Prototype Testing for a Destination Recommender System: Steps, Procedures and Implications

Andreas H. Zins[a],
Ulrike Bauernfeind[a],
Fabio Del Missier[b],
Nicole Mitsche[a],
Francesco Ricci[b]
Hildegard Rumetshofer[c], and
Erwin Schaumlechner[d]

[a] Institute for Tourism and Leisure Studies
Vienna University of Economics and Business Administration, Austria
{zins, bauernfeind, mitsche}@wu-wien.ac.at

[b] ITC-irst
Electronic Commerce and Tourism Research Laboratory, Italy
{delmissier, ricci}@itc.it
[c] Institute for Applied Knowledge Processing
University of Linz, Austria
hildegard.rumetshofer@faw.uni-linz.ac.at
[d] Tiscover, Austria
erwin.schaumlechner@TIScover.com

Abstract

This paper describes and discusses the application of various state-of-the-art techniques to improve the design and usability of a web application in the B2C area. It is illustrated by the ongoing project of developing an intelligent destination recommender system (DieToRecs). These techniques comprise four particular evaluation steps: 1. a concept test, 2. a cognitive walkthrough, 3. a heuristic evaluation, and 4. an experimental evaluation by system users. Each section (i.e. evaluation step) addresses three areas of interest: a) the starting situation and objective, b) a short description of the applied method and the procedure, c) a brief summary of the results and the lessons learned for the general use of evaluative techniques.

Keywords: recommender system, usability, prototype testing, system evaluation

1 Introduction

At present the DieToRecs system is in the stage of a first prototype. Its development includes a number of characteristics distinguishing it from already existing recommender systems (Fesenmaier et al. 2003). First, DieToRecs is a destination advisory system based on CBR (case based reasoning). CBR is a methodology trying

to solve a problem by retrieving and using a similar, already solved case. A case is defined as any single user-system interaction history and consists of travel wishes and constraints, the travel plan (a bundle of items being interesting for the user and which is therefore collected), the user profile, and finally the outcome of one or more travel plans. Second, a collaborative filtering system, based on similarity of sessions rather than the classical correlation of votes, is employed enabling to reuse travel plans built by similar users. Finally, the system is a conversational tool meaning that the user should feel like interacting with a human being. Queries and suggestions follow successively to enable a vivid question and answer process. Interactive query management is employed to handle queries more efficiently. The system helps the user to redefine queries: they are relaxed or tightened in order to display a desired number of ranked results.

Following the concept of a user-centred usability engineering (Manhartsberger & Musil 2002) the system development was subdivided into four phases:

- Development of the decision model: collection of use cases and system features, development of a realistic model of the user decision process, including the modelling of destination.
- First prototype design and development: the key technologies and the decision model provide input for the design and the development of the first prototype, a fully operational recommender system with key components such as dialogue management based on the tourist decision model, similarity based queries, filtering using a user model, and user activity logging.
- Prototype management and evaluation: experimental tests determine the statistical and practical significance of the improvements brought by each single technique.
- Final recommendation system and framework: at this stage, log data collected with the first prototype will be used for user profile learning and for tuning the various filtering techniques implemented.

