6.8 User Studies in IR

Nicola Ferro (University of Padova, IT), Norbert Fuhr (Universität Duisburg-Essen, DE), Kalervo Järvelin (University of Tampere, FI), Noriko Kando (National Institute of Informatics – Tokyo, JP), and Matthias Lippold (Universität Duisburg-Essen, DE)

The goal of information retrieval (IR) is to best serve a user information need by presenting him/her with a list of documents (information objects) potentially relevant to this need. This calls for specific evaluation methodologies which take into account the user, since determining the quality of a produced ranking, i.e. the effectiveness of a system, is directly depending on the user notion of what is satisfactory for his/her information need.

This setting is quite different from what we have, for example, in databases, where queries are exact and the correctness of results is not an issue, putting the emphasis on efficiency rather than effectiveness.

Therefore, it becomes central to understand what reproducibility is and how it can be achieved when users are in the loop.

6.8.1 Methodological Background

6.8.1.1 Experiments in psychology

The knowledge acquired in psychology is based on empirical results of experiments. An experiment is a research method in which one or more independent variables (IV) are manipulated to determine the effect(s) on a dependent variable. Other relevant factors need to be controlled in this setting. For instance, in the case of a user experiment in information retrieval, the independent variable could be a different search algorithm and the dependent variable could be the time to finish the search.

Psychological experiments needs to fulfil three criteria: **validity**, **objectivity** and **reliability**.

6.8.1.2 Validity

They need to be valid, which it is when the measures what it claims to measure is really measured. A problem could be that some participants might not be paying attention during the experiment, because of a lack of motivation. In some cases a manipulations check, which tests the attention of the user can be useful.

6.8.1.3 Objectivity

Objectivity is also important. An experiment has to be objective in two ways, the result of the experiment should not be influenced by the experimenter and that the interpretation of the data should not depend on the examiner.

6.8.1.4 Reliability

An experiment has to be reliable. When you repeat your experiment or another person repeats your experiment should come to a similar result. To ensure reliability, scientists have to specify their experimental design, they have to describe the conditions, under which the experiment is conducted and share information about the participants. The material and the raw data of the experiments needs to be stored and shared on demand by the corresponding author.

6.8.1.5 Reproducibility crisis in psychology

In a recent study (Open Science Collaboration, 2015) the results from 100 experiments from four top journals could just be partially replicated. That started a big discussion about the reasons.

6.8.1.6 Reasons for failed replication

Theoretical reason can be in the theories selection itself. If you have an **ill-defined theory**, which does not specify the outcome of the experiment and you use the result of the experiment as evidence for you theory, then the result did not matter and most likely can not be reproduced. For the IR experiments it might be necessary to define for which population the tools are produced and if the result can be generalized for all possible users. Older people might use the search engines in a different way than students do, which usually are the participants of the experiments.

Another theoretical threat are post theories and **post hypotheses or predictions**. If the hypotheses and the theoretical background are selected after the result of the experiment is known, you can not claim that you knew before. When this is happening the probabilities and the p-values are wrong.

Concerning the methodology, this is also a problem in psychology. Researchers rely almost exclusively on the p-value and **do not consider the effect sizes**, which are more important. The question in IR should not primarily be, is there a difference, but how big is the difference and would the user actually notice this difference. Furthermore, a lot of experiments are conducted with **low statistical power**, so the effect in this kind of experiment might not be the real effect and a replication can not find this result.

6.8.2 Context of User-oriented IR Evaluation

In IR, we have different kinds of user studies:

- laboratory experiments, where users are observed in the lab
- in situ observation of users at their workplace
- living labs, where the researcher analyses the system logs and possibly also manipulates the system employed by the users for their daily work.

Besides these types of experiments, there are studies that focus mainly on data collection methods, for which the discussion below only partially applies:

- exploratory user studies,
- focus groups, where researchers interview users
- longitudinal studies of users.

For discussing the reproducibility issues for the specific case of user studies, we follow the PRIMAD model (see Section 6.1) described above:

- Research objective is the research question to be addressed. In most cases, this part should also include the hypotheses to be tested with the experiment described in the remainder of the research paper.
- Model relates here to the experimental settings, which are used for testing the hypotheses specified before, So, besides the type of study, also the relevant aspects of the settings that refer to the research objectives are part of the model

- Implementation and Platform correspond here to the environment in which the study was carried out. Besides the system used for the study, also the group of users participating in the study as well as the exact conditions under which they participated belong to this aspect.
- Actor is the experimenter. In cases where the experimenter has direct contact with the users, the actor might have influence on the results of the study. Thus the actor should be kept constant throughout the study
- Data has a twofold meaning in user studies. First, there is the data that comprises the so-called testbed, like the document collection, the tasks carried out by the users, etc.. Second, there is the observation data collected throughout the study (thus, the user is regarded here as a data generator))

For enabling reproducibility, a researcher should share this context with other users to the maximum extent possible. Research objective and method are usually described in the research paper. In the past, the main research objective was the effectiveness of the methods investigated. Nowadays, also other aspects are considered, which are either more closely related to the actual user task, or to more subjective factors such as user satisfaction or engagement (which, in turn, can be measured via different variables). The more factors are considered, the more it becomes important to state the research hypotheses before actually carrying out the study, in order to achieve statistically valid results.

