

人机交互的软件工程方法 —— 评估之观察用户

主讲教师：冯桂焕

2020年秋季



背景

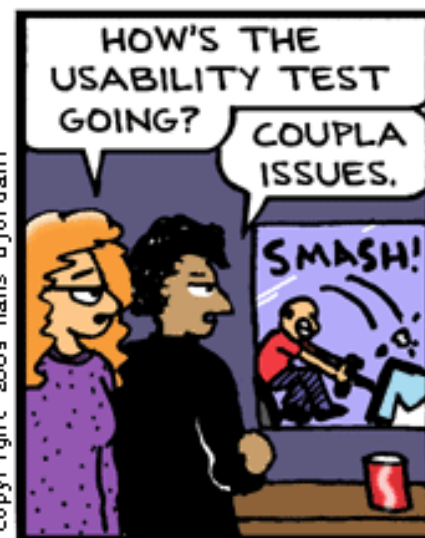
- 观察用户怎样工作是极其重要的可用性方法
 - 用户并不总能客观和完整地描述产品的使用情况
 - 访谈得到的信息可能不真实
 - 用户有可能忽略一些细节
 - 忽视了细节的重要性
 - Heath和Luff在伦敦地铁控制室的观察证明有助于改进系统设计
- 观察法是所有可用性方法中最简单的方法
 - 涉及看和听两个方面
- 观察适用于产品开发的任何阶段
 - 初期：理解用户的需要
 - 开发过程：检查原型
 - 后期：对最终产品进行评价



Bug Bash by Hans Bjordahl



Copyright 2005 Hans Bjordahl



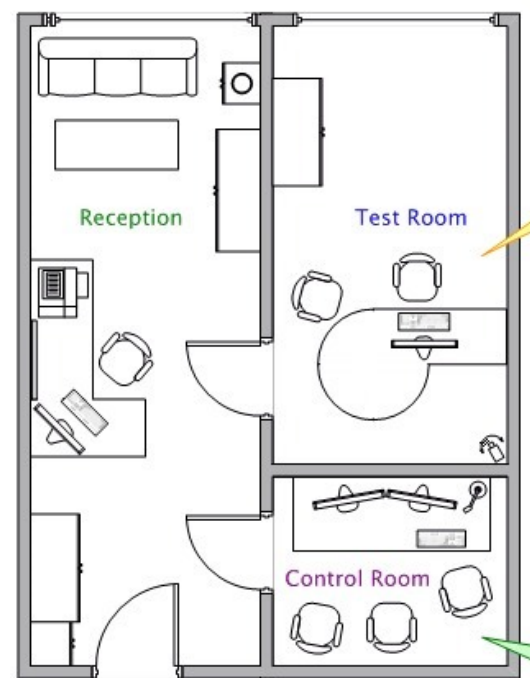
<http://www.bugbash.net/>

观察方式

- 观察可以发现一些意想不到的用户操作方式
 - 举例文字处理软件
 - 把“模板”当做一类特殊文件来处理
- 两种方式
 - 真实环境中的观察
 - 观察者既可作为旁观者，也可作为参与者
 - 重点是应用的上下文
 - 受控环境中的观察
 - 观察者不能作为参与者
 - 重点是研究用户执行任务的细节
 - 二者差别不大
 - 有时前者模仿后者的测试条件
 - 实地观察也可作为实验室观察的补充

