User Model User-Adap Inter (2007) 17:1–4
DOI 10.1007/s11257-006-9025-2

EDITORIAL

Introduction to the special issue on statistical and probabilistic methods for user modeling

David Albrecht · Ingrid Zukerman

Published online: 13 February 2007

Statistical and probabilistic models are concerned with the use of observed sample results to make statements about unknown, dependent parameters. In user modeling, these parameters represent aspects of a user's behaviour, such as his or her goals, preferences, and forthcoming actions or locations.

Recent technological advances, in particular increased computational power, together with anytime, anyplace access to computers, and the information explosion associated with the Internet, provide new opportunities for information dissemination and information gathering. On one hand, people have access to large repositories of information in digital form. On the other hand, information providers can find out more about their users' requirements by logging people's activities. This mixture of vast electronic content and increased knowledge about people's actions provides an opportunity to harness statistical and probabilistic models to build user models that support the delivery of personalized content.

This usage of statistical and probabilistic models has been manifested in UMUAI for the last ten years. Particularly noteworthy are the articles in the Special Issue on Machine Learning for User Modeling (1998); the survey articles by Zukerman and Albrecht and by Webb et al. in the 10-year anniversary issue, respectively on predictive statistical models for user modeling (Zukerman and Albrecht 2001), and on machine learning for user modeling (Webb et al. 2001); Burke's survey on recommender systems (Burke 2002); and Pierrakos et al.'s survey on Web usage mining (Pierrakos et al. 2003). These articles identified several challenges that user modeling presents to statistical and probabilistic modeling techniques. We classify these challenges into three categories: (1) limitations of current user modeling approaches, (2) dynamic nature of user modeling data, and (3) efficiency considerations.

D. Albrecht (✉) · I. Zukerman (✉)
Faculty of Information Technology, Monash University, Clayton Vic 3800, Australia
e-mails: David.Albrecht@infotech.monash.edu.au, Ingrid.Zukerman@infotech.monash.edu.au

Limitations of current user modeling approaches

The two main approaches employed in statistical user modeling are content-based and collaborative. In the content-based approach, the behaviour of a user is predicted from his or her past behaviour, while in the collaborative approach, a user's behaviour is predicted from the behaviour of other like-minded people.

The main shortcoming of content-based approaches is that the features selected when building a content-based model have a substantial effect on the usefulness of this model (Zukerman and Albrecht 2001). Features that are too specific yield a system that is useful only for repetitive behaviours, while features that are too general yield predictions of debatable usefulness. An inherent limitation of collaborative approaches is that they make predictions about the behaviour of a single user from observations of many users, hence they do not make predictions tailored to the behaviour of individual users.

Dynamic nature of user modeling data

Data related to user modeling changes constantly. New users may appear, and new items may be introduced. Additionally, it is often the case that the behaviour of users changes over time. These traits respectively cause two problems for current statistical modeling techniques: the *sparse data problem* and *concept drift*.

The *sparse data problem*, which is related to the *cold start* problem, represents situations where there is insufficient information to categorize users or items. Both the collaborative and the content-based approach have difficulty predicting the behaviour of users about whom there are few observations. Similarly, the collaborative approach cannot predict accurately the behaviour of users regarding items that have been insufficiently observed. In contrast, the content-based approach can make such predictions on the basis of the features of these items.

Concept drift represents the change over time of the attributes that characterize a user. This means that "historical" models, acquired early may become inadequate or obsolete.

Efficiency considerations

Efficiency issues pertain to the ability of statistical models to cope with vast amounts of data during all stages of the user modeling process, viz model building, prediction generation, and online adaptation. The last two stages demand the delivery of results in real time.

The contributions in this Special Issue address these and other challenges. Lekakos and Giaglis address the sparse data problem when modeling users' preferences. Domshlak and Joachims also model users' preferences, but they focus on ordinal preferences expressed by means of qualitative statements. On-line personalization has been considered in two spoken language understanding systems—a voice-enabled web browser (Chickering and Paek) and a Command and Control system for mobile phones (Paek and Chickering)—and in a student modeling system that focuses on test-item selection (Guzmán et al.). Additional challenges considered by the contributions

in this Special Issue pertain to coupling human and automated resources in a spoken dialogue system (Horvitz and Paek), and identifying navigation stages in a web site (Hollink et al.).

All the contributions in this volume have considered complexity or efficiency issues in some way. For example, Domshlak and Joachims emphasize computational tractability during model construction, and Lekakos and Giaglis consider the time complexity of making predictions. Guzmán et al., and Chickering and Paek have similar concerns during on-line adaptation, with the former employing item-selection algorithms that reduce data requirements and computational costs, and the latter focusing on a quick adaptation of their predictive model.

