Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 4–21

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work

Ann Blandford [a,*], Anne Adams [a,1], Simon Attfield [a], George Buchanan [a,2], Jeremy Gow [a], Stephann Makri [a], Jon Rimmer [b], Claire Warwick [b]

[a] UCL Interaction Centre, University College London, Remax House, 31-32 Alfred Place, London WC1E 7DP, UK
[b] School of Libraries, Archives and Information Studies, University College London, Gower Street, London WC1E 6BT, UK

Received 7 April 2006; received in revised form 28 December 2006; accepted 15 January 2007
Available online 6 March 2007

Abstract

The strongest tradition of IR systems evaluation has focused on system effectiveness; more recently, there has been a growing interest in evaluation of Interactive IR systems, balancing system and user-oriented evaluation criteria. In this paper we shift the focus to considering how IR systems, and particularly digital libraries, can be evaluated to assess (and improve) their fit with users' broader work activities. Taking this focus, we answer a different set of evaluation questions that reveal more about the design of interfaces, user–system interactions and how systems may be deployed in the information working context. The planning and conduct of such evaluation studies share some features with the established methods for conducting IR evaluation studies, but come with a shift in emphasis; for example, a greater range of ethical considerations may be pertinent. We present the PRET A Rapporter framework for structuring user-centred evaluation studies and illustrate its application to three evaluation studies of digital library systems.

Keywords: Digital library; Usability evaluation; HCI; Case study

1. Introduction

One of the priorities in setting up any evaluation project is to choose appropriate evaluation techniques and construct a plan of the evaluation. Within the Information Retrieval (IR) tradition, there are some well established approaches to evaluating the performance of retrieval algorithms (e.g. Tague-Sutcliffe, 1992) and, more recently, there has been an increasing focus on user-oriented evaluation criteria and methods for evaluating IR systems within the context of user–system interaction (Interactive Information Retrieval) (e.g. Borlund, 2003). Important as these evaluation criteria are, they continue to focus largely on algorithm evaluation; there are

[*] Corresponding author. Tel.: +44 20 7679 5288; fax: +44 7679 5295.
E-mail address: A.Blandford@ucl.ac.uk (A. Blandford).

[1] Present address: Institute of Educational Technology, The Open University, Milton Keynes, MK7 6AA, UK.

[2] Present address: Department of Computer Science, University of Swansea, Singleton Park, Swansea, SA2 8PP, UK.

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.01.021

other criteria that need to be considered if IR systems are to be truly useful within the context of users' broader activities. People using IR systems are most commonly retrieving information in support of some larger task such as writing a news article or an essay, preparing a legal case or designing a novel device. Evaluating a system in terms of its fitness for purpose in this broader sense demands a different approach to evaluation from the methods that have become established within the IR tradition. In this paper, we present a framework for planning user-centred evaluation studies that set systems within the context of information work. We illustrate the application of the framework to evaluations of various digital libraries (DLs).

Digital libraries are coming into widespread use to support information work. While DLs are not simply "IR systems", they are an important class of systems within which IR algorithms are routinely implemented, and effective information retrieval is one essential feature of DLs. DLs typically bring together various subsystems to deliver information access and management facilities for users. There is no agreed definition of what a DL is; as Fox, Akscyn, Furuta, and Leggett (1995, p. 24), note, "The phrase "digital library" evokes a different impression in each reader. To some it simply suggests computerization of traditional libraries. To others, who have studied library science, it calls for carrying out of the functions of libraries in a new way". What matters for the purpose of this paper is that DLs are systems that enable users to retrieve information, and that they can be evaluated in terms of how well they address users' needs.

Just as the term "digital library" is used in different ways by different people, so the term "evaluation" is interpreted in different ways by different communities. Within the IR community, evaluation is most commonly summative – that is, the outcome of an evaluation is summative measures (e.g. of precision and recall) of how "good" a system is. Within the Human–Computer Interaction (HCI) community, evaluation is more commonly formative – that is, the outcome of an evaluation is a description of how users interact with systems that highlights ways in which those systems could be improved. Formative evaluation can consider the "system" at various levels of granularity; as discussed more fully below, we take an inclusive view of evaluation as covering various aspects from details of implementation through to understanding how computer systems support work in context. The work reported here is based on the formative approach. The contrasts between these approaches are discussed below.

2. Background

To set the work on PRET A Rapporter in context, we consider evaluation from three different perspectives: the evaluation tradition within IR; the evaluation tradition within HCI; and approaches that have been taken to evaluating digital libraries. In doing this, we compare the evaluation cultures, to identify strengths and limitations of each, and show how they have influenced the evaluations of DLs.

2.1. An overview of IR evaluation

The classic approach to IR evaluation is the "Cranfield paradigm", within which as many variables as possible (including the database of documents over which retrieval is to be performed and the set of queries) are controlled in order to measure algorithm performance on criteria such as precision and recall. Tague-Sutcliffe (1992) presents a detailed, and highly cited, methodology for conducting an evaluation study within this paradigm.

Tague-Sutcliffe (1992) highlights three criteria that any evaluation study must satisfy:

- Validity, which she defines (p. 467) as "the extent to which the experiment actually determines what the experimenter wishes to determine". She highlights inappropriate measures (e.g. using a Likert scale to measure user satisfaction) and user populations (e.g. student users to represent professionals) as possible causes of low validity.
- Reliability, which she defines (p. 467) as "the extent to which the experimental results can be replicated" – typically by another experimenter.
- Efficiency, or "the extent to which an experiment is effective" (p. 467) relative to the resources consumed. This is an issue that is explored further by Toms, Freund, and Li (2004).