南京大学
NANJING UNIVERSITY

人机交互的软件工程方法
—— 评估的基础知识

主讲教师：冯桂焕
fgh@software.nju.edu.cn
2012年春季

背景

- 评估总是需要的
- 什么是评估？
  - 系统化的数据搜集过程
  - 目的是了解用户或用户组在特定环境中，使用产品执行特定任务的情况
    - 例如，用户能否找到特定的菜单项？图像是否有用，是否吸引人？产品是否引人入胜？
- 评估不是设计过程中一个单独的阶段
  - 优秀的交互设计师应掌握如何在不同的开发阶段评估不同的系统

评估目标

- 评估的优点
  - 能够在交付产品之前（而不是之后）修复错误
  - 设计小组能够专注于真实问题，而不是假想问题
  - 工程师们能专心于编程而不是争论
  - 能够大大缩短开发时间
  - 销售部门可获得稳定的设计
- 评估的目标
  - 评估系统功能的范围和可达性
  - 评估交互中用户的体验
  - 确定系统的某些特定问题

评估原则

- 评估应该依赖于产品的用户
  - 与专业技术人员的水平和技术无关
- 评估与设计应结合进行
  - 仅靠用户最后对产品的一两次评估，是不能全面反映出软件可用性的
- 评估应在用户的实际工作任务和操作环境下进行
  - 根据用户完成任务的结果，进行客观的分析和评估
- 要选择有广泛代表性的用户
  - 参加测试的人必须具有代表性

评估范型和技术

- “范型”与“技术”
  - 范型与具体学科相关，对如何评估有很大影响
    - 可用性测试是一种评估范型
  - 每种范型有特定的技术
    - 可用性测试的技术有观察、问卷调查、访谈等
- 评估范型
  - 快速评估
  - 可用性测试
  - 实地研究
  - 预测性评估

快速评估

- 设计人员非正式地向用户或顾问了解反馈信息，以证实设计构思是否符合用户需要
  - 可在任何阶段进行
  - 强调“快速了解”，而非仔细记录研究发现
    - 如在设计初期了解用户对新产品的意见、在设计末期了解用户对图标设计的看法等
  - 得到的数据通常是非正式、叙述性的
    - 可以口语、书面笔记、草图、场景的形式反馈到设计过程
  - 是设计网站时常用的方法
- 基本特征：快速

可用性测试

- 20世纪80年代的主导方法
- 评测典型用户执行典型任务时的情况
  - 包括用户出错次数、完成任务的时间等
- 基本特征
  - 是在评估人员的密切控制之下实行的
- 主要任务
  - 量化表示用户的执行情况
- 缺点
  - 测试用户的数量通常较少
  - 不适合进行细致的统计分析

实地研究

- 基本特征
  - 在自然工作环境中进行
- 目的
  - 理解用户的实际工作情形以及技术对他们的影响
- 作用
  - 探索新技术的应用契机
  - 确定产品的需求
  - 促进技术的引入
  - 评估技术的应用
- 分类
  - 评测人员作为“局外人”
  - 评测人员也可作为“局内人”或测试用户

预测性评估

- 专家们根据自己对典型用户的了解（通常使用启发式过程）预测可用性问题
  - 也可使用理论模型
- 基本特征
  - 用户可以不在场
  - 使得整个过程快速、成本较低
- 启发式评估是典型的预测性评估方法
  - 注意：启发式原则应定制
  - 可能误导设计人员
    - 且有些结果可能并不准确

评估范型比较

评估范型	快速评估	可用性测试	实地研究	预测性评估
用户角色	自然行为	执行测试任务集	自然行为	用户通常不参与
控制权	评估人员实施最低限度控制	评估人员密切控制	评估人员与用户合作	评估人员为专家
评估地点	自然工作环境或实验室	实验室	自然工作环境	类似实验室的环境，通常在客户处进行
适用情形	快速了解设计反馈。可使用其他交互范型的技术，如专家评测	测试原型或产品	常用于设计初期，以检查设计是否满足用户需要，发现问题，发掘应用契机	专家（通常是开发顾问）检查原型，可在任何阶段进行。使用模型评测潜在设计的特定方面
数据类型	通常是定性的非正式描述	量化数据，有时是统计数据。可采用问卷调查或访谈搜集用户意见	应用草图、场景、例证等的定性描述	专家们列出问题清单，由模型导出量化数据，如两种设计的任务执行时间
反馈到设计	通过草图、例证、报告	通过性能评测、错误统计报告等为未来版本提供设计标准	通过描述性的例证、草图、场景和工作日志	专家列出一组问题，通常附带解决方案建议。为设计人员提供根据模型计算出的时间值
基本思想	以用户为中心，非常实用	基于试验的实用方法，即可用性工程	可以是客观观察或现场研究	专家检查以实用的启发式原则和实践经验为基础，采用基于理论的分析模型

