

Evaluation of an on-vehicle adaptive tourist service

Luca Console, Cristina Gena, Ilaria Torre

Dipartimento di Informatica, Università di Torino
Corso Svizzera 185; I-10149 Torino, Italy
Email: {lconsole, cgena, ילותorre@di.unito.it}

Abstract. This paper describes methodology and results obtained in the evaluation of a system that provides personalized tourist information onboard cars. With a PC simulator, using a layered sampling strategy and strong statistic metrics to compare the system suggestions to the users answers, we analyzed several dimensions of adaptation (user preferences, context risk, etc.).

1. Introduction

The goal of providing large amounts of information at the driver's fingertips is becoming more and more important in the last decade. The dashboards of modern cars include, and in some cases integrate, an on-board computer, a GSM telephone, possibly with GPS, a navigation systems, besides more common devices such as radio, CD/DVD player, etc. The presence of these electronic systems, and the fact that people are spending more and more time on cars, suggested car manufacturers the design of new systems for providing to the various types of services on board the car (news, information about facilities and tourist locations, etc). The availability of these kinds of systems represents an opportunity but could also become a problem for the driver: the services may be very useful or even necessary, but the use of the systems may distract the driver causing serious dangers for active and passive safety. This led car manufactures and providers of these systems to study the design of the systems from the point of view of ergonomics and human-machine interaction. In the last few years some researchers suggested that *adaptation* and *personalization techniques* can play a very important role in vehicle on-board applications and can significantly contribute to solve the problems previously mentioned. Believing in the potentials of such techniques applied to this area, in the 2000 we started the study of a framework for on-board adaptive systems and implemented a prototype, MASTROCARONTE, which exploits adaptation and personalization techniques to provide *tourist information* to a driver (see [3], [4] for details).

Given the complexity of adaptive systems, due also to the discretionary choices they unavoidably carry out, evaluation is considered a very important subject in the user modeling and adaptive systems community (see [1], [2], [7], [9]). As recommended by the paradigm of user-centered design [6], it should be performed already in the first design phases to get immediate feedbacks from users. Thus, we started an evaluation exercise of the first prototype of MASTROCARONTE. aimed at showing that indeed adaptation and personalization can contribute to the achievement

of *two major goals*: *i*) checking whether the first items suggested by the system are in accordance with the user's preferences and needs and *ii*) whether the mode and format of the presentation is in accordance with the user's features and contextual situation (especially the driving conditions).

The main aim of this paper is to report the results obtained from this evaluation exercise. In particular, the paper is organized as follows. Section 2 sketches the framework and architecture of MASTROCARONTE. Section 3 describes the evaluation while Section 4 reports the results of the evaluation. Section 5 concludes the paper.

2. The system under evaluation

In the past two years we defined a framework and architecture for on-board adaptive systems, implementing a prototype application MASTROCARONTE, for providing tourist information to the driver. For details about the architecture of the system see [3], [4]. For the purpose of this paper, it is worth spending a few more words on the user model, the interface and on the rules used by the content adaptation agent and interface agent to rank items and to decide the presentation format.

The *USER MODEL* contains several pieces of information about the user and is initialized starting from stereotypes about the behavior of the Italian population¹, which provide probabilistic information about the level of interest of classes of users for aspects such as: traveling, art, history, nature, visiting museums. They also provide information about the propensity to spend and consume, whose values highly contribute in the recommendation of hotels and restaurants. Finally, the user model contains information regarding the user's receptivity (estimated from parameters such as the user's age, her stereotypical classification, familiarity with electronic devices and interfaces and visual problems, etc). The initialization is performed off-line, e.g., at the dealer, storing the result on a user's personal smart card. These initial estimates are then revised by the system by performing analyses on the actual behavior of a specific user: MASTROCARONTE tracks the actions of the user and updates her model based on statistic taken from this log of actions.

The *INTERFACE* for presenting information to the user includes two media: audio speakers and a screen. As regards the screen we defined a number (five in the prototype) of presentation styles. Each style defines a format (number of items to be displayed, fonts, colors) as well as the number of extra services that the user can access (i.e., calling a restaurant/hotel, asking for the route to reach it, etc) and that can be very useful for getting feedback about the user's behavior.

