数据收集与处理调研

我们调研了数个国内的网课平台例如：中国大学 MOOC、imooc、新东方网课平台等，最后，我们选择了最熟悉也是实现可能性最高中国大学 MOOC。

中国大学 MOOC 的用户个人中心界面（具体图 1）由用户信息（包括用户名、身份、关注数、粉丝数、主题/回复数、学习时长）、课程信息（包括课程名、课程学校、课程人数）两个模块组成。其中调研结果显示网站用户进行关注的积极性低，所以关注数和粉丝数没有推荐价值，不进行登录的话无法查看具体课程的学习进度，也无法看到该名学生的学校信息。

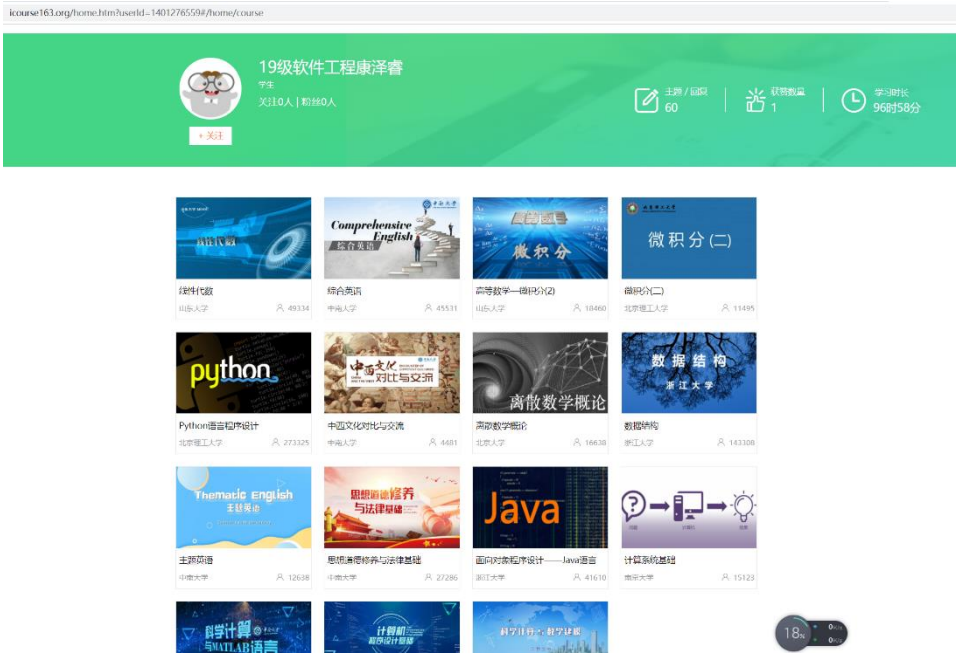


图 1 中国大学 Mooc 用户中心界面

为了根据行为轨迹进行用户之间的推荐我们暂时先选取了可以直接抓取到的几项数据：用户名、主题/回复数、学习时长、课程名、课程学校来编写我们的爬虫代码。

为了编写爬虫代码我们进行了技术调研，编程语言因为 python 的语言特性，我们的首选是 python，向服务器发送请求可以用 python 内置模块 urlib 或者外部库 requests，其中 requests 库继承了 urlib 和 urlib2 的所有特性。源码解析可以用正则解析，但是这种方法比较麻烦，所以我们可以用 bs4 库下的 beautifulsoup，这个解析器方便、快捷。