评估技术

- 观察用户
  - 有助于确定新产品的需求
  - 也可用于评估原型
  - 挑战：如何在不干扰用户的前提下观察用户，以及如何分析大量数据
- 询问用户意见
  - 简单，调查用户数量从几个到几百不等
- 询问专家意见
  - “角色扮演”方式评估
  - 同时专家会提出解决方案

评估技术-2

- 测试用户的执行情况
  - 可比较不同设计方案优劣
  - 通常在受控环境中进行
- 基于模型和理论，预测界面的有效性
  - 常用技术如GOMS模型和KLM模型等

评估范型和技术的关系

评估技术	评估范型			
	快速评估	可用性测试	实地研究	预测性评估
观察用户	观察用户实际行为的重要方法	使用摄像和交互日志的记录方式，可做进一步分析，以找出问题，了解操作步骤，计算执行时间	实地研究的核心方法。在现场研究中，评测人员与测试环境相融合；在其他类型的研究中，评测人员只做客观观察	——
询问用户意见	与用户和潜在用户讨论，可采用个别会谈、集体会谈或专门小组的形式	通过问卷调查了解用户满意度，也可通过访谈了解更多详情	评测人员可采用访谈的形式，与用户讨论观察到的问题。现场研究可采用现场访谈	——
询问专家意见	专家评估原型的可用性（提供“评估报告”）	——	——	在设计初期，专家使用启发式原则预测界面的有效性
用户测试	——	在受控环境中，测试典型用户执行典型任务的情况，是可用性测试的基本方法	——	——
用户执行情况分析模型	——	——	——	使用分析模型预测界面的有效性，或比较用户使用不同设计方案的执行效率

区分评估技术的因素

- 评估在周期中的位置
  - 设计早期阶段的评估更快速、便宜
- 评估的形式
  - 实验室环境or工作环境
- 技术的主客观程度
  - 技术越主观，受评估人员知识的影响越大
    - 如认知走查等
- 测量的类型
  - 与技术的主客观性有关
    - 主观技术：定性数据
    - 客观技术：定量数据

- 提供的信息
  - 低层信息：这个图标是可理解的吗？
  - 高层信息：这个系统是可用的吗？
- 响应的及时性
  - 边做边说法可及时记录用户行为
  - 任务后的走查取决于对事件的回忆
- 干扰程度
  - 直接响应测量可能会影响用户表现
- 所需资源
  - 设备、时间、资金、参与者、评估人员的专业技术及环境等

评估技术比较

方法	生命周期阶段	用户人数	主要优点	主要缺点
启发式评估	早期设计，反复设计过程的“内循环”	无	能发现单个可用性问题，能发现熟练用户碰到的问题	没有涉及真实的用户，故无法再用户需求方面有“惊人发现”
绩效度量	竞争性分析，最终测试	至少 10 人	硬性数据，对结果容易进行比较	不能发现单个可用性问题
边做边说	反复设计，形成性评估	3~5 人	准确了解用户的错误想法，测试费用低	用户感到不自然，熟练用户感到很难用语言表述
观察	任务分析，后续研究	3 人或以上	生态有效性；发现用户的真实人物；建议系统功能与特征	很难约定安排，实验人员无法控制
问卷调查	任务分析，后续研究	至少 30 人	发现用户主观偏好，容易重复进行	需要进行问卷预答（避免出现误解）
访谈	任务分析，后续研究	5 人	灵活，可以深入了解用户观点和用户体验	耗时，难以进行分析、比较
焦点小组	任务分析，用于参与	每组 6~9 人		分析起来困难，有效性低
使用过程记录	最终测试，后续研究	至少 20 人		
用户反馈	后续研究	上百人	跟踪用户需求和想法上的变化	需要专门部门来处理回复

评估方法组合

- 评估方法的组合取决于项目待评估的具体特性
- 常用组合
  - 启发式评估+边做边说等用户测试技术
    - 专家可通过启发性评估排除显而易见的可用性问题
    - 重新设计后，经用户测试，反复检查设计的效果
  - 访谈+问卷调查
    - 先对小部分用户进行访谈，确定问卷中的具体问题
- 启发式评估vs.用户测试
  - 前者不需要用户参与
  - 二者发现的可用性问题不同，可以互补

- 评估步骤

DECIDE评估框架

- 六个步骤
  - 决定评估需要完成的总体目标
  - 发掘需要回答的具体问题
  - 选择用于回答具体问题的评估范型和技术
  - 标识必须解决的实际问题，如测试用户的选择
  - 决定如何处理有关道德的问题
  - 评估解释并表示数据