The *RANKING OF ITEMS* provided by the server is based on a set of rules that make two kinds of evaluation. The first one provides a score to each hotel or restaurant, according to the user's characteristics, namely, her propensity to spend (coming from the predictions in her model) and an estimation of her preferences, such as preference for a type of food or a type of place (computed with another set of rules). The second

¹ We defined the stereotypes starting from a psychographic study (Sinottica Eurisko) about the Italian population. Examples of descriptive features are: age, school degree, job, geographic area, etc.

evaluation regards the contextual information (time of the day, distance, type of area – metropolitan, extra urban, etc.) and is used to weight the previously selected items.

The *SELECTION OF THE INTERFACE*, medium and format, is more complex as it involves several parameters, such as the user receptivity and preferences (but also estimates of the user tiredness), the driving conditions (e.g., speed, traffic), contextual information (e.g., time of the day, weather, etc.). This is achieved by means of a set of rules which operate in steps, first deciding the medium, then the format.

In the following we show, as an example, the layouts displayed by the system with two different contexts: *i*) a university student with high level of receptivity (young, familiar with electronic devices, no visual problems, etc.), medium/low propensity to spend, while driving at a low speed with no traffic (see Fig. 1); *ii*) the same student, while driving at a high speed, but with non traffic and in a straight way (see Fig. 2).

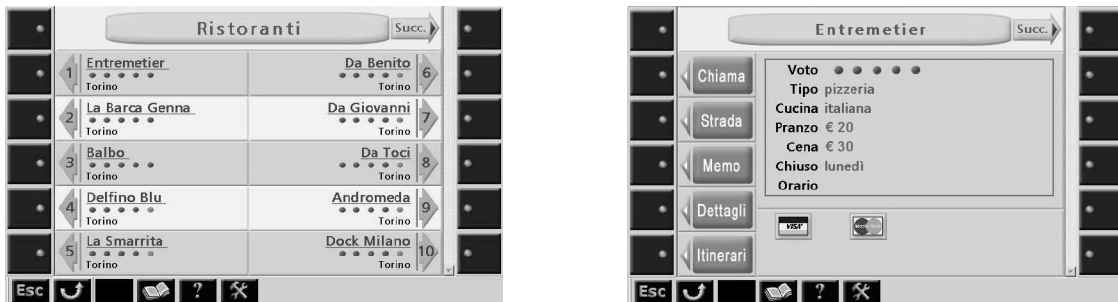


Figure 1 – Recommendations for the user 1 (university student) in context 1 (low risk level)



Figure 2 – Recommendations for the user 1 (university student) in context 2 (medium risk level)

3. The evaluation and its methodology

In the UM community the importance of systems evaluation has been strongly advocated (see [1], [2], [7], [9]) and now it is a shared principle. As regards the automotive environment, in our opinion, it requires considering *several aspects*: first of all, of course, the matching between the real users preferences and the features of the items suggested by the system. Second, the correct weight to external conditions, like distance, time pressure, etc., in the selection of the items. Finally and most

important, the analysis must evaluate if the system adapts its content, presentation and behavior in order to, and respecting the requirement of, being safe and not intrusive. Since the system is still under development, we performed *i) a formative evaluation*, which is aimed at checking the initial choices and getting clues for future revision, concerning the Knowledge Base (KB) implementation and the correctness of adaptation rule; *ii) a predictive evaluation*, based on HCI experts estimation, concerning the interface design choices.

To obtain reliable users' data we needed an accurate and quick way to collect self-reported information from target users. Thus we decided to exploit a questionnaire we personally distributed to 107 users identified following a non-probabilistic² blocking sampling, where the population is divided into layers related to the variables that have to be estimated and containing each one a number of individuals proportional to its distribution in the target population. We identified eight groups characterized by different age, sex, education, job, technology expertise, geographic area, etc. (that are the same descriptive data used by the system to classify users). Every group identifies a potential user. For instance group s1 (5% of our sample) is characterized by age: 26-35, sex: male, education: high school, job: autonomous workers, technology expertise: medium, etc.; while group s8 (23% of our sample) is characterized by age: 36-45, sex: male and female, education: high school/degree, etc..