The environment usually can only be shared partially (mainly the system), while most other aspects (e.g. the users, the hardware, etc.) should be described at a reasonable level of detail in order to ease reproducibility. The same holds for the actor.

For the data, sharing testbeds is widely accepted nowadays, since the state of the art does not allow yet to characterize testbeds to such an extent that an independent researcher would be able to create a comparable testbed that could be expected to give the same results. The observation data, on the other hand, is essential for verifying the claims of a research paper. To a limited extent, it also can be used for simulation studies, depending on the degree of interactivity involved in the study (in classical IR experiments, the only data of this kind are relevance judgments).

6.8.3 Barriers/Obstacles

The research on Information Retrieval (IR) using computer started in 1950s and is said that IR is the first area in computer science using the human judgement as a success criteria of the technology [1]. This makes IR interesting and complex, and therefore the IR community has a strong tradition on evaluation to cope with how users incorporated in the experiment and testing, and make the reasonable comparison across the systems and the algorithms in the same system. Moreover, the commercial online search services started in 1960s and then the issue of working with real users in an interactive environment came up.

Since the Cranfield project in early 1960 [2], researchers constructed and shared testbeds called “test collections” which consist of the three types of data: document collections, the set of search requests, and the static set of human relevance judgment for each search request on the document collection. Such re-usable static testbeds were shared by the community as an infrastructure for the comparative evaluation and as one of the major elements for reproducibility of the experiments.

However, the technology and the society evolved tremendously: interactive online search for various purposes by ordinary persons became pervasive in everyday life, the various data collections including various web services and the social media are enhanced, search on multiple devices for multi-tasking become more common. The traditional evaluation paradigm (based on the batch-style one-time judgments)can not cover all the problems in the IR research and we are facing various new challenges and obstacles to make the research reproducible:

For studying users in interactive IR, there are various barriers and obstacles for reproducibility:

- Privacy/limitations of anonymizing
- Confidentiality
- Volatility of data (live streams, when the same situation never happens again, etc)
- Validity of the data: the data is so multidimensional that it is difficult to ensure the external validity of the experiment. This complexity is present also in IR test collection, and even more if we consider dynamic test collections.

Online web services are generally based on algorithms using user behaviour data in some way. This data is intrinsically rich in privacy and often includes confidential information. With interactive research IR systems, the situation is similar. Although various research efforts have targeted anonymization, there are still limitations, and these make it difficult to release the user behaviour data for external research groups, which, in turn, hampers reproducibility.

Large-scale users logs are generated in commercial search services and substantial studies on modeling and predicting users behaviour have been conducted based on these data, but again, the underlying data is not accessible for other researchers. Not only user modeling studies, but also various operational search mechanisms exploit user behavior data in the search and ranking algorithms, thus making it difficult to reproduce these methods.

To tackle the problems of the document data with privacy information and/or copyright problems, various evaluation-as-a-service approaches have been proposed and some of them were implemented successfully. However, these are still not sufficient for all the data produced by the users in-situ and lab environments.

For volatility, IR experiments can be conducted on live streaming data or commercial search services, in which the data and algorithms are continuously changing and the same data will never obtained again. Also, user experiments can not be "re-run" with the same users as the users learn from the previous experience.

In IR, interactivity and user behaviour or search experience through whole search sessions (or sometimes even a task involving multiple sessions) become more important, in order to consider real-world contexts. Various algorithms and softwares to support such interactions have been studied and proposed. The data obtained from the users in such task-based or whole session-based studies are highly complex, comprising e.g. the nature of the tasks conducted as well as the characteristics of each user. More research is needed for developing a framework that is able to describe such complex, multi-dimensional data as well as for devising methods for proper scientific analysis of the data collected.

