In *Proceedings of the NL 2002 Conference* (Berlin, Germany, May 2002).

Optimizing and profiling users online with Bayesian probabilistic modeling

Petri Nokelainen, Henry Tirri, Miikka Miettinen, Tomi Silander
Helsinki Institute for Information Technology,
P.O.Box 9800, FIN-02015 Helsinki University of Technology
firstname.lastname@hiit.fi

Jaakko Kurhila
Department of Computer Science
P.O.Box 26, FIN-00014 University of Helsinki
jaakko.kurhila@cs.helsinki.fi

Abstract

One solution to build adaptive educational material is to model the user with a questionnaire before he/she enters the system, and then use this information to carry out adaptation of the platform. For example, users that are profiled could be offered personalised links to resources based on their metacognitive strategies or intrinsic goal orientations. These machine understandable beliefs of the profiles of different users could then be updated by collecting additional information with on-line questionnaire in regular intervals. An adaptive on-line questionnaire system EDUFORM is based on intelligent techniques that optimize the number of propositions presented to each respondent. In addition EDUFORM creates an individual profile for each respondent. The adaptive graphical user interface is generated automatically (e.g., propositions in the questionnaire, collaborative actions and links to resources), and profile analysis and the related selection of order of the propositions is performed with Bayesian probabilistic modeling. Preliminary testing implies that the obvious advantage with EDUFORM is that the questionnaires are usually significantly shorter compared to traditional non-adaptive questionnaires. The empirical results show that after reducing dramatically the number of propositions (from 50-60%) one is still able to control the error ratio (12-22%). In the context of course feedback from a web-based course, the model construction in the Profile creation phase can offer he1p for teachers to find differences among the various learner groups so that different versions of the web course can be prepared to suit the individual needs of the group. The correct profile information of the respondent is in most cases obtained already with less than 33% of the original proposition set.

Main goal

The main goal of this paper is to describe the design and implementation of a software module, called EDUFORM[1], a Web-based data gathering tool, which performs adaptive and dynamic optimization of the number of questionnaire propositions during the actual data gathering process. This is achieved by probabilistic modeling techniques that allow for profiling the respondents based on the data gathered. EDUFORM uses probabilistic Bayesian modeling [1] to create the respondent profiles, and these models can be used to optimize dynamically the set of propositions that are showed to the user in order to maximally extract the information. It should be observed that although we are discussing the adaptive techniques in the context of (course) questionnaires, many of the features used in this restricted evaluation task can be directly applied to wider context of modern computer-based learning environments [2]. The educational problems investigated in this study are two-fold:

1. A great number of questionnaires, both on paper and electronic form, are designed with "one size fits all" - principle. Equipped with numerous propositions, usually around one hundred, along with some inadequate propositions related to the theory or underlying model, they prolong the answering process decreasing internal, external and contextual validity.

2. Learning environments are not effectively profiling learners which would allow the systems to promote collaborative and cooperative learning, or provide possibility to develop adaptive user interfaces and personalized contents and reference to additional resources.

---

[1] http://eduform.cs.helsinki.fi/software.html

The purpose of our research is to study the viability of intelligent data analysis techniques (in particular probabilistic modelling) to address the above concerns.

Data set and instrumentation

The instructional data consists of questionnaire results from 1800 students of a Finnish polytechnic institution. The data set was collected in December 2000 with both traditional and Bayesian optimized Motivated Strategies for Learning Questionnaire (MSLQ) [3]. The same organization with partly the same respondents (sub-sample of 460 students) is our target on the next measurement. Motivational profiling instrument (query) in this study is based on the MSLQ, which is developed on the basis of motivational expectancy model [4]. MSLQ measures both motivational factors and learning strategies, and has been adapted to the needs of the research field of Finnish vocational education [5]. The motivation section (A) of MSLQ consists of 28 items that were used to assess students' evaluation of the course, their beliefs about their skills to succeed in the course, and their anxiety about tests in the course. The learning strategy section (B) includes 40 items concerning student's use of different cognitive, metacognitive and resource management strategies. A 5-point Likert-scale ranging from 1 (Not at all true of me) to 5 (Very true of me) was used for all items. The initial order of items was randomized.

