

# Machine Learning in User Modeling

Christos Papatheodorou

Division of Applied Technologies, NCSR "Demokritos", 15310 Athens, Greece  
papatheodor@lib.demokritos.gr

## 1 Introduction

It is generally recognized that information systems are becoming more complex and, therefore, intelligent user interfaces are needed to improve user interaction with these systems. Furthermore, the exponential growth of the Internet makes it difficult for the users to cope with the huge amount of available on-line information. The challenge that information providers and system engineers face is the creation of adaptive (Web-based) applications, as well as the development of "personalized" retrieval and filtering mechanisms. Responses to this challenge come from various disciplines including machine learning and data mining, intelligent agents and multi-agent systems, intelligent tutoring, information retrieval, etc.

User modeling (UM) aims to make information systems really user-friendly, by adapting the behaviour of the system to the needs of the individual. The importance of adding this capability to information systems is proven by the variety of areas in which user modeling has already been applied: information retrieval, filtering and extraction systems, adaptive user interfaces, educational software.

Machine learning (ML) techniques have been applied to user modeling problems for acquiring models of individual users interacting with an information system and grouping them into communities or stereotypes with common interests. This functionality is essential in order to have a useful and usable system that can modify its behavior over time and for different users. ML techniques are promising in cases where very large sets of usage data are available, like WWW sites, and other Internet applications (news filtering, digital libraries, search engines, etc.). In such applications, the information providers need a tool to help users in selecting useful information from the plethora of information on the Net.

This chapter is based on the work presented in the Workshop on "Machine Learning in User Modeling", which was organised in conjunction with the ECCAI Advanced Course on Artificial Intelligence for 1999 (ACAI '99). The aim of the Workshop was the presentation of the current trends and activities in exploiting Machine Learning techniques for the creation of adaptive user interfaces, the construction of user communities and stereotypes, the customized web usage and for the development of educational hypermedia. Pat Langley was invited speaker and the rest of the Workshop consisted of seven papers, which could be grouped in "clusters" corresponding to the topics: *Adaptive User Interfaces*, *Agent-based systems*, *Student Modeling and Web Mining*.

The following section presents an overview of UM and describes the general contribution of ML in UM. Section 3 discusses the current research efforts in the above-mentioned topics, as they were presented in the workshop. Finally section 4 presents some conclusions and suggestions for future research.

## 2 Machine Learning in User Modeling

A user model consists mainly of knowledge about the individual preferences which determine the user's behaviour. As in the most knowledge based approaches, two are the main issues that user modelling faces: The user model representation and acquisition.

User models could be represented as preferences sets, first order predicates or groups of predicates, plans, decision trees, etc. Furthermore, a user model may contain personal information about the user, such as his/her age, occupation, etc. The latter type of information is not directly necessary for the adaptation of the system to the user, but may be used to categorise the user into a *stereotype*, which in turn allows the system to anticipate some of the user's behaviour. Stereotypes have been introduced by Rich [34], as a means of organising the users of the system with common behaviour, into meaningful groups. Personal information about the users of a system is not always available and therefore the construction of user stereotypes may not be possible. This is especially true of visitors to a Web site. In that case, the organisation of users into groups with common interests can still be useful. Such a group of users is termed by Orwant [26] a *user community* and corresponds to a stereotype missing the personal information.

Many user model acquisition methods have been presented in the literature. These methods depend on the user model definition and representation and vice versa. For example plan recognition methods are used when user models are represented as plan hierarchies or libraries. In general, the user model acquisition methods could be grouped in two families: The explicit and the implicit (non-invasive) methods. The first family creates the user models by asking the users directly about their interests and preferences or by allowing them to specify and modify their own model. These preference setting methods could be found in many modern software packages. The implicit methods estimate and infer user models by tracking the user behaviour and interaction.

Educational software uses student modelling techniques to personalise the learning process, i.e., to make it adaptive to the student skills and background knowledge, as well as to predict the student's actions [40], [17]. Student modeling precedes historically user modeling and thus one could argue that it was one of the "causes" for the development of UM technology.