在调研之后我们做出了以下的部署：

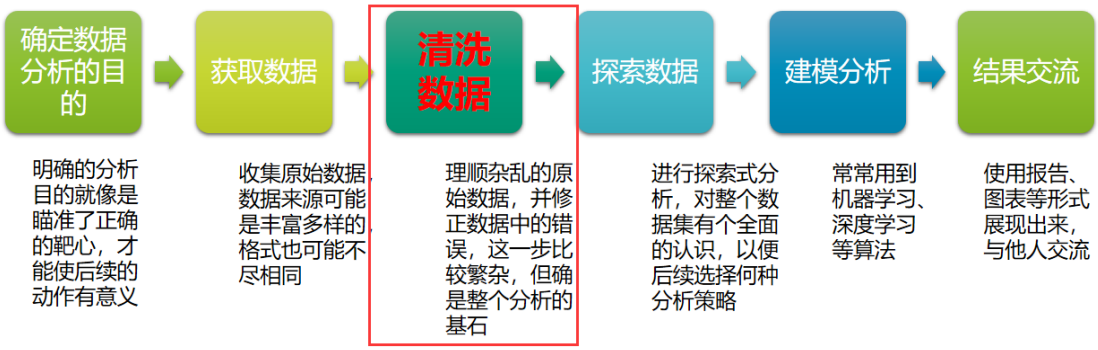
(1)编程语言：python

(2)向服务器发起请求：外部库 requests

(3)解析器：beautifulsoup

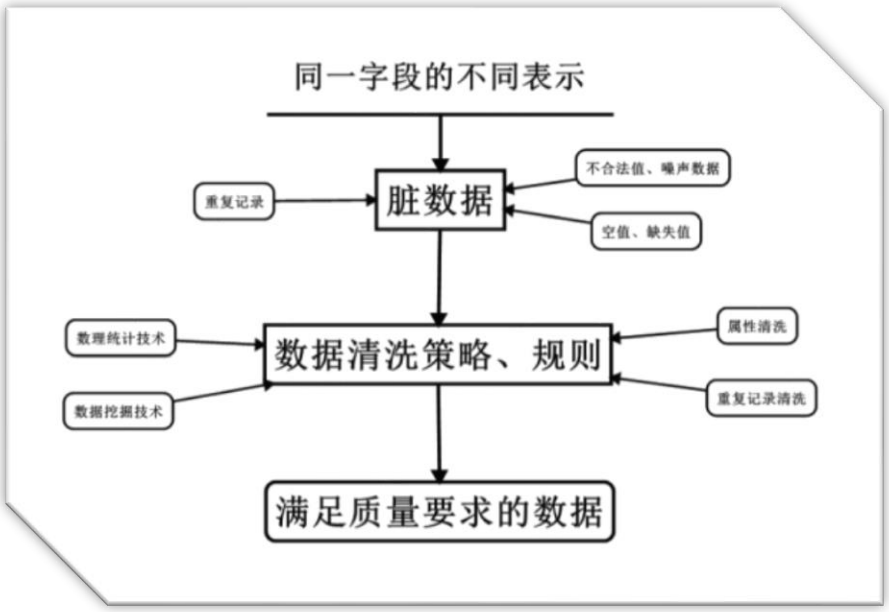
(4)保存数据：Mysql

得到数据后，对数据进行处理本推荐系统算法实现功能的核心要点，数据处理的整个流程如下：



而数据清洗是其中极其重要的一部分，其直接关系到能否对数据进行下一步分析的问题。

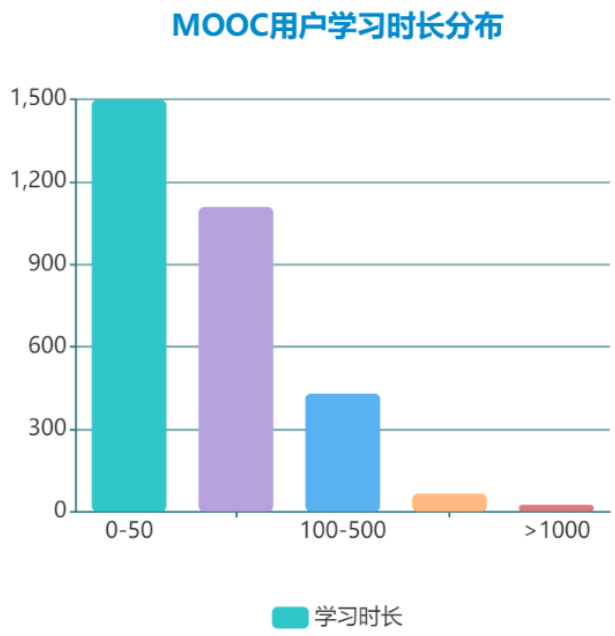
数据清洗是指从记录表、表格、数据库中检测、纠正或删除损坏或不准确记录的过程，简单来说，数据清洗就是把“脏数据（残缺数据、错误数据、重复数据、不符合规则的数据等）”变为“干净的数据（即可以直接带入模型的数据）”



