User-Centric Computing for Human-Computer Interaction NPTEL-MOOCS L26

Dr Samit Bhattacharya

Computer Science and Engineering

IIT Guwahati

Learning So Far …

- Idea of interactive computing systems
- How to design (software development life cycle)
- User-centric computing for design (the framework)
- UCC models (user models + formal models)

Basic Idea

- When we perform **a controlled experiment** to collect and analyze data on user behavior, the entire process is known as *empirical research*

Basic Idea

- Term not specific to study of human behavior only - any study involving observation-based data collection and analysis, for any purpose, is empirical research
- Data we get from observation is **empirical data**

Basic Idea

- Empirical research not easy –to obtain *reliable* empirical data that lead us to *reliable* conclusion, we have to follow a systematic process (consisting of series of stages in sequence)

Stages

- Broadly, four (or five) stages
  - Identification of research question(s)
  - Determination of variables
  - Design of experiment
  - Analysis of empirical data
  - There is also a fifth stage: building of a model, if that is what we want

Understanding the Stages

- In empirical research, we seek answer to one or more questions
  - E.g., "how good my system is" - not necessarily a *good* question
  - Finding a good question difficult and requires expertise
- We should always start by clearly specifying one or more **good research question(s)**

Understanding the Stages

- Next, we need to identify *variables* - to remove ambiguity in observation
  - E.g., unless we specify variables, we do not know what to observe
- We need variables to record observations for later analysis

Understanding the Stages

- Third stage: **experiment (study) design** - refers to planning and execution for the study

Understanding the Stages

- Getting **right users** and in **right numbers** essential - otherwise, results may not be *reliable*
- It is also necessary to determine **appropriate tasks** in a **suitably controlled** environment - without that, we may again end up with unreliable data

Understanding the Stages

- Finally, we should employ **appropriate** data analysis techniques to extract conclusions

Understanding the Stages

- For building UCC models, we have to follow separate procedure to determine the model (fifth and final stage)

Illustrative Case Study

- Suppose we are interested in building **a computational model for our aesthetic judgment behavior**
  - Important topic since there is direct relation with usability (measure of **satisfaction)**

Illustrative Case Study

- We wish to collect empirical data to build the model
- **First step – research question**

Research Question

- Consider the research question

  RQ1: *How our aesthetic judgment depends on the interface?*

Research Question

- With RQ1, we are supposed to observe the relationship between the user's aesthetic judgment behavior and the interface

Research Question

- To observe the relationship, we set up an experiment

Research Question

- We present an **interface** to the user and ask him/her to judge its aesthetics
- We change the interface and ask the user to judge it again
- **We repeat the process few times** (say for five interfaces) and complete our experiment

Research Question (Recorded Observation)

Experiment log

Observations for user #1

Interface #1: good aesthetic

Interface #2: poor aesthetic

Interface #3: not very good but not very bad either

Interface #4: good aesthetic

Interface #5: very bad aesthetic

Research Question

- Observation reveals *aesthetic judgment depends on the interface*
  - Something well known - study does not reveal anything new
- We are not likely to get answer to RQ1: **how** they are related

Research Question

- This is because there are **elements of vagueness** in both the question (RQ1) and the observations

Research Question

- Idea of interface is vague
- Idea of aesthetic judgment behavior is vague
- User opinions can be vague (e.g., user judgment for interface #3)

Research Question

- We are **dealing with observations** that are **difficult to interpret** due to the **vagueness inherent** in the research question

Research Question

- We are **dealing with observations** that are **difficult to interpret** due to the **vagueness inherent** in the research question

Research Question

- Let us define an interface to be a collection of *objects* (images, text blocks, headings, tables, animations …)
  - **Interface represented in terms of number of objects**

Research Question

- We reformulate the research question

  RQ2: How our aesthetic judgment depends on the number of objects an interface has?

RQ2 - Observations

Experiment log

Observations for user #1

Interface #1 (2 objects): good aesthetic

Interface #2 (7 objects): poor aesthetic

Interface #3 (5 objects): not very good but not very bad either

Interface #4 (3 objects): good aesthetic

Interface #5 (10 objects): very bad aesthetic

Research Question

- From the recorded observations, we can come to the conclusion

  Aesthetic judgment goes to the negative side as the number of objects increases

Research Question

- We can do even better

Research Question

- Let us define aesthetic judgment as a *score* on a rating scale (say between 1 to 5) - higher the score, the better is the aesthetic
- We can rephrase the research question as in RQ3

Research Question

RQ3: How the aesthetic score (in a scale of 1-5) depends on the number of objects an interface has

RQ3 - Observations

Experiment log

Observations for user #1

Interface #1 (2 objects): aesthetic score = 4

Interface #2 (7 objects): aesthetic score = 2

Interface #3 (5 objects): aesthetic score = 3

Interface #4 (3 objects): aesthetic score = 5

Interface #5 (10 objects): aesthetic score = 1

Research Question

- Now easier to analyze data and answer RQ3
- We can actually **obtain a mathematical relationship** between the score and interface

Research Question

- Idea is simple
- In a 2D graph, plot along X-axis the N_I (no of objects) values and the corresponding AS (rating) values along the Y-axis
- Next, perform a **regression analysis** on the data points to obtain an equation

$$AS = f(N_I)$$

Research Question

- Alternatively, we might use the observations to build the training data and apply a learning approach to get a learning-based model

Research Question

- So, what is there in RQ3 that makes it **better** than RQ1
- TWO major differences

Research Question

- **Lack of ambiguity**
  - We are specifying that we wish to capture aesthetic judgment in terms of a number (the rating) rather than leaving it to the imagination of the users
  - We are also defining the interface in terms of the number of objects it contain - helps to differentiate between interfaces