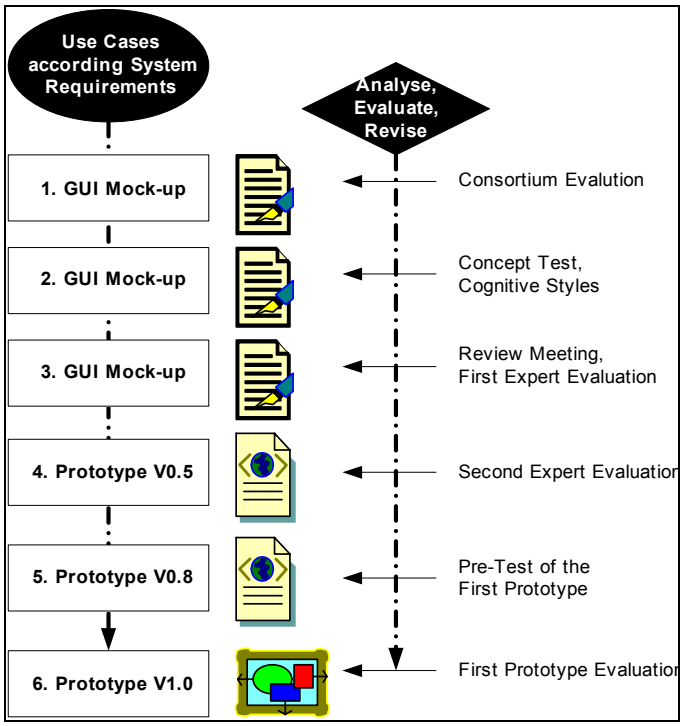

Fig. 1. Re-engineering process to build the GUI in DIETORECS

In order to achieve a really superior recommender system the software quality had to be reviewed and evaluated at the various developmental stages. With regard to quality standards we refer to the views of product quality in terms of the appropriate ISO/IEC norms and definitions (ISO/IEC 9261, FCS 9216-1, ISO 9241-11). These go beyond the ergonomic concept of usability and are aimed at improving the effectiveness, productivity and satisfaction a user perceives in a specific working environment (Bevan 1999). To cover a maximum of perspectives a series of steps of formative evaluation have been taken so far in addition to the continuous checks and adaptations undertaken by the developers of the DieToRecs system.

Figure 1 sketches the stages in the development process combined with evaluative milestones carried out to adjust the system functionality and usability to the requirements determined during the initial use case phase. The following sections of this paper present and discuss the evaluation steps 2, 3, 4, and 6: 1. a general concept test challenging two different interface options, 2. a cognitive walkthrough, 3. a heuristic evaluation, and 4. an experimental evaluation by system users. Each section (i.e. evaluation step) addresses three areas of interest: a) the starting situation and objective, b) a short description of the applied method and the procedure, c) a brief summary of the results and the lessons learned for the general use of evaluative techniques.

2 Concept test with a horizontal prototype

From literature review and an additional observational study the user model for the destination recommendation system was elaborated. One of the basic premises was that the system has to serve users with different decision styles (Grabler & Zins 2002). Therefore, a concept test (Dalgleish 2000) had been conducted in an early stage of the prototype development (Pearrow 2000). The purpose was to investigate the potential advantages of two effects of the human-computer interaction: 1. giving the user the choice to select among two alternative navigational options A (more sequential) and B (more pictorial-holistic), and 2. classifying the user in advance into one of two broad categories of cognitive styles (A': analytical and B': holistic) to direct her/him to the potentially more suitable navigational option. For the empirical test a horizontal prototype (Nielsen 1993; Rudd & Isensee 1994) or so-called demonstrator (i.e. a not yet fully operable system of related web pages to collect or identify the user's travel preferences and wishes) had been developed and presented to 176 test persons (internet users only). As the graphical user-interface was already in an advanced stage and responded to a limited set of keyboard and mouse inputs it can be seen as a high-fidelity prototype (Walker et al. 2002).

Results have been encouraging and in favour of offering two alternative system accesses: a classical interface with check boxes and structured input fields and another more holistic approach using sketches of already existing travel bundles for revealing someone's travel preferences and constraints. The analysis highlighted that users should be classified in advance to one of the corresponding cognitive styles. This is based on the observations that asking them in advance and let them choose anyway between two interface options leads to a substantial rate of a misleading self-selection. The probable consequences are a reduced user satisfaction and in the worst case a lost costumer.

For the further DieToRecs development it was decided to follow the encouraging direction of creating two different entrance gates: one for the more analytical, decompositional process of communicating someone's travel preferences and contingencies and another supporting a more holistic approach based on fuzzy cues and responses from which the user gets some inspiration for the concrete travel planning task. Still unresolved problems and areas are: 1. finding and applying well performing and not boring a priori classification instruments to detect the user's appropriate cognitive style, 2. testing multiple interface alternatives to better address the inhomogeneous audience 3. comparing the performance and user evaluation of competing fully functional recommender systems with alternative presentation and interaction designs as proposed by Rumetshofer et al. (2003).

