Inf Retrieval (2011) 14:204–207
DOI 10.1007/s10791-010-9149-1

BOOK REVIEW

Diane Kelly: Methods for evaluating interactive information retrieval systems with users

Foundation and Trends® in Information Retrieval, vol 3, nos. 1–2, pp. 1–224, 2009, ISBN: 978-1-60198-224-7

Hideo Joho

Published online: 3 November 2010

Methods for Evaluating Interactive Information Retrieval Systems with Users by Diane Kelly is the first textbook to extensively focus on the methodologies of user studies in Information Retrieval (IR). This 224-pages book consists of 15 chapters covering a range of topics from the development of cognitive aspects in IR, to the foundation of research, to IR test collections, to the ethical issues of user studies.

Chapter 1 opens the book by succinctly describing the problem in learning how to carry out a user study with interactive IR systems, that is, the experimental procedures are often not documented in the literature. Then Kelly reminds us that IR systems are fundamentally interactive and they should be evaluated from a user's perspectives. Also, it was pointed out that IR has lacked the literature on the research methods and experimental designs when compared to other disciplines. This highlights the value of this book. Chapter 1 also contains a number of recommended readings on related topics in IIR and concludes with the outline of the book.

In Chap. 2, Kelly presents a model to define Interactive Information Retrieval (IIR) using a continuum where one end has system focused studies (TREC-style studies) and the other end has human focused studies (Information-Seeking Behavior in Context). In between, several types of studies are placed along with the references. This continuum itself can be quite useful for those who are not familiar with IIR. An archetypical IIR study was described as evaluation of a system or search interface feature developed to support not only relevant document finding but cognitive activities associated with search. A holistic approach to evaluation based on system performance, interaction, and usability was described as another characteristics of the archetypical IIR study.

Chapter 3 provides further background to the theme of the book. The first aspect is the historical development of cognitive perspectives in IR research. For instance, De Mey's *cognitive viewpoint in IR*, which was proposed in 1977, is introduced as the first significant alternative to system or algorithm viewpoint in IR. The second aspect is the description of TREC tracks that were user-oriented such as Interactive Track, HARD Track, and ciQA (complex interactive QA) Track. Readers will significantly benefit from Kelly's experience

H. Joho (✉)
Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Japan
e-mail: hideo@slis.tsukuba.ac.jp

as the frequent participant as well as the organiser in TREC for an excellent overview of the historical development in TREC's user-oriented tracks.

The next three chapters describe the foundation of research, namely, research approaches (Chap. 4), research basics such as hypotheses, variables, and measurement (Chap. 5), and experimental design (Chap. 6). When reviewing submissions in conferences and journals, it is relatively easy to see whether or not the authors have solid understanding of these foundational aspects of research. Therefore, these basics are crucial not only for conducting user studies but for *presenting* your work. There are textbooks which cover the research basics, but many IIR-related examples used in these chapters will allow you to contextualise the significance of such aspects in user studies.

Chapter 7 discusses sampling. It describes the characteristics and techniques of different types of sampling. It also discusses the recruitment of participants, and terminology related to users, subjects, and participants. This chapter is slightly weak in my opinion. This is not because sampling is insignificant in IIR research, but the significance of sampling becomes only evident in a later chapter on data analysis. Also, there is a relatively lengthy description of a sampling method which is not always applicable to IIR studies. The discussion on the terminology looked slightly detached from the rest of the chapter, although I could understand why it was there.

After several chapters on relatively generic research topics, the book returns to the IR-centred issue such as test collections in Chap. 8. Those who have experience on user studies but not in IR will find this chapter particularly useful. The first half of this chapter describes the basic components of IR test collections in the context of TREC and different types of corpus such as newswire texts, web, personal collections. The second half describes tasks and topics in IR. It also discusses the pros and cons of artificial information needs and natural information needs. Simulated work tasks are also introduced in this chapter.

Given that we managed to design an experiment and decided what corpus and tasks to use, the next question would be how to collect data during the experiment. This is discussed in Chap. 9. A number of data collection techniques such as Think-loud, stimulated recall, prompted self-report, observation, logging, questionnaires, and interviews are described and their advantages and disadvantages are discussed. Of those, logging and questionnaires have more focus than the others. In particular, it emphasises the difference between server-side logging and client-side logging, and the impact of question designs on participants' subjective assessments.

