题 目：基于人工策略及机器学习算法的网页分类器模型的设计与过滤系统的实现

学院：软件学院 专业：软件工程 学生姓名：XXX 学号：XXXXXXXX

项目概述：

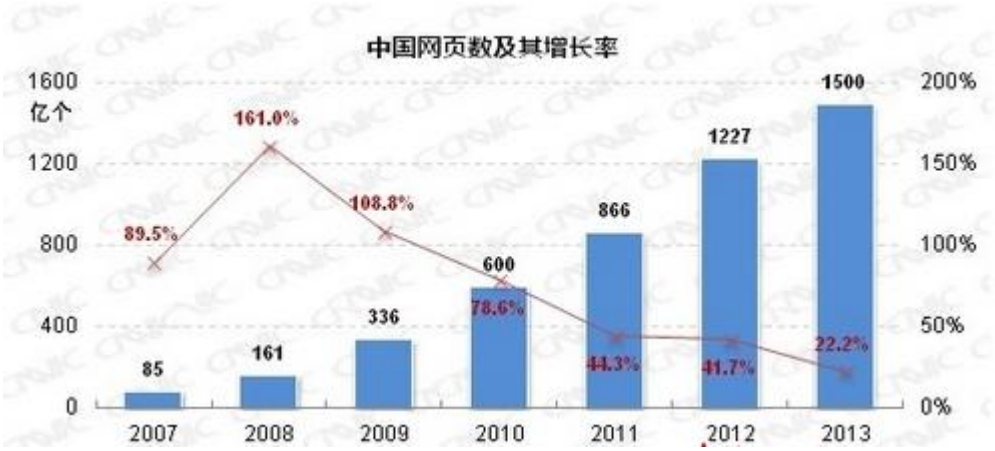
项目来源：

题目来源于本人在百度在线网络技术（北京）有限公司实习的有效性控制相关项目，该项目主要通过设计开发策略模型（包括人工策略模型及机器学习模型）用于识别各种低质垃圾网页（如作弊页，空页面，报错页等），并将垃圾网页从百度百灵库（网页库）中删去，以提升网页库的有效性，垃圾占有率，并为百度节省大量服务器空间。且将模型上线至 EC（计算中心）模块后，可用作过滤垃圾网页，防止垃圾网页入库。

项目背景：

现如今中国互联网飞速发展，网页数量也不断增多。

中国互联网信息中心(CNNIC)发布《第 33 次中国互联网络发展状况统计报告》(以下简称《报告》)。《报告》显示，截至 2013 年 12 月，中国网页数量为 1500 亿个，相比 2012 年同期增长了 22.2%。单个网站的平均网页数和单个网页的平均字节数均维持增长。[1]




然而搜索引擎需要抓取、存储、索引这些日益增长的网页，单从存储来看，就需要不断增多的物理空间，而导致节省这些存储空间变得迫不及待。然而，为了保证搜索引擎的覆盖率，也就是让有价值的网页能够被搜到，搜索引擎必须存储这些有价值的网页，然而在搜索引擎抓取互联网网页的时候，为确保覆盖率而尽可能的抓取网页，导致抓取的网页中，无价值或者低价值的网页也被存储而占用极大空间！所以删除，过滤这些垃圾网页也成为必要之举。

垃圾网页是利用搜索引擎运行算法的缺陷，采取针对搜索引擎的作弊手段，使其获得高于其网络信息质量排名效果的网页．垃圾网页欺诈的现象在万维网环境中非常盛行。如果搜索引擎技术不能很好地解决垃圾网页过滤的问题，致使搜索结果中存在大量对某些搜索需求来说毫无意义的垃圾网页，那么必然给用户应用搜索结果带来不便和麻烦，从而影响搜索服务的应用。[2]

项目开发意义:

项目启动前，百度网页库网页数已达 1400 亿之多，空间消耗巨大，且经抽样调研发现垃圾网页占网页库的 40%之多，不利于下游建库、索引等工作，为优化网页库有效性，启动该项目。通过在 EC（计算中心）模块中设计开发策略模型标记并从网页库中删除垃圾网页以提高百度网页库有效性，不仅打击已入库的低质垃圾网页，还能将模型上线至网页入库前的入口，用于过滤垃圾网页，而非等到垃圾网页入库占据大量空间后再进行删除。