On the other hand, information retrieval and filtering systems aim to deliver to the user those documents that are relevant to his/her information requirements, whereas information extraction systems aim to deliver specific facts from those documents. NewT is a system that helps the user filter Usenet Netnews [22] according to his/her interests. Brajnik and Tasso [8], [9] present the exploitation of the user modelling

shell UMT in the prototype information-providing system Tourist Advisor (TA). Kay [19] describes the Movie Advisor project, which operates on a database of movie descriptions, suggesting movies that should appeal to a specific user. Doppelgänger [26] is a user modelling system that gathers sensor data about users, draws inferences on those data and makes the results available to applications. Firefly's agent software groups users' interests based on their similarities in order to suggest buying opportunities to specific customers based on their similarity to other customers. UMIE [5] is a Web-based prototype user modelling component (see <http://www.iit.demokritos.gr/UMIE>), that filters the data extracted from an information extraction system according to the users' models.

Machine learning (ML) methods have been applied to user modeling problems, mainly for acquiring models of individual users interacting with an information system, e.g. [7], [33], [32], [3]. In such situations, the use of the system by an individual is monitored and the collected data are used to construct the model of the user, i.e., his/her individual requirements. By definition, the development of intelligent and adaptive user interfaces presupposes the design of specific modules, responsible for learning user models by tracing the interaction of the system with the users [20], [10], [11].

Moreover, ML techniques have been used in many agent-based and multiagent systems which aim to discover and recommend information from Web sites like FAB [2], Syskill & Webert [29], WebWatcher [18], Amalthea [23] and Oyster [25]. Perkowitz and Etzioni [30], [31] deal with the development of adaptive Web-sites, which recommend (on the fly) pages relevant to the interests of their visitors. Specifically, they provide a conceptual cluster mining method that analyses the Web server usage data (logs).

Furthermore, ML for UM has been used in digital library services, like IDL [15], [14], and in news filtering systems, like News Dude [6]. The main goal of these systems is to learn and revise user profiles as well as to propose which information on a given topic would be interesting to a user.

Most of this work “revolves” around the World Wide Web, which provides a wealth of information to filter, as well as a famous interaction mechanism. We could classify the filtering methods into two families. The first one is *content-based* filtering in which the system accepts information describing the items, and learn to predict which items fit to the user models. Syskill & Webert and IDL are representatives of this family. The other family is the *collaborative* or *social* filtering, in which the system updates the user models and predicts items that fit to them, based on feedback from many different users. Typical representatives of this approach are FAB, Firefly and amazon.com. The works dealing with the formulation of user communities [26], [27] have similarities to the latter family. However there are systems (e.g. [3]), which adopt a hybrid philosophy in recommending items based on user models.

Apart from information discovery and filtering systems, Langley [20] refers to a variety of systems that use ML in order to adapt to particular situations that a user faces, such as travel assistance, routing in an area, emergencies reaction and scheduling. Furthermore, such techniques have been applied in intelligent tutoring systems [37], [12]. Finally, ML techniques have been applied successfully to the prediction of a users' actions. Significant research efforts on intelligent assistance

and/or tutoring view the user behaviour as a plan and learn the patterns of users' actions which achieve particular states [4]. Specifically, there are significant contributions concerning the usage of decision theory and Bayesian belief networks in plan recognition [1], [39].

### **3 Overview of the Workshop Papers**

The workshop consisted of the invited talk of Pat Langley and seven papers [41], presenting Machine Learning applications for the development of adaptive user interfaces, agent-based information discovery, student modeling systems, as well as for analyzing the behaviour of Web users.

#### **3.1 Adaptive User Interfaces**

Pat Langley in his invited talk [21] defined adaptive interfaces as "software artifacts that improve their ability to interact with users by constructing user models based on partial experience with each of them". Furthermore, he considered information filtering and recommendation systems as typical examples of adaptive interfaces and he presented a review of "content-based" and "collaborative" filtering systems as well as efforts, which combine the two approaches. Moreover, he described innovative applications of recommendation systems such as an Adaptive Route Advisor, which suggests alternative routes between a driver's current location and his destination. This system takes into account a weighted combination of parameters like the distance, the estimated driving time, the number of turns and intersections, etc. and proposes the routes that best fit to the current user model. Another application that was presented is INCA, which helps persons who are responsible in allocating resources in response to emergencies that involve hazardous materials. Finally, Prof. Langley compared the domain of adaptive user interfaces with other paradigms. He argued that the main difference between adaptive user interfaces and other cognition computational models, is that the former deal primarily with the content of human decision, while the latter focus on the process of human thought and decision making. He closed his talk providing a list of challenges and open questions that Machine Learning faces in creating Adaptive Interfaces.