The next two chapters are the longest chapters in this book. Chapter 10 (27 pages) looks at measures, and Chap. 11 (50 pages) focuses on data analysis. As the volume indicates, readers are not expected to fully understand the contents of these two chapters in the first time, especially when you are not familiar with IIR or statistical tests.

Chapter 10 is a great place to appreciate the richness of IIR evaluation. Kelly argues that four basic classes of measures have emerged in IIR over time, namely, contextual, interaction, performance, and usability. This chapter gives a detail discussion on the measures for each of the four classes. There are also links to the data collection methods described in the previous chapter. Not all measures introduced in this chapter are well-known in IIR, but you might find an appropriate measure to answer your research question from this wealthy list of measurements.

Chapter 11 describes qualitative and quantitative data analysis with an emphasis on the latter. Selection of appropriate statistical tests depends on many factors of collected data such as the experimental design, normality of population, distribution of sample data, number of things to compare, number of measures to use, sample size, etc. Understanding

exactly how each of statistical tests works and why is even more difficult. Like the review on Chaps. 4–6, there are many textbooks on data analysis, but the examples on such books are not necessarily applicable to IR. The IIR examples used in the chapter will give you a better idea of when and how to apply a particular method to your study. The IIR community is still generous about the assessment of statistical tests reported in papers. Although we are not a statistician, each of us will need to be more conscious about statistical tests to improve the situation.

The concepts of validity and reliability are introduced in Chap. 12. Different types of validity and its trade-off relationship with reliability were discussed in detail. An important emphasis here is that our selection of a certain method, instrument, and protocol is closely related to the validity and reliability of the research and its findings. It was also pointed out that there are some established measures for the two attributes of research in other disciplines that can be potentially useful for IIR studies.

Chapter 13 discusses human research ethics. It was emphasised that although physical harm to participants of IIR evaluations is less of issues when compared to medical studies, this should not be a reason for being ignorant of ethical issues of research. This section explains the role of institutional review boards, ethical principles, and how the ethical issues could emerge in IIR studies. For example, participants' activities could be captured by visited websites through cookies, personalised photo collections might contain the picture of other people, and long-term strategy for the management of experimental data.

This chapter could potentially discourage people to perform user studies, because ethical issues are much less of an issue in system-oriented research. However, I was glad to find this chapter. The check-points suggested in this chapter can increase your confidence in designing a user study that is more privacy-aware and socially acceptable.

A number of opportunities for IIR research was discussed in Chap. 14 as outstanding challenges and future directions. In this chapter, Kelly introduced many variant IR problems such as multimedia search, personal information management, XML retrieval, Question Answering, cross-language retrieval, personalised search, collaborative IR, and mobile search as the areas where there is room for IIR to make significant contributions.

The development of sharable collections was also raised as an important future direction in IIR. This not only includes the document collections but also sharable search/work tasks. An equally significant issue was an infrastructure to share some of our experimental data. Individual user studies can afford only few participants, but it often results rich annotated data which can be great complement to search engine logs. We need a repository for interaction data.

The last topic in this chapter was measures. Kelly reminds us the dynamism of relevance and iterative searches in interactive studies, and encourage us to further develop the evaluation measures which better incorporates these two fundamental attributes of IIR. A recent trend of using indirect measures such as physiological sensors was also discussed in this chapter.

Finally, Chap. 15 concludes the book by emphasising the importance of diversity in the approaches to future development of IIR evaluation.

In conclusion, it was a great pleasure to review the book on IIR evaluation. As Kelly mentioned in the last chapter, an evaluation of systems with users is a complex process and everything cannot be covered by a single book. Therefore, readers might find some parts of this book insufficient to fully understand and need to look elsewhere. Nevertheless, this is the book you should read first if you are interested in the evaluation of interactive IR systems with users.

The proportion of researchers who work in IIR is still limited in the IR community when compared to system-oriented research. Given that IR is often a core subject in Computer Science, one thing we can do is to attract more young CS researchers, who do not necessarily have extensive training of experimental design, to involve in IIR research. I hope that this book will encourage more researchers to conduct user studies in IR and to further develop its evaluation methodologies. It is difficult to carry out a user study without any mistakes, but this book will help you reduce the number of errors significantly.

Finally, I wish to see more books written on this topic to facilitate innovative work in IIR.