3 Cognitive walkthrough and first heuristic inspection

A completely different kind of evaluation (scenario evaluation) was applied to a GUI mock-up (without functionalities) developed to allow an early qualitative assessment

of some user interface design choices (Nielsen 1993). The primary goal of this inspection was to detect substantial weaknesses of the user interface design. A single expert employed two techniques: cognitive walkthrough and heuristic evaluation. The former is a technique for evaluating the design of a user interface, with special attention to how well the interface supports "exploratory learning," i.e., first-time use without formal training (Rieman et al. 1995). According to Ivory and Hearst (2001) the expert has to simulate the user's problem solving. The latter is a technique to identify violations of heuristics (Ivory and Hearst 2001) proposed by Nielsen (1993) for a quick analysis of applications. The usability guidelines applied for this evaluation have been taken from Nielsen (2001) which have been adapted to define the following principles: (P1) know your user, (P2) reduce the cognitive work, (P3) design for errors, and (P4) keep the consistency (internal with your systems, P4int, and with respect to common practices P4ext, external).

Important improvements have been achieved following the focal critical comments on consistent labelling, navigational and menu aspects as well as design considerations. Changes resulting from the inspection were a rearrangement of the menus (new design, change of grouping, visualisation through icons, renaming to be consistent). The start page of the main area was unified with the menus and the registration process was simplified and better explained to the user. Resolution problems concerning the display of the interface were solved. Furthermore, a clearer presentation of the recommendation results was implemented and some inconsistencies in the use of terminology were eliminated.

Although the re-engineering process of the cognitive walkthrough is pretty tedious to perform, inconsistencies, and general and recurring problems could be missed, the method is appropriate in an early prototypical stage. In particular, it is possible to detect substantial weaknesses before a prototype is built. A general problem occurring with cognitive walkthrough and heuristic evaluation is the difficult position of the evaluator. He has to act as a user with the opinion of an expert, which leads to ambiguous roles.

4 Heuristic and standardized evaluation by experts

After the adjustments made based on the first inspection, a heuristic evaluation was carried out on the Prototype V0.5 (with functionalities). A major evaluation goal was to eliminate the major interface and interaction shortcomings prior to the experimental test. This step seemed to be necessary given the prototypical stage of the system. Lindgaard (1994) defines the heuristic evaluation as a detailed informal subjective usability analysis conducted by experts simulating the perspective of a typical end user. The evaluators do not follow a specific set of methods, rules or procedures; instead they rely on a set of vague guidelines. The subjective judgements of the experts are influenced by their experience and background. In addition to the cognitive walkthrough, the heuristic evaluation is an in-depth analysis collecting all occurred problems, from the highly serious to the most trivial. Due to the subjective

judgements and the missing structure of the heuristic evaluation, a standardized instrument was additionally employed to enable comparisons between the judges and to locate the actual development stage. The comparable evaluation was carried out using the Purdue Usability Testing Questionnaire (PUTQ). The questionnaire is composed of 100 questions on system interface structured by eight factors that are relevant to human-computer interaction. These factors were compatibility, consistency, flexibility, learnability, minimal action, minimal memory load, perceptual limitation, and user guidance. An essential advantage is the possibility to compute an index based on the ratings and put into relation to the possible perfect score (Lin, Choong & Salvendy 1997). The heuristic evaluation was carried out using five experienced interface judges (according to e.g. Lindgaard 1994, Galitz 2002). These evaluators had to:

- provide comparative judgements, rating the system on a variety of dimensions by the PUTQ;
- perform a detailed analysis of the general system functionality, of the interface, and of the user-system interaction aspects.

The heuristic evaluation results, quite a long list of modifications, were summarized in groups (start page, navigation, layout and design, travel planning process, recommendation process, and results) and sorted by their importance. The most critical issues were solved before the experimental user evaluation took place. A lot of changes were made. However, a few examples of detailed problems will serve for better illustration:

- the page expired too early and the pages loaded too slow
- budget range (to indicate possible travel expenses) was too small
- inconsistent use of the term “travel plan” (“travel bag” was used as well and created confusion).