To obtain the desired information we exploited the questionnaire collecting two sets of data: (a) information useful to the system to classify users and to generate recommendations and interfaces adaptations; (b) information about the real users' preferences useful to calculate the distance between system's recommendations and real users preferences. Six main topic areas were identified in the questionnaire: personal data, information about visual problems, familiarity towards computers and human-computer interfaces, food and restaurant preferences, restaurant prices preferences, hotel prices preferences. The final questionnaire was made up of 14 questions where both the questions and the answers were set. The questionnaires were filled in by the users to avoid any possible interviewer's interferences and gained a week after the distribution. The questionnaire was anonymous and introduced by a written presentation explaining the general research aims. For the items concerning personal data, visual diseases, computer and interfaces, the participants were required to tick the appropriate answer from a set of given answers. In the other questions, users had to express their level of agreement with the options concerning the given questions by choosing an item of a 5-point Likert scale.

We inserted the data set (a) in a PC simulator version of the system to generate the system responses. This version contains a service database that, for the evaluation, has been populated with tourist information about the Turin area. After having inserted the data, we analyzed the correctness of recommendations by the exploitation of two *statistical accuracy metrics* [8], MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) that are aimed at evaluating the closeness between the numerical recommendations provided by the system and the numerical ratings entered by the users for the same items. More precisely, these metrics evaluate the distance between the system predictions and the collected users opinions -set (b)- by means of rate vectors (both user's and system's items were expressed on a scale ranging from 1 to

² This sampling strategy is not probabilistic since the subjects are not randomly selected.

5). Obviously, higher values mean worst recommendations. These metrics are frequently exploited to evaluate the effectiveness of recommender and collaborative filtering systems [8]. We used a similar approach also for evaluating the accordance of the proposed layout with contextual conditions and user's cognitive capabilities. On the simulator we set two contexts (different for speed and traffic level) for each user. The layouts loaded by the system were then compared with that ones chosen by two HCI experts (again the distance is on a scale from 1 to 5).

4. Results of the evaluation

4.1. Results of the evaluation regarding item selection and ranking

First of all, we observed that in all cases the system was able to select facilities close to the user's location, giving priority to the closest ones at specific times (e.g., closest hotels at night). We can say that the requirement for MASTROCARONTE of giving a correct weight to external conditions, like distance, time pressure, etc., in the selection of the items, is satisfied. As regards the correspondence between the user's preferences and the features of the items selected by the system, the obtained results are here summarized:

Restaurants				Hotels	
MAE	RMSE	MAE*	RMSE*	MAE	RMSE
1,44	1,75	1,05	1,49	1,87	2,08

Table 1. Experimental results. (*) means without the restaurant prices predictions

As it can be seen, the results are partially satisfactory. Concerning restaurants, we noticed that the most of the distance was due to the price. The recommendations concerning prices were too optimistic: most participants chose lower price levels (notice in Table the better results obtained without calculating price prediction). The price problem also explains the higher values concerning Hotels that are recommended mainly following price estimations. As we will see, one of the possible reasons could be due to the current economic situation, which is probably different compared to that one considered by the psycho-graphic study used to build stereotypes and we have to adapt the price recommendations to the current propensity to spend. Thus, Table 1 gives a snapshot of the results and provides the important information regarding prices. However, such a piece of information and others, not represented in the table, come from a complex analysis aimed at explaining the reason of that 1,44 or also 1,05 distance.

Several are the ways that can be followed to investigate the cause, or causes, of the distance to finally finding out where the system has to be modified. In particular we have identified two types of approaches, which could be defined, for simplicity, as *bottom-up* and *top-down*. The first one begins the investigation from the data, looking for cases, or group of cases, with anomalous statistics and lets the evidences emerging from the data. The second type of approach starts the investigation with no knowledge about the collected data, makes hypotheses and then tests the component(s) of the system which can produce errors of adaptation.

4.1.1. The bottom-up approach

In order to reduce the space of possible causes of error and thus quickening the discovery of the problem, we started the evaluation with the first type of approach. To perform this analysis, we disaggregated the results by considering different groupings of the users. In particular, we considered three ways to group the subjects of the test:

- *Sampling Groups* = classes of subjects with common socio-demographic features (in our case we singled out eight groups labeled as s1, s2, s3,, s8)

- *Profile Groups* = classes of subjects having the same predicted profile. In other words, classes of subjects having the same antecedents used in the adaptation rules. The profile are transversal with respect to the stereotypes. A profile includes *i*) similar predictions of stereotypes concerning the propensity to spend, technology expertise, receptivity; *ii*) similar rules which estimates the receptivity of the driver; *iii*) similar age. We singled out fifteen groups labeled as pf1, pf2... pf15.