项目现实价值:

项目阶段性收益:

1.线上收益:

模型上线后，spider 全环流量中，经抽样调研 28.9%的网页被识别为低质 0 分，可删除，其中低质识别准确率为 99%。即本次上线模型可将大环流量的 28.9%(这些网页垃圾比例为 99%)抵挡在网页库外。

2.离线网页库收益:

刷库后经抽样调研刷库收益为 220 亿，即删除网页库 220 亿垃圾，为百度网页库节省巨大空间。

重点和难点：

1. 由于处理的网页数量巨大，单纯的采用现有的机器学习算法模型，需要处理极长的时间（由于时间复杂度极大，可视作不可能完成）。

2. 为解决机器学习算法复杂度大，处理时间极长的问题，需要通过将模型分成人工策略模型及机器学习两部分进行开发，人工策略模型也就是通过人为的制定规则筛选可能为指定页面类型的网页。

3. 在开发人工策略模型中，人为的制定规则就需要人为的发现有效特征，并指定规则。难点在于发现的特征是否能够有效识别指定页面类型，且召回足够多且准确率满足要求的（尽量 60%的准确率以上）特征及规则

4. 在开发机器学习模型中，也需要发现可能识别页面类型的特征或者特征组合。

5. 最终模型设计完成（包含保护模型）需要确保准确率 99%的情况下，尽量提高召回率，以提升收益。 如若准确率低，将可能召回一些有价值的网页，甚至是极有价值的（诸如，淘宝首页，腾讯首页等）网页导致客户（站长）受到严重损失。

设计方案：

项目背景：

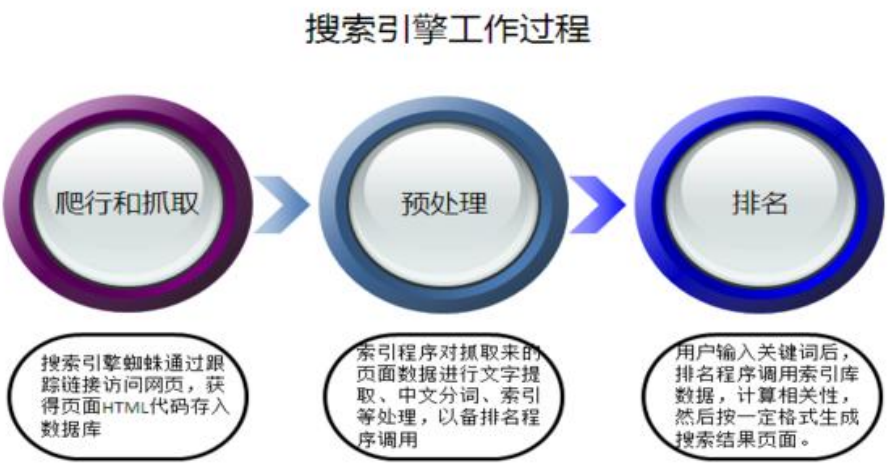
通过以上实习项目中设计开发的网页类型识别模型，加以研究如何更好的实现模型开发，并在满足高准确的要求下提升召回率。

网页类型识别模型开发已达到较成熟的阶段，继续发现模型的应用，不仅在实习项目中提及的用于识别垃圾网页，删除网页库中垃圾网页，并过滤垃圾网页的作用。还可能存在多方面的应用，诸如，识别出索引页，用以 EC 模块下游的链接发现等模块及时发现新网页并将有价值的网页收入至网页库。

理论基础：

1. 在 linux 系统下开发 C/C++程序的能力，以及熟悉 linux 下指令，编写 shell 来实现想要完成的任务。
2. 了解搜索引擎的原理及工作流程：

   搜索引擎的工作过程大致可以分成三个阶段：




一、爬行和抓取：搜索引擎蜘蛛通过跟踪链接访问网页，获得页面 HTML 代码存入数据库。

二、预处理：索引程序对抓取来的页面数据进行文字提取、中文分词、索引等处理，以备排名程序调用。

三、排名：用户输入关键词后，排名程序调用索引库数据，计算相关性，然后按一定格式生成搜索结果页面。[3]