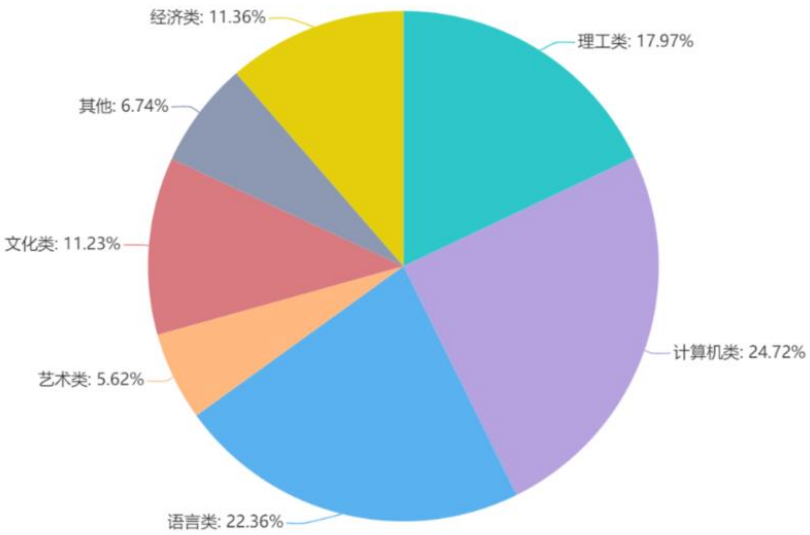
我们通过查找资料进行学习、调查、研究了解了关于数据清洗的流程如下：

由学习与分析我们了解到我们需要学习更多的相关知识技术来进行对数据的诸多处理（重复值、缺失值、异常值、文本字符串、时间格式序列）；

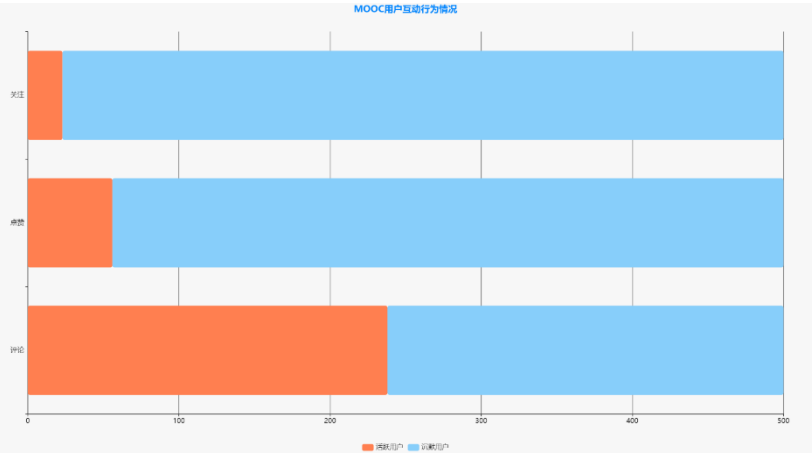
最后我们对清爬取的数据进行的一个简单的分析：



结论：用户的学习时长主要集中在学习时间比较少的区间，可见大部分学生创建了 MOOC 账户之后并没有频繁地利用 MOOC 平台进行线上学习，这和传统教学方式在当下教育体系的主导地位有关



结论：MOOC 用户的选课总体上是平均的，用户相对更倾向于选择理工类（包括物理、化学、生物等）、计算机类和语言类的课程，这可能和理工科用户基数比文科学生多有关。同时对比其他科目 MOOC 用户有更青睐在线上进行语言学习的趋势。



结论：MOOC 用户在网课学习中更倾向于进行评论，这可能和学校课程相关的硬性要求有关。但是 MOOC 用户的关注和点赞积极性很低，这说明 MOOC 用户在进行网课学习时不关注互动、没有互动意识，我们的项目正好可以针对这方面问题来改善现状，通过推荐学习伙伴来调动 MOOC 用户在平台上和其他用户的互动从而调动他们的学习积极性。