Overall, the experts gave the system a good quality total grade (PUTQ Index: 65.0 – the higher the score, the better the usability result; scale of 100), especially with respect to perceptual limitations, compatibility and learnability. Deficiencies were identified in user guidance and flexibility which mainly results from functions not available due to the prototypical status. As implication s from the heuristic evaluation additional work should be invested in typical prototype troubles such as further extension of the database, help function and error messages. The recommendation process is one of the major focuses of further developments. In particular, as seen from the detailed heuristic evaluation and the PUTQ questionnaire, more effort has to be set into the development of wording and explanations to aid the user through the recommendation process. Another focus should be given on the presentation of the result pages when the user accesses an item detailed site, asks for more recommendations, browses her current travel plan, and enters into the “Searching for inspiration” gate (the holistic interface that allows the user to navigate the travel offers exploiting visual clues, see Figure 2).

The heuristic evaluation was an important step during the re-engineering process of the first prototype. The success and richness of the observations suggests another

detailed expert evaluation during the final evaluation of the second prototype. As far as the PUTQ as standardized questionnaire is concerned it turned out to be a quite valuable tool allowing comparisons because of the standardization. On the other hand, the questionnaire is tailored to general computer interfaces and therefore, some application problems on web-based systems arose. Furthermore, some web specific problem areas remain unconsidered. It is suggested to adapt the PUTQ Questionnaire, considering research about web usability indices (e.g. Keevil 1998, Harms et al. 2000), without giving up the comprehensiveness of the PUTQ.

5 Experimental evaluation by potential users

This – so far – final step of evaluating the re-engineering process of the first prototype was conceived to involve the end user of such a destination recommender system. Its major focus was on the innovative contributions that DieToRecs is supposed to generate which are essentially the recommendation functions. Hence, the main effort of this evaluation was indeed dedicated to the implementation of the recommendation functions which are supposed to be general enough to be integrated into a variety of specific applications (web sites). The system prototype (V1.0; see Figure 1) is a tool that allowed testing of these general functions in a more controlled yet flexible way; i.e. without having to cope with the engineering problems regarding the real integration of the recommendation components into an existing application.

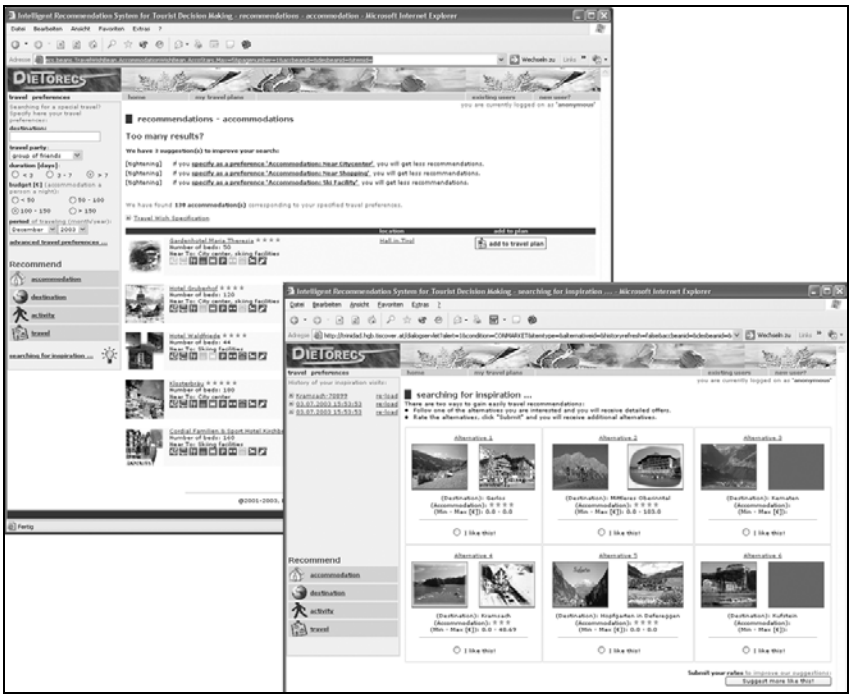

Fig. 2 Screenshots for a Feature-based Search Result and for "Seeking for Inspiration"

The approach of this experimental evaluation consisted of building a limited set of variants of the DieToRecs prototype to test hypotheses about the performance of the system on a set of dependent measures. The main hypotheses concerned the users' search and choice behaviour, and their satisfaction. They are stated as follows: The query management, the case-based ranking functions, and the other recommendation functions embedded into the DieToRecs system are able to:

- provide valuable recommendations for the user;
- help the user to construct better plans;
- enhance the efficiency of the interactive search and decision making processes involved in the plan construction;
- increase user satisfaction.