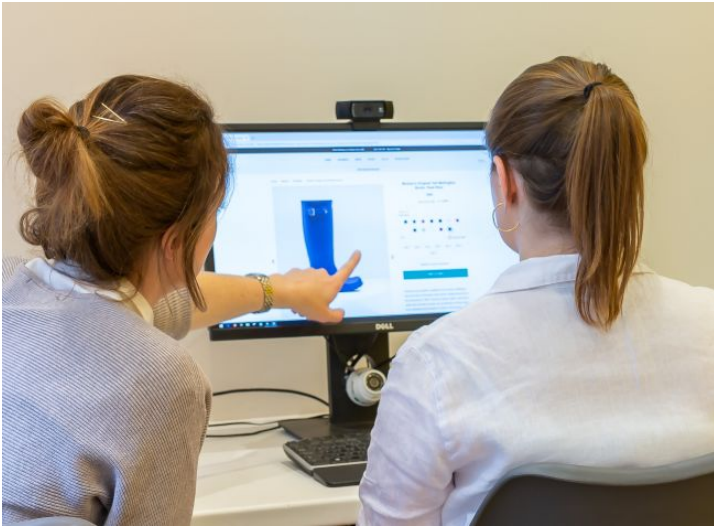
实验室观察

- 在专门为可用性测试而安装配置的固定设备的环境下进行的观察
 - 在一些情况下，在实验室中观察是唯一的选择
 - 如空间站等危险环境下的系统
- 实验室布局
 - 测试区
 - 观察区
 - 为避免干扰，二者分开
- 用户坐在家中测试
 - 如Windows Live Meeting等软件
 - 可观察到鼠标移动及屏幕内容的变化
 - 使用户处于更加自然和真实的状态



- 优点
 - 提供了可控且一致的评估环境
 - 易于分析比较
- 缺点
 - 可控且一致的评估环境
 - 人为环境、不自然
 - 可能降低测试结论的普遍性和一般性
 - 不利于观察多人之间的协作
- 具备观察测试过程的能力很重要
 - 对开发人员
 - 令开发人员感到丢脸面
 - 给他们一些启发和指导
 - 使之能够设身处地为最终用户着想





实验室观察的实际问题

- 测试地点选择，测试设备安装
 - 摄录面部表情
 - 摄录移动鼠标和击键过程
 - 更广范围内的用户肢体语言
- 测试设备检查
 - 确保设置正确，并能正常工作
- 文档准备
 - 协议书
 - 要求用户签署
 - 说明测试的目的、测试时间，并解释他们的权利
 - 让用户感觉轻松自在
 - 脚本
 - 要求用户执行的任务

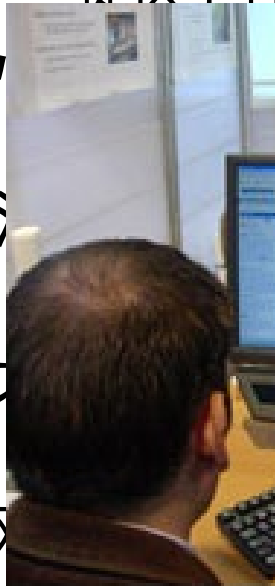
观察框架

- 观察过程发生的事件都非常复杂且变化迅速
 - 观察框架用于组织观察活动和明确观察重点
- Goetz and Lecomfte框架
 - 关注事件的上下文、涉及的人员和技术
- 人员：有哪些人员在场？他们有何特征？承担什么角色？
- 行为：人们说了什么？做了什么？举止如何？是否存在规律性的行为？语调和肢体语言如何？
- 时间：行为何时发生？是否与其他行为相关联？
- 地点：行为发生于何处？是否受物理条件的影响？
- 原因：行为为何发生？事件或交互的促成因素是什么？不同的人是否有不同的看法？
- 方式：行为是如何组织的？受哪些规则或标准的影响？