The analysis methodology: Bayesian modeling approach

EDUFORM is based on the models built from respondent data. The online questionnaire software itself is generic, and the models used by it can be produced in various ways. However, for many fundamental and pragmatic reasons the analysis in our research is based on Bayesian analysis, which is briefly outlined below. It should be remembered that our main goal is to produce respondent profiles that can be used to predict questionnaire responses after the respondent has already answered some of the questions. As a side product, interesting profile characterizations are built, and the profiles can then be used also for personalization of the course material in later phases.

Profile creation phase by finite mixtures

As stated above, EDUFORM relies on Bayesian modeling in the *Profile creation-phase.* Possible choices for model family could be the family of Bayesian networks [6] and family of finite mixtures [7]. Also Johnson's and Albert's [8] work in which they have estimated item response model parameters using Bayesian methods with prior distributions by assuming that the latent traits represent a random sample from a known population could have been a viable choice. The current version of EDUFORM relies on finite mixtures because of the criteria for terminating the questioning process in *Query-phase* can be straightforward if the user is to be profiled into a cluster (discussed in the next subsection).

In a questionnaire, it is quite natural to model the problem domain by (m) discrete variables $X_1,\ldots,X_m$ (possible continuous values discretised), and that a data d is sampled from the joint distribution of these variables. In finite mixtures we now make an additional modeling assumption that the data D can be viewed as if it were generated by K different mechanisms, all of which can have a distribution of their own. Furthermore, it is assumed that each data vector originates from exactly one of these mechanisms. Whether or not this actually is the case, is not of importance here. As we have already pointed out, model family is only a language in which we can express the constraints in data. From these assumptions it follows that the data vector space is divided into K local regions usually called *clusters* or *profiles*, each of which consists of the data vectors generated by the corresponding mechanism.

The underlying intuitive idea is that a set of data vectors can be modeled by describing a set of profiles, and then describing the data vectors using these profile descriptions. Each description gives the distribution of the variables $X_1,\ldots,X_m$ conditioned that the data vector belongs to the cluster. The cluster descriptions should be chosen in such a way that the information required to describe the data vectors in the cluster could be significantly reduced because they are similar to the "prototype" described by the profile. In such a "profile language" a data set D can be described by first giving the profile index for each data vector, and then by describing the differences between the observed and expected values.

In finite mixture models the problem domain probability distribution is approximated by a weighted sum of component distributions, where each mixture component $p(X_1 = x_1,\ldots,X_m = x_m|Y = y_k)$ models one data producing mechanism. It should be observed that the finite mixture model family is universal in the sense that it can approximate any distribution arbitrarily close as long as a sufficient number of components is used [7].

Figure 1 shows the structure of a finite mixture model in graphical form. Y denotes the latent variable, the values of which represent the clusters. When the model is

created, the number of clusters is chosen to minimise the information required to describe the data used. Variables $X_1,\ldots,X_m$ are assumed to be independent of each other given the latent variable Y.

Insert Figure 1. here

Figure 1. Graphical Bayesian network representation of finite mixture structure.

The mathematical form of a finite mixture model is

$$p(\vec{d}) = \sum_{k=1}^{K}\left(p(Y = y_k)\prod_{i=1}^{m} p(X_i = x_i | Y = y_k) \right). \quad (1)$$

Finite mixture as defined in Equation (1) is a generic model family, since we still have to fix the cluster distributions $p(X_i|Y = y_k)$. Construction of mixture models from a given data set D by using the Bayesian approach is described in articles by Kontkanen et al. [9] and Tirri et al. [10]. Naturally construction of these models can also be done using maximum likelihood approaches, however in EDUFORM we have adopted the Bayesian perspective as it allows us to use the prior information available (i.e., the theoretical framework of a questionnaire) and also helps us in the structure selection, i.e., selecting the proper number of profiles.