6.8.4 Actions to Improve reproducibility

Actions to improve reproducibility of user-oriented experiments include checklists for authors (and reviewers, editors, chairs, etc.), sample exemplary papers, method inventories, extended methodological sections in papers, and critical discussions on the components/tools/other data used. These are considered briefly below:

Checklists should be provided on the methodology applied in the study. Kelly [3] is a useful source for constructing a checklist for user-oriented IR studies. Examples of items to check are:

- Research questions and experimental design (latin square, intra/inter subject, etc.)
- Participant characteristics and the population they are claimed to represent
- Methods of data collection, including the experimental protocol, environmental conditions, and variables used in the study (how to describe, how to measure, operational definition, observables)
- Experimenter
- Retrieval systems and their interfaces
- Methods for data analysis, including assumptions of statistical analyses (and adjustments if assumptions are not met)
- Degree of control on the system by the experimenter

Exemplary papers representing various types of user experiments could be offered in some community-based repository and annotated for their strong features, see also next section.

Inventories of typical variables various types of user experiments and standard ways to operationalize and measure them in different (sample) study settings could be provided by the community), as further discussed below.

Methodological sections could be emphasized in document templates, author guidelines and review guidelines. More space might be allocated to these sections and or authors encouraged to provide methodological appendixes or technical reports.

Finally, authors could be encouraged to *critically discuss* how suitable the set of tools and collections is to answer the research questions, what claims can the tools/collection support, describing the generalizability of the findings on the basis of the tools/collection that have been used.

6.8.5 Community Support to Reproducibility

In order to embody the vision described above and foster reproducibility in user-oriented studies, the involvement of the research community is crucial and it should consist of two complementary actions:

1. Support to the creation of shared resources;
2. Taking up and implementation of shared practices.

When it comes to *shared resources*, we can foresee several examples of them:

- **Inventories:** in order to streamline the reproducibility process, there is a need for catalogues accounting for the most appropriate experimental designs, the kind of independent and dependent variables you typically encounter in these settings, how to describe and measure such variables, the proper data analysis methodologies and statistical validation methods to apply to these variables in the different experimental designs, and so on;
- **Do's and don'ts:** in order to facilitate the understanding and adoption of the above facilitators of reproducibility, real and hands-on examples of appropriate and inappropriate ways to carry out user-oriented experiments are needed to clearly explain why a seemingly appropriate experimental setup is or is not working as expected. This could be partnered with a selection of exemplary and well-known papers, which should be annotated and enriched with links and explanations related to the above inventories, in order to clarify the researcher how and when to apply a given approach by means of concrete and remarkable case studies;
- **Repositories:** the adoption of shared repositories to gather collections of documents, interaction data, tasks and topics, and more is a key step to extend the reach of reproducibility in user-oriented experimentation;

- **Data formats:** the development of commonly understood and well-documented data formats, which can be extended to specific needs, as well as the introduction of proper metadata (descriptive, administrative, copyright, etc.) to model, describe, and annotate the data and the experimental outcomes is a crucial factor in lowering the barriers to reproducibility in user-oriented experimentation.

The methodological instruments, the checklists, the critical discussions, the different kinds of shared resources previously described are all key "ingredients" for successfully reproducing user-oriented experiments but the actual catalyst is the systematic and wide adoption by the community of *shared practices*, effectively exploiting all of these "ingredients", as also discussed in Sections 6.4 and 6.7.

References

1 W. B. Croft. Information retrieval and computer science: an evolving relationship. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2–3. July 28 – August 1, 2003, Toronto, Canada.

2 C. Cleverdon. The Cranfield tests on index language devices. Aslib Proceedings 16, 6:173–194. 1967.

3 D. Kelly. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1–2), pp. 1–224. 2009.

7 Open problems

7.1 Open Research Problems in Reproducibility

Carole Goble (University of Manchester, GB) and Daniel Garijo (Technical University of Madrid, ES)

When referring to reproducibility, we can distinguish two main types of research agendas, each with their scope and social implications. There is a **macro research agenda**, which consists of the topics of interest of the main funding agencies, and a **micro research agenda**, which would consist of the particular topics for new PhD students. While the macro agenda is influenced by the political tendencies of the moment, the micro agenda is influenced by the particular interests of researchers. Reproducibility initiatives may work fine for specific domains, but they may collapse when applying them at a macro level. Since most of the people in the group did not belong to funding agencies, the discussion focused on the micro agenda.

Regarding the social implications of reproducibility, an agenda should be issued in terms of productivity. Reproducibility can be seen as an investment for productivity, and part of its agenda should study and make explicit the correlation between these two features. Another challenge is addressing how the quality and quantity of the research work is affected by reproducibility. Currently, when given the opportunity, a researcher will choose to publish two publications rather than a highly reproducible one. **It is important to be able to show the long term value of high quality reproducible work**.

Another important aspect to take into consideration is the analysis of infrastructure, which includes the improvement of record keeping. The best way of holding trusted resources is to convince institutions to get involved. **Labs, companies and people are temporary,**