- *Prediction Groups* = classes of profiles, namely subjects belonging to some profiles, for which the system produced the same recommendations (we singled out five groups labeled as pr1, pr2,...). These are clearly related to the previous ones: each prediction group includes a set of profile groups. In our case, for example, we have that $pr2 = pf3 \cup pf6 \cup pf10$. The basic idea of this analysis is that by comparing the behavior of the systems for these groups, it may be possible to get hints to understand for which reasons the system does not provide a good advice on some users.

Analysis of Prediction Groups. As a first result of this deeper evaluation, we noticed that the recommendations changed significantly according to the different predictions groups of users: some groups received better recommendations than others (see Table 2). As noticed above, the five *prediction groups* cluster participants with different socio-demographic features (they crossed the initial sampling groups) but common recommendations. For instance, people belonging to group pr2 are 26-35 years old and have a medium propensity to spend. Within group pr2 there are males and females, with different education levels and different professions (employees, autonomous workers, managers). Concerning the results, the table suggests that either the accuracy of classification or the prediction for pr5 and pr4 is more problematic than for other groups.

Prediction Groups	Restaurants		Hotels	
	MAE	RMSE	MAE	RMSE
group pr1	1,46	1,78	2,05	2,27
group pr2	1,08	1,40	1,00	1,27
group pr3	1,29	1,53	1,85	2,05
group pr4	1,54	1,85	1,92	2,11
group pr5	1,72	2,05	2,02	2,24

Table 2. Prediction groups' results

In order to have a better evaluation of these results, we wanted to understand if these differences were uniform among all the *Profile groups* that belong to the same *Predictions group* or not.

Analysis of Prediction Groups, divided for Profile Groups. The idea was that, in case we observed some *Profile groups* with very high MAE with respect to the belonging *Prediction group*, we could hypothesize that the negative result of the

Prediction group is due to one only *Profile group*. Thus, it becomes possible to limit the probable causes of error: it can be a classification problem or a personalization problem regarding only the specific values of that *Profile*. The results of this evaluation are reported in Table 3. As you can see, we have computed the distances MAE and RMSE between each *Profile Group* and the belonging Prediction Group, and then we calculated the Standard Deviation (SD) inside each *Profile Group*. We decided to exclude (also in the following table) from such a computation, *Profile Groups* with less than five subjects. The result was that many groups could not be classified, making all the remaining statistics not significant.

Predictions groups	Profile groups	Restaurants			Hotels		
		MAE	SD	RMSE	MAE	SD	RMSE
group pr1	Group pf1	1,66	0,40	1,90	2,13	0,14	2,31
group pr1	Group pf2	1,43	0,29	1,74	1,99	0,23	2,23
group pr1	Group pf4	1,39	0,21	1,76	2,06	0,89	2,40
group pr2	Group pf3	0,98	0,16	1,37	0,90	0,15	1,13
group pr2	group pf6	0,83	0,13	1,26	1,02	0,16	1,30
group pr3	group pf7	0,86	0,21	1,17	2,05	0,22	2,19
group pr3	group pf8	0,82	0,10	1,13	1,45	0,21	1,64
group pr4	group pf13	1,14	0,27	1,63	1,94	0,34	2,13
group pr5	group pf17	1,27	0,21	1,74	1,97	0,19	2,22

Table 3. Prediction and Profile groups' results

We already said that the aim was discovering if, inside a *Prediction group* with high MAE, there is a *Profile group* very deviating from the average MAE. See for example the profile pf1 of the group pr1. Its MAE for restaurant is 1,66 and the MAE of the belonging *Prediction Group* is 1,46.

- If we consider the SD of pf1, we see that it is 0,40. This means that the group is not very homogeneous. The analysis of the SD is important because it allows to understand if the deviation of the group is due to a single isolated case or not.

- Once we have seen that it is not an isolated case to determine the deviation, the problem becomes understanding the cause of the deviation of the Profile group, i.e. if it is a problem of classification or of personalization. This second evaluation can be done easily, because we know exactly the values of the dimension used by the system to classify (the values that identify the Profile). For example, for the group pf1 we judged that the adaptation rules were right defined. So we were reasonably sure that it was a classification problem.