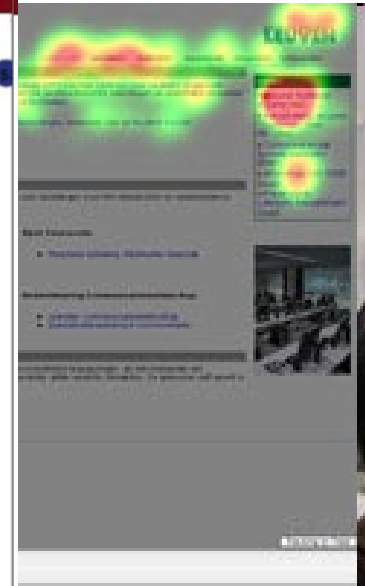
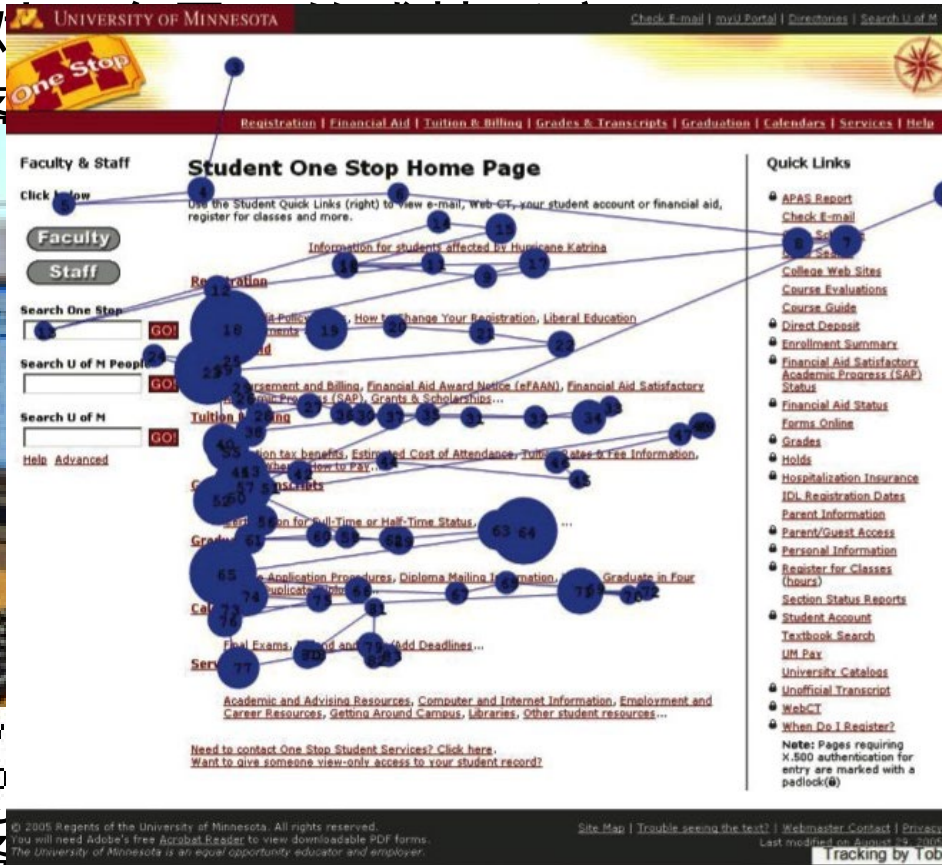
- Robson框架
 - 有助于组织观察和数据搜集活动
- 空间：物理空间及其布局如何？
- 行为者：涉及哪些人员？人员详情？
- 活动：行为者的活动及其原因？
- 物体：存在哪些实际物体（如家具）？
- 举止：具体成员的举止如何？
- 事件：所观察的是不是特定事件的一部分？
- 目标：行为者希望达到什么目标？
- 感觉：用户组及个别成员的情绪如何？

生理反应监控

- 决定用户对
- 提供了客



Left: Tobii Eye
Right: Hotspot



完成任务

观察中的问题

- 不知道用户在想什么
 - 只能根据观察到的现象去揣测
 - 举例：使用Web搜索引擎检索Alan Turing的专著
- 解决方法
 - 让用户“边做边说”
 - 要求被测试人说出自己的想法以及想要做的事情
 - 帮助观察人员了解他们的思考过程
 - 当用户沉默时，观察人员可以提醒用户
 - 优点：简单、只需要很少的专业技术
 - 缺点：不自然，可能改变人们执行任务的方式
 - 举例：<http://www.e-beam.com>

合作评估

- 两位用户共同合作，以便他们互相讨论、相互帮助
 - 限制少，评估者很容易学会应用
 - 鼓励用户对系统提出批评
 - 在出现不清楚的情况时，评估者能够澄清容易混淆的地方
 - 能提示许多信息
 - 尤其适合评估面向儿童的系统
 - 也适用于多人共享系统的评估

实验室观察

- 能够使得研究人员更好地分离多个可能的影响因素，从而能够得出更准确的研究结果
 - 对具有良好协调性的人机界面所起到的重要作用
 - 举例：键盘布局的试验研究
 - 使用相似的训练方法，标准的基准任务，普通的说明出错率的依赖程度以及测试平凡用户的策略
- 观察者对被观察者的影响取决于观察类型和观察技巧
 - 只对用户的某些行为感兴趣
 - 选择作为旁观者观察
 - 如了解不同性别学生使用计算机的时间差异
 - 如想了解计算机及其他设备如何影响学生们的交流
 - 则选择作为参与者进行观察