Research Question

- **Measurable quantities**
  - We cannot measure *aesthetic judgment* - possible if we tell it is a number in a scale of 1-5
  - Similarly, we cannot measure an *interface* but we can measure number of objects in an interface
  - Therefore, we have **replaced unmeasurable concepts** with **measurable quantities**

Research Question

- We call R3 **testable** and **R1** as **non-testable**

Tradeoff

- Our aim - frame *testable* questions
- Problem
  - Testable questions designed to seek answer to *specific* queries
  - Such questions may lack *generalizability*

Tradeoff

- Example – RQ3
  - Objective is to determine relation between *a rating score on a 5-point scale* (representing aesthetic judgment) and *number of objects* (representing interface)
  - Will that represent **the relationship** between our aesthetic judgment and the interface

Tradeoff

- 5-point rating scale not the only way to represent judgment

Tradeoff

- 3-point scale may be more convenient for some (e.g. poor aesthetics = 1, average aesthetics = 2 good aesthetics = 3)
- Some may like larger scale with more spread (say a 10-point scale) (e.g. “bad”, “very bad”, “not so bad”, “slightly better” and so on)

Tradeoff

- With change in rating scales, observations change with a corresponding change in the relationship

Tradeoff

- Similarly, number of objects need not be the only characteristic of an interface that affect aesthetics

Tradeoff

- **Type of objects** (image, text or animation) should be important (e.g., interface #1 has text and image objects, interface #2 has only text objects and so on)
- **Geometric positioning (layout)** of different object types on the interface may also be important – a third factor

Tradeoff

- A more *appropriate* relationship should consider these factors

$$AS = f(number, type, layout)$$

We might frame an **even better** research question

Tradeoff

RQ4: *How the aesthetic score (in a scale of 1-10) depends on the number of objects, object types and the layout of the objects an interface has?*

Tradeoff

- We do not know which rating scale is the best
- Neither we know how many interface features influence aesthetics
- Therefore, we cannot say answer to RQ4 will lead to **the relationship**

Tradeoff

- Relationship is between a rating (representing a specific behavior by a specific group of users) in a pre-defined scale and the three quantities (number of objects, object types, object layout) representing specific features of an interface that influence aesthetics

Tradeoff

- This is in contrast to RQ1, which encompasses everything (and thus untestable since we do not know what to test)

Tradeoff

- However, if we can somehow get the answer to RQ1, we are supposed to get the **true** relationship

Tradeoff

- In scientific terminology, this is known as "validity" of the research question

Tradeoff

- The *extent* to which the observations made for a research question depends on the test condition is known as the "internal validity" of the question

Tradeoff

- The *extent* to which we can generalize the conclusions drawn from the observations is called the "external validity" of the question

Tradeoff

- A trade-off
  - We cannot frame questions that are based on generalized concepts (e.g. RQ1) - those are likely to be untestable
  - If we go for more specific questions (i.e., RQ4), we might get testable questions - however, we may not get the *true* answer

Tradeoff

- We can balance the trade-off by framing *multiple* testable questions

Tradeoff

- For example, let us assume there are three factors influencing interface aesthetics: number of objects (N), object types (T) and object layout on the interface (L)
- We are not sure which rating scale to use to record user judgment (since we do not know)
- We can actually frame three testable questions instead of one

Tradeoff

RQ4 (as before): *How the aesthetic score (in a scale of 1-10) depends on the number of objects, object types and the layout of the objects an interface has?*

RQ5 (modified form of RQ3): *How the aesthetic score (in a scale of 1-3) depends on the number of objects, object types and the layout of the objects an interface has?*

RQ6: *How the aesthetic score (in a scale of 1-5) depends on the number of objects, object types and the layout of the objects an interface has?*

Tradeoff

- We captured three different judgment behaviors in terms of the three rating scales in the questions
- We now perform empirical research for each separately and find out **three relationships**
- These relationships can be used to conclude about **the relationship -** not possible with *any one* of the research questions

Tradeoff

- There is a *positive correlation* between the testable questions and the untestable question
  - We are likely to arrive at a generalized answer for an untestable question from the specific answers to multiple testable research questions

Tradeoff

- A better approach than having **only untestable question and user feedback**

Basic Idea

- "Testable research questions" are more popularly known as "research hypothesis" in the domain of behavioral research

Basic Idea

- We start with two hypotheses: *null* hypothesis and *alternative* hypothesis
  - Both originate from same testable research question

Example

- We can frame two *hypotheses* from RQ4
- **H_0**: *The aesthetic score (in a scale of 1-10) does not depend on the number of objects, object types and the layout of the objects an interface has.*
- **H_1**: *The aesthetic score (in a scale of 1-10) depends on the number of objects, object types and the layout of the objects an interface has.*

Example

- We are no longer posing any question - ‘?’ at the end is gone
- Apart from that, there is one important difference - a single question gave rise to two hypotheses

Example

- In $\mathbf{H_0}$, called the **null** hypothesis, we are essentially stating that the test condition is not going to affect the outcome (judgment)
  - Typically, opposite to what we set out to establish (effect of the test condition on the observations)
- $\mathbf{H_1}$, called the **alternative hypothesis**, is just the opposite - we are stating that test condition does affect outcome

Note

- In an empirical research, we aim to find *statistical evidence* to *refute or nullify* null hypothesis and *support* alternative hypothesis

Note

- Hypotheses are relevant in the context of design evaluation

Example

- Reconsider hypotheses from RQ4
  - **H_0**: *The aesthetic score (in a scale of 1-10) does not depend on the number of objects, object types and the layout of the objects an interface has.*
  - **H_1**: *The aesthetic score (in a scale of 1-10) depends on the number of objects, object types and the layout of the objects an interface has.*