The paper by Schwab and Pohl [35] described LaboUr, a recommendation system, which learns user profiles from ELFI, a Web-based system that provides information concerning research funds. In particular ELFI is organized into hierarchies of research topics, e.g. mathematics, computer science, etc., and funding types, such as grant, fellowship. The system displays the contents of a topic by listing links to so-called detailed views (DVs) of relevant funding programs. A DV consists of the available data about a research program, i.e. an abstract, the covered topics, the funding type, etc. The authors implemented a simple Bayesian classifier and they modified the known k-Nearest Neighbor algorithm to classify the set of information objects (DVs), selected by the users. Most probable or close DVs are considered as a class, which

could be recommended to a user. Finally, seeking to determine explicitly the users' interests, the authors employed statistical methods to find the DVs features that are important to an individual user.

Semeraro et al. [36] described an intelligent Web-based digital library, named CDL, which incorporates Machine Learning techniques both for document analysis and classification, as well as for user modeling. Users can access CDL in several ways, even through a dial-up connection to Internet, and interact with it via either a form-based, a tree-based, or a topic-map interface. The topic map provides a global view of the semantic content of the documents by showing the document topics. The topics have been defined incorporating standard thesaurus building techniques and represented by vectors. The navigation in the map or in the tree interfaces results into the automatic query composition. Regarding user modeling, CDL incorporates the supervised learning algorithm C4.5 as a module of the "Application Server" which is embedded in the Learning Server component. C4.5 takes as input the data stored in the log files, and classifies the users interacting with CDL, inducing a decision tree as well as a set of rules. Each entry of the interaction log files is labeled indicating three main registered user categories: Novice, Expert and Teacher. Furthermore, each class is associated with one of the mentioned interfaces. In this manner the CDL architecture uses Machine Learning techniques for managing user models and providing a weak form of user interface adaptivity.

### **3.2 Intelligent Agents and Multi-agent Systems**

Fragoudis and Likothanassis [16] provided a complete comparative study of the user modeling approaches of the most well-known information discovery intelligent agents, such as Lira, Letizia, Amalthea, Arachnid, Webmate, Webace, Syskill & Webert, Webwatcher, Siteseer, Law and Fab. They stated that Intelligent agents help people in two different ways: They can provide assistance while browsing or they may autonomously discover documents of interest, based on the results of a query to some search engine. The main conclusion of this work is that "none of the referred solutions seems to outperform the others and in any case the result is the same: Intelligent systems that help people on their battle against information overload".

The paper by Moukas [24] described "Amalthea", a multi-agent user modeling and information discovery system. Amalthea is implemented by creating an "ecosystem" of two kinds of cooperating agents: the Information Filtering Agents which are responsible for the acquisition and maintenance of the user profiles and the Information Discovery Agents which monitor and adapt user profiles to the various information sources. Both documents and user preferences are represented using the Vector Space Model. The results of the users' queries on existing search engines are used to define the interestingness of the documents. The user receives only the most "interesting" documents and rates them in a scale from 1 to 7, providing the ecosystem with relevance feedback. Amalthea assigns to each user model a population of Information Filtering and Information Discovery agents. The user profiles are updated dynamically by the Information Filtering Agents, which, evolve by three operators:

cloning, crossover and mutation. The agents' survival, offspring generation, or diminishment depends on the values of two measures: the *individual* and *overall fitness*. Agents with low individual fitness are purged and new ones are created, while the evolution rate is determined by the overall fitness of the ecosystem.

### 3.3 Student Modeling

Chiu and Webb [13] dealt with two issues confronted during the student behavior prediction process. First, what mechanisms could be employed in order to increase the predictions without degrading prediction accuracy and second, for contexts in which accuracy is of primary importance, how the prediction accuracy could be further improved. The paper improved previous work of the authors known as the Dual model. This approach utilizes temporal data in order to improve the prediction rate of a student modeling system. Specifically, Dual model creates a temporal model, namely *fresh model* - built using data from the most recent observations - in addition to a conventional model, which is referred to as an *extended model* - inferred from data of all historical observations. When a Dual-model system predicts a student's future actions, both models will be consulted. In this paper the authors, aiming to further increase the prediction rate, propose an alternative Dual-model strategy that consults the two models in parallel. Moreover they present and evaluate three new variants of Dual-model, which improve its prediction accuracy.

### 3.4 Web Mining

Spiliopoulou et al. [38] described WUM (Web Utilisation Miner), an algorithm which keeps track of all traversals of the paths through the pages of a site and aggregates them in a graph. The authors specify MINT, a mining language, which forms SQL-like queries to extract useful information from the "aggregated" graph. Specifically, MINT provides "interesting" navigation patterns (paths). Finally, the authors define the concept of "interestingness" according to the structure of a MINT-query.