现场观察

- 指在用户的实际环境中观察用户在使用软件时的情况
 - 是发现同使用环境有关的问题的最佳手段
- 举例：超市收银系统
 - 工作环境非常吵杂
 - 收银员一般是站着操作
 - 必须尽快地为每一个顾客结账
 - 某些顾客可能不想要某个已经扫描过的商品
 - 某些顾客在结账时发现自己还想买一样东西，于是先把已经扫描过的东西放下，转身回去继续购买
 - 某些商品的标签打印得不清楚，条码扫描不起作用，收银员必须手工输入商品信息

- 举例：超市收银系统
- 观察结论
 - 操作效率要非常高
 - 使得收银员可以非常快速地完成各种常用操作
 - 要尽可能预防各种操作失误的产生
 - 可以不必太多关注系统的可学习性
 - 假定收银员一定是经过了良好的培训后才允许上岗的
 - 屏幕上显示信息时要一目了然
 - 让收银员可以轻松、正确地识别出各种信息

现场观察的问题清单

- 明确初步的研究目标和问题
- 选择一个框架指导观察
- 决定数据记录方式
 - 笔记、录音、摄像，还是三者结合
 - 确保设备到位并能正常工作
- 评估后，尽快与观察者或被观察者共同检查所记录的笔记和其他数据
 - 研究细节，找出含糊之处
 - 人的记忆能力是有限的，最好24小时内回顾数据

- 数据记录时，应区分个人意见和观察数据
 - 明确标注需要进一步了解的事项
- 培养良好的合作关系
 - 取得被观察对象的认可和信任
 - 避免只关注某些人，应注意小组的所有成员
- 处理敏感问题
 - 避免让观察对象产生被冒犯的感觉
- 注重团队协作
 - 通过比较不同评估人员的记录，得到更为可信的数据
- 应从不同的角度进行观察，避免只专注于某些特定行为

几个难题

- 要观察多久
- 如何根据紧凑的开发期限和开发人员的技能相应修改现场研究技术
- 如何降低噪音、测试中断及其他易使注意力分散的外界干扰
 - 如在测试进行的关键时刻可能会有电话打进来，或者用户在测试开始之前被请去开会了
 - 方案一：健壮的评价设计
 - 方案二：将测试协议设计成包含“有计划的干扰”
 - 以评估用户在中断当前任务后是否能很容易地返回，并知道如何将工作继续做下去

注意事项

- 观察人员自始至终应尽量保持安静
 - 让用户感觉不到观察人员的存在
 - 保证用户操作和平时工作的状态一样
- 当用户的操作令观察人员无法理解时
 - 需要打断用户，请他对所做的某些操作进行解释
 - 或把用户莫名其妙的操作行为记录下来
- 观察初期，应该拒绝用户的任何帮助请求
 - 希望观察用户在没有系统专家指点的情况下如何操作
 - 待评估完成后为用户提供适当帮助

观察与访谈相结合

- 观察方法只能展示用户做了什么，而无法知道用户为什么这样做
 - “知其然而不知其所以然”
- 应该在记录数据之后再结合其他方法
 - 如访谈
 - 请用户详细讲述记录里面任何可能引发可用性问题的地方
 - 如对一个没用过系统某个功能的用户，询问为什么没有使用某项功能
- 让用户面对记录数据时应非常小心
 - 避免让用户产生被监视的想法

数据记录

- 可以根据研究人员的专业素质及环境、项目的特点来选取合适的方法
- 纸笔记录
 - 最原始、最廉价
 - 前提
 - 对观察对象有一定的了解，从而有明确的观察侧重点
 - 优点
 - 事后对观察结果进行分析的工作量小
 - 缺点
 - 观察者容易疲劳，而且记录速度有限
 - 建议将记录者和评估者分开

- 音视频记录

- 适用场合

- 对于观察对象不太了解，或者是需要观察的内容较多
 - 特别是在采用边做边说法时

- 音频记录

- 信息全面，没有任何遗漏，便于事后详细分析
 - 没有可见记录，转录数据非常烦琐
 - 常用于提示重要细节或作为情景说明的辅助材料

- 视频记录

- 能够看到参与者正在做的事情
 - 但要始终让参与者停留在视觉的范围内很困难

- 缺点

- 所含的信息量很大，所以数据分析非常耗时

日志和交互记录

- 间接观察的适用场合
 - 直接观察可能影响用户
 - 或者评估人员无法在现场进行研究
 - 可根据搜集到的数据，推断实际情形，并找出可用性和用户体验方面的问题
- 优点
 - 体现了用户是如何完成真实任务的
 - 使得从工作在不同环境下的大量用户那里自动收集数据变得相当容易
 - 适用于用户分散、无法当面测试的情形
 - 如互联网应用和网站设计项目等

