Encyclopedia of Information Science and Technology

Second Edition

Mehdi Khosrow-Pour
Information Resources Management Association, USA

Volume VIII
T–Z

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey • New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of information science and technology / Mehdi Khosrow-Pour, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set of books represents a detailed compendium of authoritative, research-based entries that define the contemporary state of knowledge on technology"--Provided by publisher.
ISBN 978-1-60566-026-4 (hardcover) -- ISBN 978-1-60566-027-1 (ebook)
1. Information science--Encyclopedias. 2. Information technology--Encyclopedias. I. Khosrowpour, Mehdi, 1951-
Z1006.E566 2008
004'.03--dc22
2008029068

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

User Profile Modeling and Learning

Evangelia Nidelkou
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Vasileios Papastathis
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Maria Papadogiorgaki
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ben Bratu
Motorola Ltd, France

Myriam Ribiere
Motorola Ltd, France

Simon Waddington
Motorola Ltd, UK

INTRODUCTION

A major theme of Information Science and Technology research is the study of personalization. The key issue of personalization is the problem of understanding human behaviour and its simulation by machines, in a sense that machines can treat users as individuals with respect to their distinct personalities, preferences, goals and so forth. The general fields of research in personalization are user modeling and adaptive systems, which can be traced back to the late 70s, with the use of models of agents by Perrault, Allen, and Cohen (1978) and the introduction of stereotypes by Rich (1979). With the wide progress in hardware and telecommunications technologies that has led to a vast increase in the services, volume and multimodality (text and multimedia) of content, in the last decade, the need for personalization systems is critical, in order to enable both consumers to manage the volume and complexity of available information and vendors to be competitive in the market.

BACKGROUND

The goal of personalization is to endow software systems with the capability to change (adapt) aspects of their functionality, appearance or both at runtime to the particularities of users to better suit their needs. The recent rapid advances in storage and communication technologies stress the need for personalization. This need is more evident in consumer-oriented fields, like news content personalization systems, recommendation systems, user interfaces, and applications like home audiovisual material collection and organization, search engines in multimedia browsing and retrieval systems, providing services for personalized presentation of interactive video content. Among these applications, some are Web-based, but there are also versions for PDAs and mobile devices (Tuoriniemi & Parkkinen, 2007) and mobile devices.

In this article, current approaches of user modeling and user profile representation are discussed, and then the focus is on methods for automatic learning of user models and profiles. The presented learning approaches cover a wide range of machine learning (vector-based or probabilistic) methods and also extend to support the most recent advances in personalization systems such as collaborative filtering, ontology-based user modeling and user social context.

OVERVIEW OF LEARNING AND ADAPTATION METHODS IN PERSONALIZATION SYSTEMS

User Modeling–User Profile Representation

User modeling describes the process of creating a set of system assumptions about all aspects of the user, which are relevant to the adaptation of the current user interactions. This can include user goals, interests, level of expertise, abilities and preferences. The most reliable method of user modeling is by explicit entry of information by the user. In most practical systems, this is too time-consuming and complex for the user. Hence implicit user modeling, based on analysis of past and current user interactions, is critical. The user profile is a machine-processable description of the user model.

The information included in user profiles can be divided into a number of categories such as user demographic information, semantic interests, context and location information, and privacy and user interface preferences (Heckmann & Krueger, 2003). Semantic preferences reflect user preferences for particular content topics. User interests and semantic user preferences are the most important source of information widely used in the personalization systems. More specifically, user interests are distinguished between *short-term* that are determined by a particular user interaction or current context, and *long-term* interests which are determined by the user behaviour and preferences over a longer period of time. User interests can also be classified into *gradual* (as a result of user experience), *abrupt* (as a result of an external stimulus) or *repetitive*. Loeb (1992) mentions two types of repetitive changes, *repetitive but predictable* (according to time of day) and *repetitive but unpredictable* (according to user mood).

There is a variety of structures and paradigms that have been used in the academic literature and in commercial personalization systems for the representation of the knowledge and information concerning the user, including the ones listed below. *Attribute-value pairs* are a fundamental data representation in many computing systems and applications. The advantage of such a structure is that it is an open-ended data structure, thus allowing for future extension without any need for modification. In such situations, all or part of the data model may be expressed as a collection of tuples (attribute name, value), where each element is an attribute-value pair. Several attempts have been put forward to standardize this type of user information structure, such as the IEEE Personal and Private Information (PAPI) (PAPI, 2002) and IMS Learner Information Package (LIP) (IMS, 2001).

The *vector space model* (VSM) is an algebraic model used for information filtering, information retrieval, indexing and relevancy rankings. It resembles the attribute-value pairs, but it has a more mathematical structure, in the sense that each element (term, or generally attribute) has a corresponding value or weight representing it and the vector has length and direction, both used, for example, in a similarity metric. The space of all vectors is often called vector domain or domain model. It has been extensively used in documents retrieval and indexing (Salton, Wong, & Yang, 1975). This representation approach has also been followed in a variety of personalization systems (Billsus & Pazzani, 2000; Lawrence, Almasi, Kotlyar, Viveros, & Duri, 2001; Ricci, Arslan, Mirzadeh, & Venturini, 2002).