Query phase

As a result of the Profile creation-phase, a number of clusters have been identified in the sample data. A user answering the questions in EDUFORM eventually falls into one of these clusters. An attempt is made to reduce the required amount of answers significantly, while retaining the usefulness of the data acquired. The reliability of the predictions made by EDUFORM is discussed in another article [11].

The order in which the questions appear in EDUFORM is based on maximising the amount of information gained for profiling. *Kullback-Leibler distance* [12] is used to measure the difference between the current distribution and the distribution, which would result if the user gave a particular answer. For each of the remaining questions and their possible answers, the distance is calculated and weighted by the probability of the answer. As a result, the question with the maximum expected effect to the cluster distribution can be identified. At any moment, the finite mixture model knows the probability of the individual belonging to each of the clusters, as well as the probabilities of the alternative answers to the remaining questions. Figure 2 demonstrates an example of the log file of the user's actual answers and predicted answers. Every line represents a proposition in a questionnaire. The first column states the questionnaire name. The next column tells the number of a particular proposition. The next five columns represent the probabilities of a given answer. If the number is 1.0, the user has actually answered to the proposition and chosen manually that particular option. The last two rows in Figure 2 show other figures than 0.0 or 1.0 indicating that the potential answer of the user is predicted.

Insert Figure 2. here

Figure 2. The first six probability distributions of propositions in an adaptive questionnaire.

In the current experimental version of EDUFORM, questions are presented one-by-one until the probability distribution of the most likely cluster exceeds .80. Once this condition is met, the user is told he or she has provided the necessary information, and asked if the user would like to improve the accuracy of his or her profile by answering the remaining questions. An individual whose answering patterns are very different from the regularities captured by the model may have to answer all of the questions. If the clusters have been named and explanations written for them, the profile can be used for providing immediate feedback to the users.

Examples of EDUFORM interface

The EDUFORM user interface is shown in Figure 3. The propositions are on the middle part of the screen, and as seen, the seventh proposition has inspired the user to write an open comment regarding the proposition.

Insert Figure 3. here

Figure 3. EDUFORM user interface.

The Figure 4 presents a dynamic situation where user has actually given 24 responses (gray areas) and EDUFORM has inferred 24 responses (black areas) and 37 propositions are still undecided (white areas).

Insert Figure 4. here

Figure 4. EDUFORM questionnaire optimization.

The visualization of the current learner profile (groups of learners), is shown in Figure 5. The users are divided into different groups of learners based on their answers on the questionnaire. In this example the user profile gives an estimate where the learner is most likely to fit

into groups four or one, but the groups two and six are very unlikely.

Insert Figure 5. here

Figure 5. EDUFORM questionnaire current profiling state.

Empirical results

EDUFORM was tested with a sample of 66 students from a Finnish polytechnic Institute. The data was collected with EDUFORM in February 2001. Once profiling information during answering process was clear, EDUFORM gave each respondent a chance to move on to next part and skip remaining propositions, or, alternatively, finish answering questions of the current part. Those respondents who skipped were categorized as members of "Group 1" (Adaptive) and those who wanted to give all answers by themselves were members of the "Group 2" (Non-adaptive). Table 1 shows that group 1 has only seven participants (10.6 %) in part A (versus 57, 86.4 %), but already 23 (34.8 %) in part B. It should be observed that the first two parts of the questionnaire require more work from the respondent, containing mostly abstract propositions, than the remaining two which measure more practical matters. It is interesting to see that the size of group 2 (All propositions answered) grows in the last two parts of the questionnaire (62.1 % in both). Only 22 students (33.3 %) answered all 116 questions.

Table 1. Descriptive statistics of Group 1 (Adaptive) and Group 2 (Non-adaptive) of the adaptive educational questionnaire.

Insert Table 1. here

We learn from Table 2 that the total number of propositions needed to complete the questionnaire averaged from 67 (58 %) to 114 (98 %). Time elapsed during answering process varied from 6.1 minutes to 23.8 minutes showing time saving of at least 3.2 minutes, compared to non-adaptive electronic questionnaire. We estimated that the traditional paper version of the same questionnaire should be finished within twenty minutes. The least time savings were observed in part A (average 3.9 minutes versus 5.7 minutes) and the most in part C (average 1.7 minutes versus 3.1 minutes).