The paper by Paliouras et al. [28] proposed a methodology for constructing user communities (i.e., groups of people with common interests) from Web logs and they applied it on the logs of the ACAI'99 Web site. The methodology deals with three major issues: The first issue is that of data engineering, which includes selecting the right representation for the training data and reducing the dimensionality of the problem. Regarding this issue, the authors propose the representation of access sessions as transitions between pages, or as bags of pages. The second issue concerns the selection of the clustering method. The authors construct user groups using the conceptual clustering algorithm COBWEB. The last issue, in which they pay substantial attention, is the characterization of the community models, i.e., the construction of meaningful communities. For this purpose, they introduce *frequency increase*, a metric to decide which are the representative features for each community of visitors. One of the results of this paper is that representing access sessions by means of transitions between pages produces interesting navigation patterns for the

community models. Finally the authors conclude that Web usage analysis is much more insightful than the approach of examining simple usage statistics of a Web site and that the employment of machine learning and user modelling techniques are very promising tools for this analysis.

## 4 Discussion

After the paper presentations an interesting discussion took place, debating the following issues:

1. The impact of ML in the development of adaptive interfaces and UM software.
2. What kinds of ML techniques are needed in UM.
3. Whether it is worth-while using "heavy" ML algorithms in the development of on-line interactive systems such as tutoring systems, or personalized Web-based applications.

The conclusions of this discussion could be summarized as follows:

1. ML offers a suite of powerful techniques either for user model acquisition or for user community induction.
2. ML techniques support complex decision making tasks and improve the prediction quality of a UM software.
3. UM should focus on the utilization of rapid and on-line learning techniques. This means that accurate and descriptive user models should be generated from small sets of training cases and that the models should be updated each time a user interaction occurs or at least before the next session.

## References

1. Albrecht, D.W., Zukerman, I. and A.E. Nicholson: 1998. Bayesian Models for Keyhole Plan Recognition in an Adventure Game. *User Modelling and User-Adapted Interaction* 8, 5 - 47.
2. Balabanovic, M. and Y. Shoham: 1997. Content-Based, Collaborative Recommendation. *Communications of the ACM* 4 (3), 66-72.
3. Basu, C., Hirsh, H., and W. Cohen: 1998. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. *Fifteenth National Conference in Artificial Intelligence*, Madison, Wisconsin, MW.
4. Bauer, M.: 1999. From Interaction Data to Plan Libraries: A Clustering Approach. *International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 962-967.
5. Benaki, E., Karkaletsis, V. and C. D.Spyropoulos: 1997. Integrating User Modelling Into Information Extraction: The UMIE Prototype. *Sixth International Conference on User Modelling*, 55-57.
6. Billsus, D. and M. Pazzani: 1999. A Hybrid User Model for News Story Classification *Seventh International Conference on User Modelling*, Banff, Canada, 99-108.
7. Bloedorn, E., Mani, I. and T. R. MacMillan: 1996. Machine Learning of User Profiles: Representational Issues. *Thirteen National Conference on Artificial Intelligence*, Portland, Oregon, 433-438.

