称、公众号或网站的图片标识、简介、隶属学校	发布者 id	内容描述：发布者主要指校园各个部门的官网、各个社团的公众号等
4	类别	类别 id、类别名称	类别 id	
5	标签云	标签 id、标签名称、标签访问量	标签 id	
6	学校	学校 id、学校名称、学校简介	学校 id	

表格 1 初步需求分析的数据字典

- 数据之间联系的描述

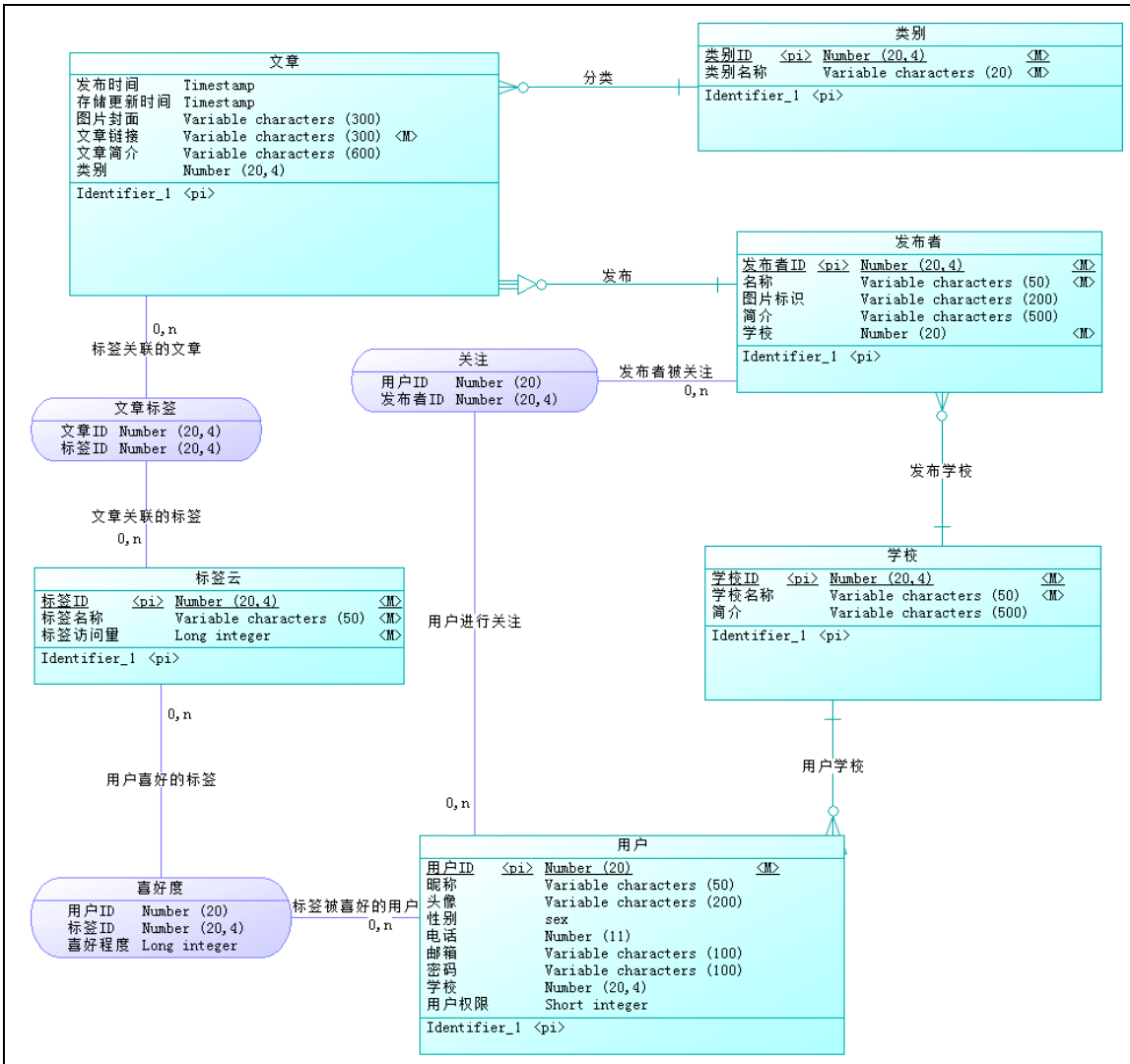
用户能够关注制定的校园公众号的推送

手机用户的浏览记录，通过浏览记录的分析能够生成用户对某类型或者标签的喜好。

每片文章可以对应一种类型和多个标签，每个标签也可以标记多篇文章

用户和发布者都隶属于不同的学校，每一篇文章都要对应制定的发布者主题

4）初步创建信息系统的关系型数据库概念模型



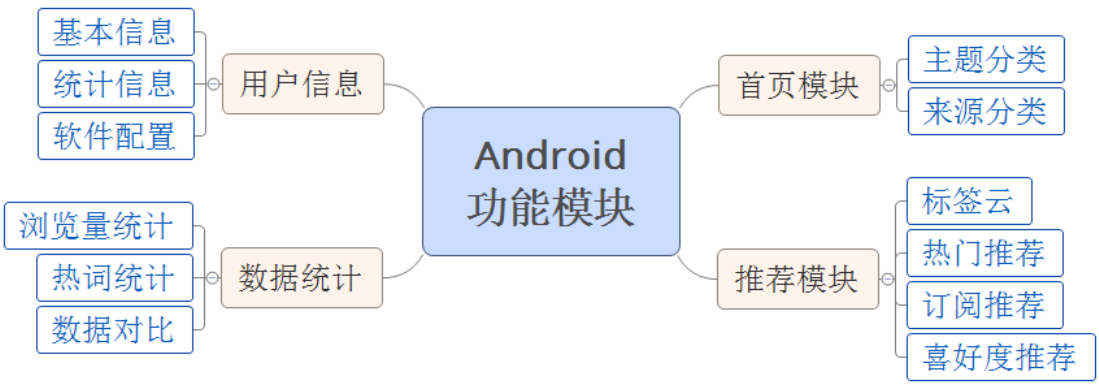
图片 11 关系数据库的概念模型

关系数据库的概念模型为移动客户端能够用到的数据，不够完善，仍旧需要根据数据分析和挖掘、网络爬虫等部分需求，不断丰富完善数据库概念模型。

5. 设计 android 开发界面

Android 中开发 app 的目的是希望能够更加友好的实现与用户的交互，让用户在浏览信息的过程中获得更好的体验。

1）Android 中实现的功能模块



图片 12 移动应用的功能模块

2）Android 开发实现

Android 的开发者提供了大量开源的类库实现网络通信、图片处理、界面布局的功能，并且有多种框架能够快捷便利的组织代码，完成开发任务。

在实现用户界面时，注重考虑用户的体验，使用一些比较先进的交互设计思想，Android 的界面开发有很多开源的界面库，如 GreenDroid、SlidingMenu、Cards 等，能够实现形形色色的 Android 的界面。