Table 2. Comparison of Group 1 and Group 2 by the number of propositions answered and the time elapsed.

Insert Table 2. here

The results of the profiling phase provide a way to both explorative (profile creation phase) and confirmatory (the final profile) comparisons of the theoretical dimensions to those found from the data. Theoretical dimensions found from the part A of the profiling phase support Pintrich's original theory, these results will be reported in [11].

Conclusion

We have described an adaptive on-line questionnaire system EDUFORM. The software is based on intelligent techniques that optimize the number of propositions presented to each respondent, and in addition creates an individual profile for each respondent. EDUFORM´s adaptive graphical user interface is generated automatically (e.g., propositions in the questionnaire, collaborative actions and links to resources), and profile analysis and the related selection of order of the propositions is performed with Bayesian probabilistic modeling.

Preliminary testing implies that the obvious advantage with EDUFORM is that the questionnaires are usually significantly shorter compared to traditional non-adaptive questionnaires. This can help to raise the answering percentage if the questionnaire is seemingly long and tedious, such as course feedback questionnaires in the universities. It is possible that since the process of filling in the questionnaire becomes shorter, the answers can be more accurate because the user is not exhausted with the long list of questions. The empirical results show that after reducing dramatically the number of propositions (from 50-60%) one is still able to control the error ratio (12-22%).

In the context of course feedback from a web-based course, the model construction in the Profile creation phase can offer he1p for teachers to find differences among the various learner groups so that different versions of the web course can be prepared to suit the individual needs of the group. The correct profile information of the respondent is in most cases obtained already with less than 33% of the original proposition set.

For future work the statistical techniques explored here are one possible solution to provide an intelligent agent to intermediate knowledge between collaborating students [13] as well as adaptation and personalization of the learning material.

References

1. Bernardo, J. and Smith, A. (2000). Bayesian Theory. John Wiley & Sons: New York. 2nd ed.

2. Dillenbourg, P. (Ed). (1999). Collaborative-learning: Cognitive and Computational Approaches. Elsevier: Oxford.

3. Pintrich, P. (2000). The Role of Motivation in Self-Regulated Learning. In P. Pintrich and P. Ruohotie (Eds.) Conative Constructs and Self-Regulated Learning, pages 31-50. Learning and Change Series of Publications: Saarijärvi.

4. Garcia, T. and Pintrich, P. (1994). Regulating Motivation and Cognition in the Classroom: The Role of Self-Schemas and Self-Regulatory Strategies. In D. H. Schunk & B. J. Zimmerman (Eds.), Self-Regulation of Learning and Performance: Issues and Educational Applications. Erlbaum: Hillsdale, N. J.

5. Ruohotie, P. (2000). Conative Constructs in Learning. In P. Pintrich and P. Ruohotie (Eds.) Conative Constructs and Self-Regulated Learning, pages 1-30. Learning and Change Series of Publications: Saarijärvi.

6. Cowell, R., Dawid, P.A., Lauritzen, S. and Spiegelhalter, D. (1999). Probabilistic Networks and Expert Systems. Springer: New York.

7. Titterington, D., Smith, A. and Makov, U. (1985). Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons: New York.

8. Johnson, V. and Albert, J. (1999). Ordinal Data Modeling. Springer: New York.

9. Kontkanen, P., Myllymäki, P. and Tirri, H. (1996). Predictive Data Mining with Finite Mixtures. In Proceedings of The Second International Conference on Knowledge Discovery and Data Mining, pages 176-182. Portland, OR, August 1996.

10. Tirri, H., Kontkanen, P. and Myllymäki, P. (1996). Probabilistic Instance-Based Learning. In L. Saitta, Machine Learning: Proceedings of the Thirteenth International Conference, pages 507-515. Morgan Kaufmann Publishers: San Fransicso.