Three variants were to be tested:

- DTR-A: interactive query management only, i.e., supporting the user in case of query failures (too many or no result) but not using a case base of previously built travel plans and therefore not providing any recommendation support via sorting (Ricci et al. 2002);
- DTR-B: single item recommendation with interactive query management and ranking based on a case base of 25 cases extracted from the database of the Austrian National Guest Survey 1997/98;
- DTR-C: this variant allows a user to navigate among complete travel recommendations in a simple and effective way (starting from the link "Seeking for inspiration"). Six travel examples are shown at each page. Then the user is requested to provide a feedback on the presented alternatives in a simple form ("I like this" vs. "I do not like this"). Finally, the system updates the proposed alternatives by means of the feedback provided by the user, and the similarity-based retrieval in the case base is performed again.

The experimental user evaluation (Chin 2001) was based on two weakly structured trip planning tasks to a pre-specified geographical region. One task was to be performed on one DieToRecs variant while the other one (similar but not identical) had to be carried out on a commercial system already existing for more than a decade. The sequence was crossed throughout the sample of 47 test persons. The random assignment of participants to different experimental conditions and the possibility to manipulate the independent variables were two basic features of the experimental approach. The goal of this kind of evaluation was to understand which system performs better (in terms of user satisfaction and/or successful task completion) and why. The evaluation was based on subjective quality-of-use measures adapted from the PSSUQ instrument (Lewis 1995) and on objective measures derived from the logging data of the complete experiment (for the DieToRecs variants only).

The results (documented in detail in a separate conference paper) indicate that the already implemented advanced recommender functions do better support the user to solve a given travel planning task. The subjective measures (in terms of overall

satisfaction, ease-of-use/learnability, and efficiency/outcome) exhibited a consistent improvement across the variants: from the naïve query based variant to the more complex inspiration inducing variant. The objective log-based behavioural data did not reveal that clear picture. However, the direction for a continuous development towards a second prototype with an even enhanced array of recommender functions seems to be justified.

From the experiences of this experimental evaluation several aspects and suggestions should be mentioned. 1. Building recommender systems for such a complex product like tourism destinations and the main services a traveller regularly consumes in this place challenges the existing evaluation procedures. The simulation of a real travel planning task within a test situation immediately touches some restrictions such as the available information space, the time span for planning a trip, the seriousness of travel preferences and budget constraints. Hence, the technical feasibility of the implemented routines can be seen as a necessary but not a sufficient condition from the usability point of view. 2. The performance tests have to be embedded in an environment that reflects realistic and therefore complete applications. This requirement raises preparatory costs and comprises functionality, interface design, the quality and scope of the database of travel items as well as those of CBR cases. 3. There are no adequate user satisfaction instruments available which cover the world of recommender systems. Some additional time and resources have to be reserved for adapting, testing and improving. 4. The proposed remedies are as follows: a) increase sample size, b) adopt better measures, c) complement laboratory experiments with web experiments, and d) use simulations.

6 Conclusion

This paper highlights the complexity of the evaluation procedure a travel support system must undergo to attain a minimum acceptable level of usability. This could only be achieved by a cooperation of usability experts, real users and technology providers. The maturity of DieToRecs improved a lot during the process and we are now facing a final step of system progress, which is based on the last empirical evaluation stage and on the analysis of the session logs. This last point refers to the optimisation of the interactive query management and the ranking technologies by means of machine learning algorithms.

In general, testing recommender systems means at least one step ahead in terms of sophistication of the available evaluation instruments. The result space is not strictly limited and determined. It depends closely on the user's contingencies as well as on the design of the whole interaction process. As a consequence, different results (complete or ad-hoc assembled travel bundles) may lead to different satisfaction levels while identical suggestions from the recommender system may cause different evaluations due to different paths on which the system guided the user to the final solution.

Acknowledgement

This work has been partially funded by the European Union's Fifth RTD Framework Programme (under contract DIETORECS IST-2000-29474).