在实现 android 的逻辑功能时，使用 AndBase 或者 XUtil 框架，进行代码的组织。这样会节省自己大量的 android 开发时间，是的 android 程序的效能提升。Android 的代码框架封装了大量的 UI 控件，能够基本满足界面的设计需求；并且提供了数据库管理、图片管理的接口，提高了程序的执行效率；Android 框架提供了一整套 http 请求的工具，能够高效稳定地向服务器发送请求获取数据。

拟解决的问题

1. 爬虫相关的问题

部分网页对爬虫程序不友好，有反爬虫措施，例如：①记录和分析用户代理信息：每个爬虫在爬取网页的时候，会声明自己的用户代理（User Agent）信息，因此后台可以通过记录和分析用户代理信息来挖掘和封锁爬虫。如果单个 IP 访问超过阈值或者单个用户代理访问超过阈值，就予以封锁。②利用流量统计系统记录的 IP：在页面里面嵌入 JavaScript 的网站会通过流量统计系统记录的 IP 和服务器程序日志记录的 IP 地址进行比较，进而识别出网络爬虫，予以封锁。

2. 数据分析和挖掘中的问题

数据挖掘要处理大量的数据，加之为精确处理数据，会大大提高算法的复杂度，使得计算时间大大增加，不能很好的保证系统实时性。所以高质量的数据处理和系统实时性难以做到互利共存，往往需要降低一种来提高另一种。怎样让两者有机结合并互利共存是数据挖掘中需要解决的问题。

3. 数据库建模中的问题

在建立数据库模型的过程中，用户的需求会不断变更。我们想要实现的功能，不可能在一开始就规划完整，而每一部分需要的数据都会有所变化。由于信息采集来自不同平台，各个信息文本之间存在较大的差距，不能使用统一的数据表进行储存。