11. Niemivirta, M., Nokelainen, P., Kurhila, J., Miettinen, M., Silander, T. and Tirri, H. (2002). Studying the Role of Individual Differencies in CSCL - The Adaptive Assessment of Motivational Profiles. (In preparation.)

12. Cover, T. and Thomas J. (1991). Elements of Information Theory. John Wiley & Sons: New York.

13. Hoppe, U. and Ploezner, R. (1999). Can Analytic Models Support Learning in Groups? In P. Dillenbourg (Ed). Collaborative-leaming: Cognitive and Computational Approaches, pages 147-168. Elsevier: Oxford.

Acknowledgements

The work belongs to educational technology division of an on-going project of Complex Systems Computation Group[2] (CoSCo) named Personalized Adaptive Interfaces (PAI). The partners involved are: National Technology Agency[3], University of Helsinki[4] and various unnamed Finnish Information Technology business partners.

We would also like to thank Professor P. Ruohotie for permission to use the Finnish Vocational Education adaptation of the Motivated Strategies for Learning Questionnaire.

[2] http://www.cs.helsinki.fi/research/cosco

[3] http://www.tekes.fi/eng/default.asp

[4] http://www.helsinki.fi/english

Tables

Table 1. Descriptive statistics of Group 1 (Adaptive) and Group 2 (Non-adaptive) of the adaptive educational questionnaire.

Part A	Part B	Part C	Part D	All propositions
Frequencies (n)[a]				
7 / 57 / 64 (2)	23 / 39 / 66 (4)	13 / 41 / 54 (12)	11 / 41 / 52 (14)	23 / 22[b] / 45 (21)
Percentages (%)[a]				
10.6 / 86.4 / 97.0 (3.0)	34.8 / 59.1 / 93.9 (6.1)	19.7 / 62.1 / 81.8 (18.2)	16.7 / 62.1 / 78.8 (21.2)	34.8 / 33.3 / 68.2 (31.8)

a Group 1 (Adaptive) / Group 2 (Non-adaptive) / total number of answers (missing data).
b Omitted case(s) due to total response time below 6 minutes and/or over 60 minutes.

Table 2. Comparison of Group 1 and Group 2 by the number of propositions answered and the time elapsed.

Part A	Part B	Part C	Part D	All propositions
Group 1 (Adaptive)				
Propositions answered (n)[a]				
7 / 26 / 17.8 / 8.6	17 / 38 / 32.9 / 5.7	8 / 20 / 14.3 / 4.3	12 / 24 / 17 / 4.4	67 / 114 / 100 / 14.3
Time elapsed (s)[a]				
76 / 781 / 265.7 / 246.9	100 / 385 / 231.9 / 80.2	30 / 382 / 104.5 / 94.7	24 / 167 / 62.1 / 40.3	364 / 1430 / 760.9 / 342.1
Time elapsed (min)[a]				
1.3 / 13.0 / 4.4 / 4.1	1.7 / 6.4 / 3.9 / 1.3	0.5 / 6.4 / 1.7 / 1.6	0.4 / 2.8 / 1.0 / 0.7	6.1 / 23.8 / 12.7 / 5.7
Group 2 (Non-adaptive)				
Time elapsed (s)[a]				
87 / 578 / 227.8 / 91.3	213 / 650 / 341.4 / 124.3	77 / 696 / 187.8 / 125.2	63 / 161 / 94.5 / 27.1	556 / 1389 / 825.1 / 216.6
Time elapsed (min)[a]				
1.5 / 9.6 / 3.8 / 1.5	3.6 / 10.8 / 5.7 / 2.1	1.3 / 11.6 / 3.1 / 2.1	1.1 / 2.7 / 1.6 / 0.5	9.3 / 23.1 / 13.8 / 3.6

a Min / max / mean /S.D.

Figures

Figure 1. Graphical Bayesian network representation of finite mixture structure.

Figure 2. The first six probability distributions of propositions in an adaptive questionnaire.