- 应用举例

- 命令行式操作系统下用户发出的6112条错误命令分析
 - 30%的错误命令是由于拼写有问题
 - 表明了需要拼写校对机制来协助系统用户输入命令
- 大型计算机系统的程序员收到的3000条错误信息
 - 其中85%属于9个一般性的错误
 - 一条很糟糕的错误信息占9.8%
 - “符号未在程序中定义”
 - 缺少更多相关信息的情况下很难从根本上纠正该错误
 - 修改后，该错误信息出现的比例只占1.7%
 - 使用新的错误提示可以避免重犯同样的错误

- 包含信息

- 用户使用软件的频度是多少？
- 每次使用软件的时间有多长？
- 对于不同操作的使用频度如何？
- 哪些操作最常用，哪些很少使用？
- 用户是经常使用键盘还是鼠标来启动各个功能？
-

观察数据

- 观察得到的数据
 - 笔记、草图、相片、访谈或事件的录音录像、日志和交互记录等
 - 用户在做什么并统计用户花在任务各个部分上的时间
 - 用户的情感反应
 - 如叹气、皱眉、耸肩等
- 数据类型
 - 用于描述的定性数据
 - 描述观察到的现象
 - 用于分类的定性数据
 - 使用各种技术进行分类
 - 定量数据
 - 用于统计目的

定性分析

- 分析方法
 - 详细分析通常不必要
 - 分析每个词、短语、动作
 - 分析半小时的录像数据也需要很长时间
 - 粗略分析
 - 结合上下文研究具体动作
- 常用的方法
 - 找出关键事件，如用户遇到困难的地方
 - 特点：用户发布评论、保持沉默或表露出迷惑的神情
 - 举例：Smalltalk编程手册的两个不同版本的比较

- 内容分析（content analysis）
 - 用于“详细分析”录像数据
 - 把数据内容划分为一些有意义、而且互斥的类别
 - 不能以任何方式相互重叠、内容类别也必须可靠
 - 可靠：不同人的分类结果不能存在很大差异
 - 费时、费力、不常使用
 - 改进：层次化内容分类技术
- 会话分析（conversation analysis）
 - 用于仔细检查语义，重点是对话过程
 - 弄明白发话者想要表达什么，而受话者又是怎么样理解它的意思并做出反应
 - 可用于聊天室、虚拟社区等互联网应用，以增进对用户需要的理解

- 话语分析 (discourse analysis)
 - 关注话语的使用，而不是内容
 - 把语言视为反映心理和社会因素的媒介
 - 了解人们如何使用语言
 - 措辞上的微小改动即可改变话语的意思
 - 当你说“我正在进行话语分析”时，你实际上就是在进行话语分析.....
 - 张三认为，当你说“我正在进行话语分析”时，你实际上就是在进行话语分析.....
 - From “An interesting introduction to discourse analysis”
 - Please use the toilet, not the pool.
 - Pool for members only.
 - Please use the toilet, not the pool. Pool for members only.

定量分析

- 视频数据的标注和分析
- 统计分析
 - 平均值
 - 标准偏差
 - T检验

标记	事件
S	任务开始
E	任务结束
O	观察到异常/问题
G	用户的积极印象
Q	用户评论
X	错误或意外操作
H	用户向试验指导人员提供帮助
P	试验指导人员给用户提示
T	超时
?	待讨论操作
C	试验指导人员的评论
*	非常重要的动作

测试名称：编辑HTML文档		用户编号：5
日期：2012年8月26日 时间：10:50		页码：1/3
任务编号	消耗时间	观察到的现象
1	04:25	X: 打开错误的文件，发现错误； X: 再次打开错误的文件
	06:22	P
	07:10	T
2	11:30	Q: “希望都这样简单”
	15:20	? : 长时间犹豫，随后改正了动作
	16:13	E