4. 用户界面设计中的问题

Android 应用程序在不同机型上运行情况差异较大。

针对以上难点，项目组拟采用以下方案解决：

1. 控制爬取频率。

典型的爬虫行为是过高的访问频率。正常的用户行为通常是每 5~30 秒访问一个网页，而爬虫程序在不加限制的情况下，往往能达到每秒爬取数十甚至数百

个网页。因此，通过访问频率即可很容易地识别出爬虫行为，从而将其禁用或是通过验证码来扰乱爬虫程序。为了能够控制爬虫的访问频率，考虑用代码实现一个计时器，用来记录上一次返回 URL 的时间。每次在输出下一个待爬取 URL 时，可以计算出当前时间距离上一次返回 URL 的时间有多少秒。如果该时间间隔小于预配置的最快爬取时间间隔，则等待一定的时间，再返回下一个待爬取 URL。

2. 为了解决数据挖掘在实际应用中的实时性问题，本项目将把数据挖掘拆分为在线部分和离线部分进行处理。对于计算较为复杂的部分先通过离线部分处理，例如：BP 神经网络文本分类器中分词处理、文本特征化、神经网络的训练以及 KNN 算法中用户对内容的兴趣权重模型、用户相似度估计等，至于计算结果和呈现则交予在线部分处理。采用离线方式处理一部分计算过程可以在很大程度上缩短在线部分计算所耗时间，有效提高整个智能推荐引擎的运行效率，让用户得到方便快捷的使用体验。

3. 需要跟随用户不断变更的需求，反复进行数据库的需求分析和建模，不断完善数据库。为了实现更好的存储功能，应该将不同类别的数据表继承总的数据表。并且完善数据库之间的关系，使数据库能够满足不断变化的数据需求。同时结合数据库的专业知识，对数据库进行规范化处理，达到 BCNF 范式的要求。

4. 需要在主流的 Android 机型上进行反复测试，修改相应的代码框架，使其最终能在各个 Android 设备上达到预期的效果。

预期成果

项目预期完成以下研究成果：

1. 实现网络爬虫采集信息，能够从官网、教务系统、微信公众号等平台获得对学生有价值的信息，并进行整理。

2. 通过数据挖掘和数据分析探索数据中的价值，给用户提供更有价值的信息，并给数据发布者提供更好的数据发布方案。

4. 搭建稳定的服务器并实现完善的关系数据库。

5. 实现 Android 客户端。在客户端对信息进行智能推送和智能分类，并通过统计学的方法展示数据分析和数据挖掘的结果。

7. 整理相关的技术文档，完成专利申请。

（六） 项目研究进度安排

本项目计划分为六个阶段：

第一阶段：分组学习相关的知识和技术。查阅网络爬虫、数据挖掘和大数据分析的论文并学习 Python 语言、Android 编程、建立数据库和搭建服务器的技术，定期进行工作汇报，为该项目的开发储备知识。预计耗时 3 个月。

第二阶段：搭建服务器端并创建数据库。搭建服务器并建立相应的 web 服务，使其能协调网络爬虫、数据库访问、服务器脚本运行、响应 Android 客户端的请

求等任务。创建数据库存储爬虫和数据分析相关的数据。预计耗时 1 个月。

第三阶段：爬取信息并进行整合。利用爬虫爬取信息，并根据实际情况制定防止爬虫被封锁的策略，实时维护爬虫；对爬取到的信息进行初步的处理。预计耗时 3 个月。

第四阶段：进行数据挖掘和数据分析。结合数据挖掘和数据分析领域相关的知识，寻找数据之间的联系性，挖掘其中有价值的信息，并通过统计的方式对数据进行处理。预计耗时 3 个月。

第五阶段：开发客户端 APP。完成客户端的界面设计、程序框架的搭建和基础功能的实现。预计耗时 3 个月。

第六阶段：整体平台的联合调试。验证整体系统的功能和可用性，调查收集用户的反馈意见，并在此基础上完善、增加产品的功能。预计耗时 3 个月。

综上所述，该项目共需时 16 个月。

（七） 已有基础

1. **与本项目有关的研究积累和已取得的成绩**

目前本项目构思阶段已经完成，项目组成员已经进入相关技术的学习和研究阶段，并取得了如下的研究成果；

（1）阅读了网络爬虫、数据分析的部分论文，基本确定了爬虫设计和信息只能整合的大体思路。

（2）基本确定了 APP 需要实现的功能，对框架的选取、界面的设计有了一致的意见。

（3）团队成员专业对口，具备良好的编程基础，能够很快地掌握 Python 语言并加以运用。

2. **已具备的条件，尚缺少的条件及解决方法**

已具备的条件：

项目组成员具有一定的技术基础，熟悉与该项目相关的一些难点技术。小组成员对数据库、Android 编程都有一定了解，且具备软件开发能力。更重要的是，小组成员在进行了两年的专业学习之后，有一颗迫切希望实践的心，成员的热情为项目的实现提供了有力的保障。