- Finally it is even possible to identify the exact component in charge of the bad classification, just calculating the MAE for the different features of the advice. For example, for restaurants, the features prices, kind of food, type of places, etc. are correlated with specific user model dimensions, which can be computed using the stereotypes (e.g., propensity to spend and thus price) or the rules (type of food and of places). As an example, for pf1 we discovered a problem in the propensity to spend (stereotypes).

Analysis of Sampling Groups. We now move to the last analyses regarding the bottom-up approach. Considering the *Sampling groups*, we have again different results for different groups (see Tab. 4).

Sampling Groups	Restaurants		Hotels	
	MAE	RMSE	MAE	RMSE
group s1	1,39	1,67	1,64	1,75
group s2	1,41	1,69	1,81	2,08
group s3	1,36	1,71	1,00	1,18
group s4	1,59	1,92	2,06	2,28
group s5	1,18	1,55	1,31	1,58
group s6	1,49	1,78	2,05	2,23
group s7	1,22	1,51	1,85	2,06
group s8	1,42	1,73	1,70	1,88
not-classified	1,50	1,81	1,93	2,11

Table 4. Sampling groups' results

The *Sampling groups* are the initial 8 groups collected by the sampling strategy. It happens that 5 subjects result as not classifiable in any group. Also in this case there are differences between groups: for instance *group s6* receives the best restaurant recommendations while *group s3* receives the worst recommendations.

Finally, we compared these results with those ones obtained for the prediction groups. An ANOVA comparing the groups' results showed that the different results are due to a significant correlation between the kind of group taken into account (independent variable) and its related recommendations (dependent variable).

Predictions groups (103 subjects for 5 groups): restaurants MAE: $F(4, 98) = 9,27$, $p < 0,01$; restaurants MAE no prices: $F(4, 98) = 5,33$, $p < 0,01$; hotels MAE, $F(4, 98) = 26,83$, $p < 0,01$.

Sampling groups (107 subjects for 9 groups): restaurants MAE: $F(8, 98) = 2,74$, $p < 0,01$; restaurants MAE no prices: $F(8, 98) = 1,71$, $p < 0,01$; hotels MAE, $F(8, 98) = 9,40$, $p < 0,01$.

All these results (except for restaurants MAE no prices for Sampling groups) show significant dependencies. In summary, by analyzing the groups precisely, we could thus get clues on the parts of the KB that should be revised, but not enough for understanding the exact changes that should be brought to the system. Therefore we followed a second way of analysis, that one we called top-down approach.

4.1.2. The top-down approach

Test of the correctness of the KB. Our starting hypothesis was that each stereotype shares homogeneous preferences, behaviors and lifestyles. The inexistence of this supposed correlation could be due to these factors: i) the stereotypes are too generic and therefore they cannot be used for a specific domain; ii) the inserted data are too old and do not reflect the current situation. We could accomplish this test using the questionnaire and comparing the answers with the Eurisko classes. The answers in the questionnaire showed that users almost always selected ranges of prices lower than

those predicted by the lifestyles. This finding was also confirmed by the results shown in Table 1. Therefore the need of updating our KB, concerning the supposed propensity to spend emerged clearly. The other dimensions (technology expertise, receptivity) derived from the stereotypical knowledge seemed to be well suited for the system's purposes, and therefore non generic.

Test of possible errors in the KB implementation. The research we exploited describes the lifestyles in a qualitative way and thus we had to translate them in probabilistic values. Errors are frequent in processes like this, due to a misunderstanding in tuning the estimates. For instance we found that working young people are described as having a high propensity to spend. However, the test demonstrated that their propensity to spend is related to their current situations, so they generally prefer non-expensive restaurants even if they like to go out for dinner. As a consequence, we decided to split the propensity to spend in frequency and value, but Eurisko does not have this distinction and we have no means to know its interpretation of the dimension. A confirmation of this problem comes from the MAE calculated with respect to students. A possibility could be to define a new KB, exploiting a domain specific survey to the target population.