User name	Prop. No.	Value1	Value2	Value3	Value4	Value5
John	33	0.0	1.0	0.0	0.0	0.0
John	15	0.0	1.0	0.0	0.0	0.0
John	10	0.0	0.0	0.0	1.0	0.0
John	27	0.0	0.0	0.0	1.0	0.0
John	5	0.0149	0.0292	0.1225	0.2392	0.5939
John	11	0.0084	0.0086	0.0422	0.2451	0.6954

Figure 3. EDUFORM user interface.

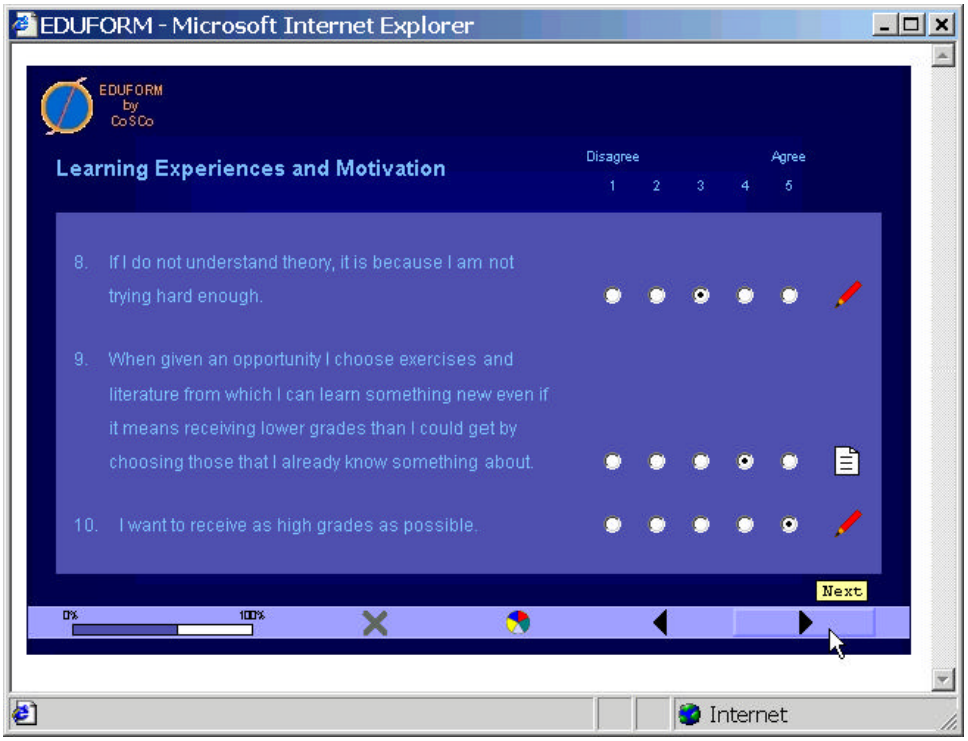

Figure 4. EDUFORM questionnaire optimization.

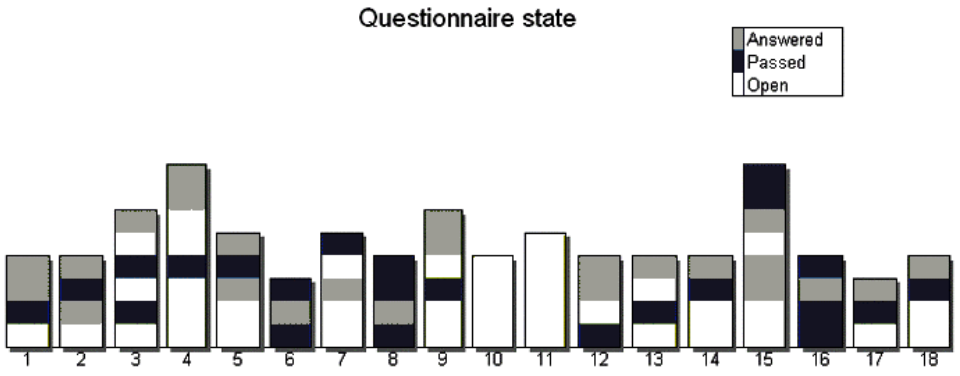

Figure 5. EDUFORM questionnaire current profiling state.