Evaluating Adaptive Systems and Applications is often Nonsense

Paul De Bra
Dept. of Math. and Computer Sciece
Eindhoven University of Technology (TU/e)
Eindhoven, the Netherlands
p.m.e.d.bra@tue.nl

ABSTRACT

In the research field of User Modeling, Adaptation and Personalization there is a strong focus on comparative evaluation. In this discussion paper we as ourselves when it makes sense to perform such evaluation and also when it is complete nonsense. We argue that especially for adaptive *systems* (and not *applications*) the typical comparitive evaluations with groups of end-users make no sense. We also argue that for applications it is difficult to perform a meaningful evaluation because it is hard to find something to compare the (use of the) application with.

CCS Concepts

•Human-centered computing → Usability testing; Empirical studies in HCI;

Keywords

usability, comparative evaluation, systems versus applications

1. INTRODUCTION

Adaptive (web-based) hypermedia [2, 8] is being used for many data-driven web-based services (like YouTube, Facebook, Amazon, etc.) and for specific expert-driven applications like museum guides (e. g. the Rijksmuseum CHIP demonstrator [1]) and on-line course texts created using e. g. Interbook [3], AHA! [7] or GALE [11, 10]. As the references show several papers describing adaptive systems have been published at the ACM Hypertext conference instead of at UMAP. So called "systems papers" have always been somewhat problematic to publish because the typical empirical evaluation with groups of end-users makes no sense. Throughout the years we can observe that the UMAP research community seems to have a strong preference for research on methods and applications and has difficulty in handling papers that merely present a new "platform".

A second issue we address in this paper is that when considering a specific application, for instance an on-line course text, it is unclear how the benefit of adaptation can be evaluated, as it is hard to do a fair comparison between applications.

2. EVALUATING ADAPTIVE SYSTEMS

When adaptive applications were introduced, mainly in the early 1990s, they were closely integrated with the technology used to run the application. A good example is ELM-ART [5], an on-line Lisp tutor which was called "an Intelligent Tutoring System" even though it was really an Intelligent Tutoring Application. ELM-ART inspired new developments, but later systems like Interbook, AHA! and its successor GALE were all realized as "platforms" in which an author could/can create adaptive on-line courses. The distinction between the application, e. g. an on-line course, and the system that makes it possible to deliver the application, is essential: end-user evaluation of the system makes no sense whereas end-user evaluation of the application may make sense but is very different.

When we first presented AHA! we were often asked "How good is the adaptation provided by AHA!?" and our standard answer: "The quality of the adaptation depends entirely on what the author of an application supported by AHA! creates." was never considered a satisfactory answer. Yet, it was and still is the only possible answer.

For the UMAP community it is of vital importance that *generic* or *general purpose* adaptive systems are developed so that researchers who wish to experiment with adaptation and with applications such as on-line courses can concentrate on the core of their research without the need to also develop underlying technology that makes the adaptation they need possible. At the same time, the community is not reluctant to accept papers describing new platform developments because such papers cannot contain an end-user comparative evaluation without considering also an application running on the platform, and actually evaluating the application, not the system.

Interbook could be used to develop and deliver on-line courses about very different topics, but all being presented in the same presentation style, using the same adaptation strategies. AHA! and later GALE went further: they allow the definition of arbitrary rules for user modeling and adaptation, allow for the conditional inclusion of fragments and objects in the presentation, and allow for the use of arbitrary presentation styles, arbitrary layout, arbitrary choice of link annotation, etc. As a result, the question "How good is the adaptation provided by AHA! or GALE?" is nonsense. The systems provide the adaptation an author defines. AHA! and GALE applications can provide excellent adaptation that greatly helps learners find their way through an on-line course and AHA! and GALE applications can also completely mislead learners and make learning much harder than if there were no adaptation at all.

There is definitely a possibility to evaluate adaptive systems: by mapping the functionality of the system to existing reference models like AHAM [6] and GAF [8] the user modeling and adaptation power of different systems could be compared. And by taking performance measurements under (synthetic or real-world) load the

ability of systems to handle large numbers of users can be compared. None of these types of comparisons have gained acceptance as being "evaluation" by the UMAP community.