申请人基本情况介绍如下：

李锦瑢，女，计算机学院计算机科学与技术 2015 级学生。获院级三好学生称号，积极参加英语竞赛、编程竞赛。2016 年 FIRA 足球机器人世界杯仿真 5VS5 项目国际冠军。自 2016 年 9 月在足球机器人基地工作有良好的团队意识，思维开阔，责任心强，对程序设计和算法有浓厚兴趣。

殷康龙，男，软件与微电子学院软件工程专业 2015 级学生。获得过校数模三等奖，参与过 web 项目开发，有一些合作开发的经验。2016 年在团体项目 FIRA 足球机器人世界杯仿真 5VS5 中，获得国际冠军。2016 年 9 月参加足球机器人基地，

编程能力得到很大提升。

白向龙，男，自动化学院自动化专业 2015 级学生。获院级三好学生称号，参加编程竞赛。2016 年 FIRA 足球机器人世界杯仿真 5VS5 项目国际冠军。对数据挖掘有浓厚兴趣，有较好的合作意识。

董金国，男，软件与微电子学院软件工程专业，2016 级本科生。2016 年第 21 届 FIRA 机器人世界杯仿真 5V5 组冠军，2016 年校“编程之星”程序挑战赛二等奖，2016 年模拟联合国校大会 Contribution Award，2016 年“外研社”英语写作比赛校内赛优秀奖。从大一刚入学加入足球机器人基地工作，积极参加英语竞赛、模联培训，积极承担多种志愿工作。有良好的团队意识，乐于助人、坚持不懈的品质。

薛建业，男，教育实验学院机械电子专业 2015 级学生。2015-2016 学年平均学分积全院第 12，综合评测全院第 13。获校三好学生称号，并获校级一等奖学金。2016 年获校级数模竞赛三等奖，FIRA 世界杯冠军，积极参加 ACM 竞赛，力学竞赛。有良好的团队意识和开拓创新的精神。

尚缺少的条件：

1）经费条件紧张。由于在相关资料文献和软件的购置方面需要大量的资金，致使项目开发的进度被延缓，开发过程受阻。经费短缺成为了项目开发的一大瓶颈。

2）理论深度不够。由于研究内容困难度大，任务量重，组员初次接触相关理论、研究深度不够，在后续的研发过程会遇到较大的阻力。

解决方法：

1）以项目经费作为项目支出，解决经费紧张问题，加快开发进度，严格按照进度安排实现项目开发。

2）加强组员的理论学习。通过查阅学习大量相关资料，深化组员的理论学习，在相关的知识背景下，拓展自己的知识面，并在组内经常开展学习交流，提升研究队伍的整体科研水平，为后续的研发提供扎实可靠的理论基础和理论依据。

三、 经费预算

开支科目	预算经费（元）	主要用途	阶段下达经费计划（元）	
			前半阶段	后半阶段
预算经费总额	20,000	**硬件采购，实验耗材，材料费，专利申报**	1,4300	5,700
1.业务费				
（1）计算、分析、测试费	4,000	**对硬件进行测试、调整**	4,000	0
（2）能源动力费				
（3）会议、差旅费				

（4）文献检索费	500	国内外的付费文献资源的使用权限购买	300	200
（5）论文出版费	1,500	论文出版	0	1,500
2.仪器设备购置费	6,000	租用服务器	6,000	0
3.实验装置试制费	3,000	损坏设备维修	1,000	2,000
4.材料费	3,000	图书资料、硬件杂项	3,000	0
5.申请专利	2,000	申请专利	0	2,000
学校批准经费				

四、 指导教师意见

导师（签章）：

年 月 日

五、 院系大学生创新创业训练计划专家组意见

专家组组长（签章）：

年　月　日

六、 学校大学生创新创业训练计划专家组意见

负责人（签章）：

年　月　日

七、 大学生创新创业训练计划领导小组审批意见

负责人（签章）：

年　月　日