分析工具

- 文本记录分析工具

- 访谈稿、现场记录、开放性问卷调查等
- NUDIST

- Non-numerical unstructured data indexing, searching, and theorizing

- Ethnograph

- 1985年发行，是第一个电脑辅助质性分析软件

- 视频分析工具

- Observer VideoPro
- Morea

- 网络日志分析工具

- Webtrends

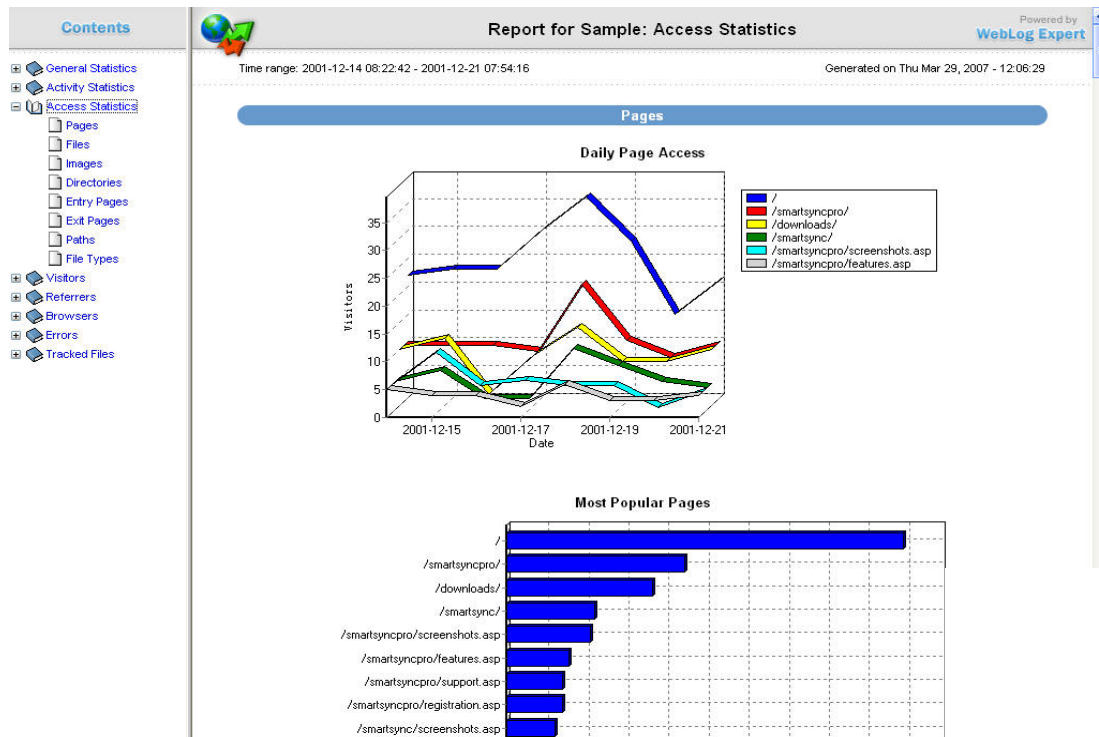
- 网站服务器端记录文件分析工具



WebLog

- WEB服务器日志文件综合分析软件

- 当前活动会话统计、文件存取统计、搜索使用情况统计、浏览器/操作系统统计、错误统计等



小结

- 直接观察
 - 实验室观察
 - 现场观察
- 间接观察
 - 日志和交互记录
- 数据记录
 - 纸笔
 - 音视频
 - 日志和交互记录
- 数据分析
 - 定性分析
 - 定量分析

关于大作业的问题

- 1. 项目的规模是否有要求？项目规模若是较大，前端页面数量较多的话，我们是否可以只选择部分页面进行修改，保证修改了的页面之间的一致性就行？
- 2. 如果需要使用真实数据，那么就存在对原有系统数据格式进行修改的可能性，比如增添 / 删除字段，而这些都会影响后端系统的逻辑实现，这部分的工作是否会纳入大作业的评分考量？还是说，就算爬取真实数据，也尽量不要修改后端逻辑与数据格式，专注于前端展示界面？

关于大作业的问题-2

- 3. 记得上课好像说不能重写界面？那么对重写应该如何定义？如果不能重写，那就是对修改的比例有所要求，这个可以修改的比例又是否有硬性规定呢？因为我们希冀于对界面进行重新设计，甚至修改界面的风格，所以自然会导致整个界面的改动，重写界面是一件很显然的事情。
- 4. 对前端的展示是否可以更换技术栈？比如原来使用jsp，现在使用React，那么界面基本就是重写了。。是否可以呢？