We now describe briefly the papers in this volume, followed by concluding remarks.

Lekakos and Giaglis propose a hybrid predictive model that employs a lifestyle model to address the sparse data problem encountered by collaborative models. They introduce the concept of a pseudo user, which combines original ratings provided by a target user with lifestyle information to fill in the missing ratings for this user. This model is extended into an integrated model that creates a pseudo user for each user in the data set. Their experimental framework also includes a meta-learning approach that selects the best model based on past performance and features of the data.

Domshlak and Joachims offer a theoretical formalism for predicting users' preferences that maps qualitative preferences into a high dimensional space. They cast preference prediction as an optimization problem, using a transformation which does not make unwarranted assumptions about the structure of a user's utility function. This problem in turn is translated into a Support Vector Machine formulation.

Chickering and Paek propose a hybrid model that applies techniques from Markov Decision Processes (MDPs) to personalize over time an initial Influence Diagram obtained collaboratively. Specifically, they adapt several explore-vs-exploit strategies used in MDPs to learn the parameters of Influence Diagrams in order to model the behaviour of particular users.

Paek and Chickering use Decision Trees to predict users' commands and contacts in a voice-enabled Command and Control system for mobile phones, and combine these predictions with the output of an automatic speech recognizer. They also perform on-line adaptation for individual users by taking into account the weight of recent experience to adjust the probabilities on the leaves of the Decision Trees.

Guzmán et al. address the problem of on-line adaptation of test items for students. They offer a hierarchical student model that links a conceptual model of a student's expertise in the subject matter with test items and results of tests. They employ Item Response Theory to update the student model, and to decide which test items to pose and when to finish a test, and propose an efficient algorithm for estimating question characteristic curves, which are required to select test items.

Horvitz and Paek apply a decision-theoretic approach to combine human and automated resources in a spoken dialogue system. Specifically, they learn Bayesian Networks to predict the outcome and duration of a dialogue, and apply a decision theoretic approach to devise policies that identify when it is best to transfer a call from a spoken dialogue system to a human receptionist.

Finally, Hollink et al. use mixture models to automatically cluster pages in a web site into navigational stages, according to the order in which the pages are visited by users. They also show how these discovered stages can be used to build problem-oriented menus that reflect coherent navigation patterns in a web site, and how it is

possible to restructure a web site to provide a more useful structure for novice users, without destroying structures that expert users have found useful.

We have come a long way since the early days of applying machine learning techniques and predictive statistical models to user modeling. On one hand, recent developments in machine learning have produced improved tools (e.g., we can now automatically learn the structure of Bayesian Networks, rather than hand-crafting these networks), and inspired novel combinations with techniques from other research areas. On the other hand, user modeling has extended the application of these techniques to new domains and tasks. Although the challenges identified at the beginning of this preface remain, the papers in this volume point towards new ways of addressing them, and have also identified new problems. In the future, we expect that user modeling will continue to benefit from new developments in machine learning and predictive statistical models, while at the same time providing challenges that will motivate new developments in these areas.

References

Burke, R.: Hybrid recommender systems. User Model. User-adapt. Interact. **12**(4), 331–370 (2002)
Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web usage mining as a tool for personalization: A survey. User Model. User-Adap. Interact. **13**, 311–372 (2003)
Salton, G., McGill, M.: An Introduction to Modern Information Retrieval. McGraw Hill (1983)
Webb, G.I., Pazzani, M.J., Billsus, D.: Machine learning for user modeling. User Model. User Adapt. Interact. 19–29 (2001)
Zukerman, I., Albrecht, D.W.: Predictive statistical models for user modeling. User Model. User-Adapt. Interact. **11**(1–2), 5–18 (2001)

Authors' vitae

David Albrecht is a Senior Lecturer in Computer Science at Monash University. He received his B.Sc and Ph.D. degrees in Mathematics from Monash University. His research has spanned several areas of mathematics, including Statistics, Analysis, Optimization, and Logic. Since 1995 he has been working at the Faculty of Information Technology. His current areas of research include plan recognition, machine learning and Bayesian statistics.

Ingrid Zukerman is a Professor in Computer Science at Monash University. She received her B.Sc. degree in Industrial Engineering and Management and her M.Sc. degree in Operations Research from the Technion — Israel Institute of Technology. She obtained her Ph.D. degree in Computer Science from UCLA in 1986. She is interested in the consideration of uncertainty in Natural Language Processing and User Modeling. Her specific areas of research are discourse planning and interpretation, and plan recognition.