在熟悉了解搜索引擎的原理及工作流程之后，才可以发现网页类型识别模型在其他方面的可能的应用并取得收益，原公司实习项目（《有效性控制》）则是在搜索引擎将网页抓取后，对网页的预处理（诸如通过 EC 模块对网页分析计算，设置字段，标明网页类型，如是否为低质垃圾网页等）供下游的建库，排名等使用。

3. 机器学习模型开发的流程及原理。 熟悉了解所应用的机器学习算法，如本次实习项目中用到的最大熵模型相关知识

最大熵原理:

最大熵原理是一种选择随机变量统计特性最符合客观情况的准则，也称为最大信息原理。随机量的概率分布是很难测定的，一般只能测得其各种均值（如数学期望、方差等）或已知某些限定条件下的值（如峰值、取值个数等），符合测得这些值的分布可有多种、以至无穷多种，通常，其中有一种分布的熵最大。选用这种具有最大熵的分布作为该随机变量的分布，是一种有效的处理方法和准则。[4]

对于分类问题(实习项目中的机器学习模型中用于对页面类型进行分类)，给定一些训练样本(x,y)，其中x表示上下文，y表示问题的类别，可根据已知的样本构建一个能够对实际问题进行准确描述的统计模型p( y | x)，用于预测未知事件。该模型的概率分布与训练语料中的经验概率分布应该相符。最大熵原理表明，x、y的正确分布应该是，在满足已知条件(约束)的情况下，使熵的分布最大，所构建的模型就是最大熵模型[5]

解决问题的方法和步骤:

1. 查阅相关文献，学习必备能力

2. 掌握独立调研能力，通过已有知识和能力，如编程、统计学，去调研相关信息，并分析所得数据
3. 了解搜索引擎相关知识
4. 学习模型设计开发流程及必备技能，如在 linux 下 C/C++语言的开发
5. 熟悉现有且较成熟的机器学习算法，熟悉机器学习模型开发流程（寻找特征，标注数据，训练语料等）

预期成果:

1. 详述并优化现有模型设计开发流程
2. 设法提升模块效率，通过调研实验，优化现有模块
3. 发现现有技术（通过在已有模块下开发页面类型识别模型）在其他方面的应用，及分析可能的潜在收益
4. 编写论文

交付物:

1. 论文
2. 网页类型识别模型
3. 调研报告，关于模型的新应用
4. 任务书： 开题报告 任务书 计划书

主要参考文献：

[1] 搜索服务中基于云计算的垃圾网页识别研究 李艳平 徐雅斌 陈俊伊

[2] 中国新闻网 http://www.chinanews.com/it/2014/01-16/5745005.shtml

[3] http://www.intertid.com/school/2013/590629.shtml

[4] http://baike.baidu.com/

[5] 最大熵模型的事件分类 于江德 1，李学钰 1，樊孝忠 2，庞文博 2

毕业设计（论文）进度安排：

| 序号 | 毕业设计（论文）各阶段内容 | 时间安排 | 备注 |
|---|---|---|---|
| 1 | 学习网页类型识别技术相关知识，收集整理数据及资料 | 2015.9.16 - 2016.1.16 | |
| 2 | 毕业设计开题报告和毕业设计任务书 | 2016.3.1 - 2016.3.4 | |
| 3 | 调研网页类型识别技术的其他可能应用，及分析预期收益 | 2016.3.5 - 2016.3.15 | |
| 4 | 分析研究不同机器学习算法对模型的影响，分析各算法的效率，及准确率、召回率的对比 | 2016.3.15 - 2016.3.20 | |
| 5 | 开始编写毕业论文 | 2016.3.21 - 2016.5.1 | |
| 6 | 准备毕设答辩 | 2016.5.1 - 2016.5.30 | |
| | | | |

| | | | |
|---|---|---|---|
| | | | |
| | | | |
| | | | |

指导教师（审核签名）：_____ 审核日期：_____年____月____日