3. EVALUATING ADAPTIVE APPLICATIONS

The core question about adaptive applications is "Does adaptation help?". This question was for instance addressed in [4], considering adaptive link annotation in an Interbook application. What typically happens in such an evaluation is that a user group is divided in two subgroups; one subgroup gets to work with the adaptive application and the other subgroup gets to work with that application with the adaptive functionality turned off. A number of performance indicators are then measured, like how many navigation steps users make, how well they perform on a test, etc. The results are somewhat predictable: the users of the adaptive application perform better and are more satisfied than the users of the "crippled" adaptive application that had no adaptation. If an author creates an on-line (hypermedia) course text and cannot use any adaptation a lot of care will go into deciding where to place which links. Users who study a course page may have reached that page through many different paths. These users will have different knowledge and knowledge gaps. When the page can link to another (related) topic, the author needs to decide carefully whether to make that link available or not. In an adaptive application the author can place the link and the system will decide, based on the user's knowledge and on prerequisite relationships, whether that user will be recommended to follow that link at that moment in time. So it is likely that the course text will contain many links that will sometimes be recommended and sometimes not. Simply making these link recommended all the time does not give the application a fair chance in any comparison with the adaptive application.

Problems and pitfalls in the evaluation of adaptive applications have already been identified, for instance by Weibelzahl [12]. Even though the title of that publication mentions "adaptive systems" it is really more about applications. In [9] it is argued that separating the evaluation of different aspects (rather than brute force enabling or disabling all adaptation) can help to pinpoint where the adaptation helps or fails. The paper [9] provides a detailed description of a layered approach to the evaluation that makes it clear that proper evaluation of an adaptive application is a huge task, not something to describe in a short section at the end of a research paper. It should not come as a surprise that many evaluations that have been made of adaptive applications are have not been performed to such a rigorous standard and are actually closer to nonsense.

4. CONCLUSIONS AND DISCUSSION

In this paper we have shown that the typical end-user evaluation with groups of users using different versions of applications 1) cannot be used at all to evaluate adaptive systems or platforms and 2) that performing a proper evaluation is a major undertaking (see [9]) when one wants to avoid comparisons that make no sense.

Two interesting discussion topics for the workshop are:

- Since we really need to have generic platforms that can be used to perform UMAP research without the need for every researcher to create their own special-purpose platform we need to discuss criteria for assessing whether a paper describing the development of a generic system is acceptable. The current practice is that UMAP researchers publish system descriptions in other venues. An incentive should be created at UMAP to embrace "systems papers".
- In answering the "Does adaptation help?" question we should have clearer criteria for the comparative evaluation to avoid the pitfall of simply comparing an adaptive with a non-adaptive version and hoping the results are not nonsense. We should define some quality standard for the applications with which we compare. The layered approach published in [9], among others, may help us get started with setting that standard.

5. REFERENCES

[1] L. Aroyo, N. Stash, Y. Wang, P. Gorgels, and L. Rutledge. CHIP Demonstrator: Semantics-Driven Recommendations And Museum Tour Generation. In G. Schreiber and K. Aberer, editors, *The Semantic Web - ISWC/ASWC 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 879 – 886. Springer, 2007.

[2] P. Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1-2):87–110, 2001.

[3] P. Brusilovsky, J. Eklund, and E. Schwarz. Web-based education for all: a tool for development adaptive courseware. *Computer Networks and ISDN Systems*, 30(1-7):291 – 300, 1998. Proceedings of the Seventh International World Wide Web Conference.

[4] P. Brusilovsky, C. Karagiannidis, and D. Sampson. The benefits of layered evaluation of adaptive applications and services. In *In*, pages 1–8, 2001.

[5] P. Brusilovsky, E. W. Schwarz, and G. Weber. Elm-art: An intelligent tutoring system on world wide web. In *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, ITS '96, pages 261–269, London, UK, UK, 1996. Springer-Verlag.

[6] P. De Bra, G.-J. Houben, and H. Wu. Aham: A dexter-based reference model for adaptive hypermedia. In *Proceedings of the Tenth ACM Conference on Hypertext and Hypermedia : Returning to Our Diverse Roots: Returning to Our Diverse Roots*, HYPERTEXT '99, pages 147–156, New York, NY, USA, 1999. ACM.

[7] P. De Bra, D. Smits, and N. Stash. The design of aha! In *Proceedings of the seventeenth ACM conference on Hypertext*, page 133. ACM, 2006, adaptive version at http://aha.win.tue.nl/ahadesign/.

[8] E. Knutov, P. De Bra, and M. Pechenizkiy. Ah 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia*, 15(1):5–38, 2009.

[9] A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, 2010.

[10] D. Smits. *Towards a Generic Distributed Adaptive Hypermedia Environment*. PhD thesis, Eindhoven University of Technology, adaptive version on http://gale.win.tue.nl/thesis/, ISBN 978-90-386-3115-8, 2012.

[11] D. Smits and P. De Bra. Gale: a highly extensible adaptive hypermedia engine. In *Proceedings of the twentysecond ACM conference on Hypertext*, pages 63–72. ACM, 2011.

[12] S. Weibelzahl. Problems and pitfalls in the evaluation of adaptive systems. *Adaptable and adaptive hypermedia systems*, 11:285–299, 2005.