1. 确定目标

- 评估目标决定了评估过程，影响评估范型的选择
- 为什么要评估？
  - 产品设计是否理解了用户需要？
  - 为概念设计选择最佳隐喻？
  - 界面是否满足一致性需要？
  - 探讨新产品应做的改进？
- 举例
  - 设计界面时，需量化评价界面质量
    - 适合进行可用性测试
  - 为儿童设计新产品时，要使产品吸引人
    - 适合采用实地研究技术，观察儿童交谈

2. 发掘问题

- 根据目标确定问题
  - 目标：找出为什么客户愿意通过柜台购买纸质机票，而非通过互联网购买电子机票
  - 问题
    - 用户对新票据的态度如何
      - 是否担心电子机票不能登机
    - 用户是否能够通过互联网订票
    - 是否担心交易的安全性
    - 订票系统的界面是否友好
      - 是否便于完成购票过程
- 问题可逐层分解

3. 选择评估范型和技术

- 范型决定了技术类型
- 必须权衡实际问题和道德问题
  - 最适合的技术可能成本过高
  - 或所需时间过长
  - 或不具备必要设备和技能
- 可结合使用多种技术
  - 不同技术有助于了解设计的不同方面
  - 不同类型数据可从不同角度看待问题
  - 组合有助于全面了解设计的情况

4. 明确实际问题

- 用户
  - 应选择恰当的用户参与评估
    - 能代表产品的目标用户群体
    - 可以先做测试，确定用户技能所属的用户群
  - 任务时间多长
    - 20分钟休息一次
  - 可在任务执行前，安排用户熟悉系统
- 设施及设备
  - 如需多少台摄像机录像，具体摆放在何位置
- 期限及预算是否允许
- 是否需要专门技能
  - 没有可用性专家

5. 处理道德问题

- 应保护个人隐私
  - 除非获得批准，否则书面报告不应提及个人姓名，或把姓名与搜集到的数据相联系
  - 受保护的个人资料包括健康状况、雇佣情况、教育、居所和财务状况等
  - 可在评估前签署一份协议书
- 指导原则
  - 说明研究的目的及要求参与者做的工作
  - 说明保密事项，对用户&对项目
  - 测试对象是软件，而非个人

- 指导原则-2
  - 对测试过程的特殊要求，是否边做边说等
  - 用户可自由表达对产品的意见
  - 说明是否对过程进行录像
    - 不能拍摄用户的面部
  - 欢迎用户提问
  - 用户有随时终止测试的权利
  - 对用户话语的使用应征得同意，并选择匿名方式

6. 评估、解释并表示数据

- 搜集什么类型的数据，如何分析，如何表示
  - 通常由评估技术决定
- 可靠性
  - 给定相同时间，不同时间应用同一技术能否得到相同结果
  - 非正式访谈的可靠性较低
- 有效性
  - 能否得到想要的测量数据
- 偏见
  - 评估人员可能有选择地搜集自己认为重要的数据
- 范围
  - 研究发现是否具有普遍性
- 环境影响
  - 霍桑效应

小规模试验

- 对评估计划进行小范围测试
  - 以确保评估计划的可行性
  - 如检查设备及使用说明
  - 练习访谈技巧
  - 检查问卷中的问题是否明确
- 小规模试验可进行多次
  - 类似迭代设计
  - 测试——反馈——修改——再测试
  - 快速、成本低

可用性问题分级

- 评估结果总是可用性问题清单，以及改进建议
- 方法一：基于量化数据的分级
  - 如多少人遇到该问题，耗费多少时间等
- 方法二：问题严重性的主观打分，取平均值
  - 0：不是一个可用性问题
  - 1：一个表面的可用性问题
    - 如果项目时间不允许，可不予纠正
  - 2：轻微的可用性问题
    - 优先级较低
  - 3：重要可用性问题
    - 需要重视，给以高优先级
  - 4：可用性灾难
    - 产品发布之前必须纠正

- 方法三：可用性分级的两个因素
  - 多少用户会遇到这个问题
  - 用户受该问题影响的程度

问题对用户的影响程度 \ 遇到问题的用户比例	少	多
小	低严重性	中严重性
大	中严重性	高严重性

- 方法四：该问题只在第一次使用时出现，还是会永远出现
  - 举例：菜单条中的下拉菜单
    - 用户从不尝试下拉用图标表示的菜单
    - 有人告诉他们后，可马上知道如何克服该不一致性问题
    - 因此该问题不属于永久性的可用性问题

小结

- 常用评估范型和技术
  - 范型和技术的区别
- 技术的选择
  - 哪些影响因素
- DECIDE评估框架
  - 6个步骤
- 可用性问题分级
  - 为避免偏差，建议综合多个评价者的意见
    - 研究发现，一位可用性专家作出的严重性评价与真实结果之间的误差在0.5以内（5分制）的概率只有55%
    - 4名专家所做评价的平均值，其概率为95%