Test of the Rules in charge of the personalization. In the system, a set of rules associates user features (age, propensity to spend) to hotels/restaurants features (price, restaurant, food, etc..). The test showed some problems. For instance, the restaurant suggestions for 25-36 years old are better than those ones for 20-25 years old. Then, within the first group the suggestions are better suited for people having a medium propensity to spend. Therefore, the aim of this analysis was to discover the associations that should to be revised, as in the case above described.

In conclusion, the exploitation of bottom-up and top-down approaches suggested a revision of both the KB and the rules. The use of a methodology problem-oriented helped us to identify the main problems and solve them.

4.2. Results of the evaluation regarding interface

Finally, let us move to the evaluation of the personalization choices as regards the format and layout of the presentation. As regards layouts, the evaluation was made as follows. We interviewed two HCI experts, asking them to suggest the better interface for groups of subjects sharing the same features used by the system to generate layout recommendations (age, technology expertise, receptivity) in a given contexts. Then we calculated the distance (MAE and RMSE) between the real system's proposals and experts' suggestions. Here the results:

Layout Context 1		Layout Context 2	
MAE	RMSE	MAE	RMSE
0,18	0,18	0,09	0,09

Table 5. Layout prediction' results

The closeness between the system's proposed layout and the HCI experts' suggestions confirmed the appropriateness of layout adaptations choices.

6. Conclusions

In their review of personalized hypermedia presentation techniques Kobsa et al. [5] divide the personalization process into three major tasks: *i*) acquisition method and primary inferences, *ii*) representation and secondary inferences, *iii*) adaptation production. In this evaluation we took into account this process by performing a layered evaluation where the evaluation is decomposed into these layers (for other layered evaluation methodologies see [9]). We performed a formative evaluation for the points *i* and *ii* and a both a formative and predictive evaluation for point *iii* (which includes both content and interface adaptations). The final results suggest and give clues for a revision of the current KB and adaptation rules concerning the final contents recommendations, while the interface adaptations obtained good evaluation results. We reached these results by applying both a bottom-up and a top-down approaches to the data analysis. This methodology can be easily re-used in similar evaluations since the underlying ideas are *i*) letting evidences emerging from data by clustering users in different and overlapping groups and looking for inconsistencies; *ii*) testing the inference mechanisms of the system and their generating hypotheses with real data. The next step in this user-centered iterative evaluation process will be a testing with subjects interacting with a version of MASTROCARONTE currently running on a Fiat Punto to have both users rating the predictions generated by the system and interacting with the on-board car system.

References

1. D. Chin. Empirical evaluation of user models and user-adapted systems. In *User Modeling and User-Adapted Interaction* 11 (2001), pp 181-194.
2. D. Chin and M. Crosby editors: Empirical Evaluation of User Models and user Modeling Systems. *User Modeling and User-Adapted Interaction*, vol. 12 (2-3), 2002.
3. L. Console, I. Lombardi, R. Montanari, M. Guagliumi, M. Salvoni, L. Tonelli, I. Torre: MASTROcarONTE, a Multiagent Adaptive System for Tourist Recommendations Onboard the car, which Observes the Needs and Tailors the Helps, In *Proc. IJCAI Workshop on Artificial Intelligence in Mobile Systems*, AAAI Press, Seattle, WA, 2001, pp. 19-25.
4. L. Console, S. Gioria, I. Lombardi, V. Surano, I. Torre. Adaptation and personalization on board cars: a framework and its application to tourist services. In *Proc. 2nd Int. Conf. On Adaptive Hypermedia and Adaptive Web-Based Systems*, Malaga (2002) pp. 112-121.
5. A. Kobsa, J. Koenemann, W. Pohl, Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review* 16(2): 111-155, 2001.
6. D.A. Norman, S.W. Draper, S.W. Draper, S.W. Draper: User centred system design: new perspective on HCI, *Hillsdale NJ, Lawrence Erlbaum*, 1986.
7. D. Petrelli, A. De Angeli and G. Convertino. A user centered approach to user modeling. In *Proc. 7th Int. Conf. On User Modeling*, Banff (1999) 255-264.
8. B., M., Sarwar, J. A., Konstan, A., Borchers, J., Herlocker, B., Miller and J. Riedl: Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In *Proceedings of CSCW '98*, Seattle, WA (1998).
9. Workshop on Empirical Evaluations of Adaptive Systems, <http://art.ph-freiburg.de/um2001/>, (2001).