8. Brajnik, G. and C. Tasso: 1994. A Shell for Developing Non-monotonic User Modelling Systems. *International Journal of Human-Computer Studies* 40, 31-62.
9. Brajnik, G., Guida, G. and C. Tasso: 1987. User Modelling in Intelligent Information Retrieval. *Information Processing and Management* 23, 305-320.
10. Brusilovsky, P., and E. Schwarz: 1997. User as Student: Towards an Adaptive Interface for Advanced Web Applications. *Sixth International Conference on User Modelling*, 177-188.
11. Chin, D.N.: 1989. KNOE: modelling what the user knows. In: A. Kobsa and W. Wahlster (eds.): *User models in dialog systems*. Berlin: Springer-Verlag, 74-107.
12. Chiu, B.C., and G. Webb.:1998. Using Decision Trees for Agent Modeling: Improving Prediction Performance. *User Modelling and User-Adapted Interaction* 8, 131-152.
13. Chiu, B.C., and G. Webb: "Dual-model: An Architecture for Utilizing Temporal Information in Student Modeling". In [41].
14. Crabtree, I.B. and S.J. Soltysiak: 1998. Identifying and tracking changing interests. *International Journal on Digital Libraries* 2, 38-53.
15. Esposito, F., Malerba, D., Semeraro, G., Fanizzi, N. and S. Ferilli: 1998. Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service. *International Journal on Digital Libraries* 2, 3-19.
16. Fragoudis, D. and S. Likothanassis: "User Modelling in Information Discovery: An Overview". In [41].
17. Giangrandi P. and C. Tasso.:1997. Managing Temporal Knowledge in Student Modelling. *Sixth International Conference on User Modelling*, 415-426.
18. Joachims, T., Freitag, D. and T. Mitchell: 1997. WebWatcher: A tour guide for the World Wide Web. *Fifteenth International Joint Conference in Artificial Intelligence*, Nagoya, Aichi, Japan.
19. Kay, J.: 1995. The um Toolkit for Cooperative User Modelling. *User Modelling and User Adapted Interaction* 4, 149-196.
20. Langley, P.: 1999. User Modelling in Adaptive Interfaces. *Seventh International Conference on User Modelling*, Banff, Canada, 357-370.
21. Langley, P.: "User Modeling in Adaptive Interfaces". In [41].
22. Maes, P.: 1994. Agents that Reduce Work and Information Overload. *Communications of the ACM* 37(7), 31-40.
23. Moukas, A.: 1997. Amalthea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. *Applied Artificial Intelligence: An International Journal* 11(5), 437-457.
24. Moukas, A.: "User Modeling in a MultiAgent Evolving System". In [41].
25. Mueller, M.: 1999. Inducing Conceptual User Models. *ABIS-99, 7. GI Workshop on Adaptivity and User Modelling in Interactive Software Systems*.
26. Orwant, J.: 1995. Heterogeneous Learning in the Doppelgänger User Modeling System. *User Modelling and User-Adapted Interaction* 4, 107-130.
27. Paliouras, G., Karkaletsis, V., Papatheodorou, C., and C. D. Spyropoulos: 1999. Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities. *Seventh International Conference on User Modelling*, Banff, Canada, 169-178.
28. Paliouras, G., Papatheodorou, C., Karkaletsis, V., Tzitziras, P. and C.D. Spyropoulos: "Learning Communities of the ACAI'99 Web-site Visitors". In [41].
29. Pazzani, M. and D. Billsus: 1997. Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning* 27, 313-331.
30. Perkowitz, M. and O. Etzioni: 1998. Adaptive Web Sites: Automatically synthesizing Web pages. *Fifteen National Conference in Artificial Intelligence*, Wisconsin, MW.
31. Perkowitz M., and O. Etzioni: 1999. Adaptive Web Sites: Conceptual Cluster Mining. *Sixteenth International Joint Conference in Artificial Intelligence*, Stockholm, Sweden, 264-269.

32. Raskutti, B. and A. Beitz: 1996. Acquiring User Preferences for Information Filtering in Interactive Multi-Media Services. Pacific Rim International Conference on Artificial Intelligence, 47-58.
33. Resnick, P. and H.R. Varian: 1997. Recommender Systems. Communications of the ACM 4(3), 56-58.
34. Rich, E.: 1983. Users are Individuals: Individualizing User Models. International Journal of Man-Machine Studies 18, 199-214.
35. Schwab, I. and W. Pohl: "Learning User Profiles from Positive Examples". In [41].
36. Semeraro, G., Costabile, M.F., Esposito, F., Fanizzi, N. and S. Ferilli: "Machine Learning Techniques for Adaptive User Interfaces in a Corporate Digital Library Service". In [41].
37. Sison, R., Numao, M. and M. Shimura: 1998. Discovering Error Classes from Discrepancies in Novice Behaviors via Multistrategy Conceptual Clustering. User Modelling and User-Adapted Interaction 8, 103-129.
38. Spiliopoulou, M., Faulstich, L. and K. Winkler: "A Data Miner Analyzing the Navigational Behaviour of Web Users". In [41].
39. Suryadi D. and P.J. Gmytrasiewicz: 1999. Learning Models of Other Agents Using Influence Diagrams. Seventh International Conference on User Modelling, Banff, Canada, 223-232.
40. Weber, G.: 1999. Adaptive Learning Systems in the World Wide Web. Seventh International Conference on User Modelling, Banff, Canada, 371-377.
41. Proceedings of the Workshop on Machine Learning in User Modeling, Advanced Course on Artificial Intelligence (ACAI '99), Chania, Greece, 1999 (<http://www.iit.demokritos.gr/skel/eetn/acai99/Workshops.htm>).