References

Bevan, N. (1999). Quality in use: Meeting user needs in quality. *Journal of System and Software*, 49 (1), 89-96.

Chin, D.N. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction* 11(1-2): 181-194.

Dalgleish, J. (2000). *Customer-effective web sites.* Upper Saddle River, NJ: Prentice Hall.

Fesenmaier, D. R., Ricci, F., Schaumlechner, E., Wöber, K., & C. Zanella (2003). DIETORECS: Travel advisory for multiple decision styles. In Frew, A. J., Hitz, M., & P. O'Connors, (Eds.), *Information and Communication Technologies in Tourism 2003* (pp. 232-241). Wien: Springer.

Galitz, W.O. (2002). *The essential guide to user interface design. An introduction to GUI design principles and techniques.* New York, Wiley Computer Publishing, John Wiley and Sons, Inc.

Grabler, K. & A. Zins (2002). Vacation trip decision styles as basis for an automated recommendation system: lessons from observational studies. In Wöber, K., A. Frew & M. Hitz (Eds.), *ENTER 2002, Information and Communication Technologies in Tourism 2002* (pp.458-469) Wien: Springer.

Harms, I., Schweibenz, W. & J. Strobel (2002). Usability Evaluation von Web-Angeboten mit dem Usability-Index. In DGI (Ed.), *Proceedings der 24. DGI-Online-Tagung 2002 - Content in Context.* (pp. 283-292) Frankfurt am Main: DGI.

ISO/IEC 9126 (1991). Software product evaluation – Quality characteristics and guidelines for their use.

ISO/IEC FCD 9126-1 (1998). Software engineering – Product quality – Part 1: Quality model.

ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDT) – Part 11 Guidance on usability.

Ivory, M.Y. & M.A. Hearst (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys* 33:4, 470-516.

Keevil, B. (1998). Measuring the usability index of your web site. Conference Proceedings of *Human Factors in Computing Systems (CHI '98)* (pp. 271-277). Los Angeles, CA: ACM Press.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human Computer Interaction* 7(1), 57-78.

Lin, H. X., Choong, Y.Y., & G. Salvendy (1997) A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology* 16:4/5, 267-278.

Lindgaard, G. (1994). *Usability testing and system evaluation. A guide for designing useful computer systems.* London, Chapman and Hall.

Manhartsberger, M. & S. Musil (2002). *Web Usability: Das Prinzip des Vertrauens*. Bonn, Galileo Press.

Nielsen, J. (2001). *Designing web usability: The practice of simplicity*. Indianapolis, New Riders Publishing.

Nielsen, J. (1993). *Usability engineering*. Boston, MA, Academic Press, Harcourt Brace & Company.

Pearrow, M. (2000). *Web site usability handbook*. Rockland, MA, Charles River Media.

Ricci, F., Blaas, D., Mirzadeh, N., Venturini, A., & H. Werthner. (2002). Intelligent query management for travel products selection. In Wöber, K., A. Frew & M. Hitz (Eds.), *ENTER 2002, Information and Communication Technologies in Tourism 2002* (pp. 448-457), Wien: Springer.

Rieman, J., M. Franzke, & D. Redmiles (1995). Usability evaluation with the cognitive walkthrough. In ACM (Ed.), CHI 95 Conference Companion 1995 (pp. 387-388) Denver, CO: ACM.

Rubin, J. (1994). *Handbook of usability testing*. New York, John Wiley.

Rudd, J. & S. Isensee (1994). Twenty-two tips for a happier, healthier prototype. *Journal of Interactions* Vol. 1 (1), 35-40.

Rumetshofer, H., Pühretmair, F. & W. Wöß (2003). *Individual Information Presentation based on Cognitive Styles for Tourism Information Systems*. In Frew, A.J., Hitz, M. & P. O´Connor (Eds.), Proceedings ENTER 2003, Springer Verlag Wien New York, Helsinki, Finland, Janurary 29-31, 2003, pp. 440-449.

Walker, M., Takayama, L. & J.A. Landay (2002). Low- or high fidelity, paper or computer? Choosing attributes when testing web prototypes. In *Proceedings of Human Factors and Ergonomics Society, 46th Annual Meeting HFES2002* pp. 661-665.