One of the earlier representation approaches in user modeling has been the use of *stereotypes*. Stereotyping consists of creating a set of prototypical user profiles that represent the features of classes of similar users (Rich, 1979). Instead of keeping an individual model for each user, users are classified into the stereotypical description that best matches their individual characteristics, from which they inherit additional properties and rules.

The need to automatically learn user profiles has given rise to the use of more complicated representation methods such as the *classifier-based models*. These are based on decision trees, neural networks, inducted rules and Bayesian networks. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch corresponds to one of the possible values for this attribute (Cho, Kim, & Kim, 2002). In contrast to the limited decision trees representation range, artificial neural networks can represent real-valued, discrete-valued and vector-valued functions. The classifier-based models often take as input the *usage history and ratings*. The usage history is a log of the user transaction or interaction with the personalization system, which can be seen as a form of implicit user profile. It is a very practical model used in learning and adapting the user profile (Kang, Lim, & Kim, 2005).

Finally, the recent emerge of the Semantic Web technologies has led to *ontology-based representation* in user profiling. The Semantic Web vision of a next generation Web provides the mechanisms to identify those resources that better satisfy the requests not only on the basis of descriptive keywords but also on the basis of knowledge. The most common ways of representing semantic user profiles are the ontology-based and description logic based representations (Baldoni, Baroglio, & Henze, 2005). In recent work, semantic Web languages, such as Resource Description Framework (RDF), Ontology Web Language (OWL) are used to represent users and their semantic preferences. Gauch, Chaffee and Pretschner (2003) exploit hierarchical structures in ontologies to imply generalizations of user preferences upward in topic hierarchies (e.g., interest in football implies interest in sports).

Automatic Acquisition and Adaptation of User Models–User Preferences

The different representation approaches lead to a variety of methods used for the automatic acquisition and adaptation of user models, which are being presented in this section. Acquisition and adaptation models of user interests is a research area of steadily increasing importance, as it allows intelligent computer systems to adapt to users' information needs in a personalized way. Several machine learning approaches exist to build user profiles, such as Bayesian classifier, nearest neighbor, decision trees, neural networks and genetic algorithms (Pohl & Nick, 1999).

The classifier-based models are related to the use of *Classifier-Based (Statistical) Learning Methods*. The method most widely used in user profile learning is the Bayesian learning, which provides a probabilistic approach to inference. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data. Bayesian learning methods are among the most practical approaches. The Naïve Bayes algorithm is the simplest form of probabilistic model for learning and classifying. It can easily be estimated by training data and in some cases it outperforms other learning methods (Billsus & Pazzani, 2000, pp. 147-180).

Decision tree learning is a method for approximating discrete-values target functions, in which the learned function is represented by a decision tree (Cho, Kim, & Kim, 2002).

In the case of the stereotypical user models, there are different methods of *Learning Stereotypical Sequences of User Interactions* depending on the purposes of the recommender systems: *Collaborative recommendations systems* are based on demographic, geographic and semantic information and they have manually predefined user-stereotypes. Other systems first build a data base of user profiles using the usage log (Azman & Ounis, 2004). In *marketing-based recommendations systems*, the demographic and usage log data can be used to discover rules that capture the truly personal behaviour of a user by means of data mining algorithms. Grouping similar profiles and creating a cluster representative set of rules firstly avoids the privacy problems and secondly reduces the computation during the recommendation process (Wei, Moreau, & Jennings, 2005). Finally, the *social matching recommender systems* match people to each other instead of recommending items to people. The system first creates different set of similar users and then builds a model (i.e., stereotype profile) for each set of users (Terveen & McDonald, 2005).

The user profiles represented with vector space models are related to learning methods based on the user's *relevance feedback*. The term feedback is normally used to describe the mechanism by which a system can improve its performance on a task by taking account of past performance. Adaptive systems using relevance feedback have to choose how relevance feedback should be represented, acquired and used. There are three different methods for representing the relevance feedback. *Boolean relevance* describes whether a document is relevant or not relevant (Objective feedback) vs. useful or not useful (Subjective feedback) to the user. *Multi valued relevance* has been proposed by Bookstein (1983) where the possible relevance classes might be: Very Relevant, Relevant, Indifferent, Irrelevant, Very Irrelevant. In

Quasi-Ordered Relevance, the Document Preference Relation (Wong & Yao, 1990) method relies on a quasi-order of documents. For each pair of documents, the user can either prefer one to the other or have no opinion.

Once the relevance feedback is acquired, there exist multiple formulas that propose to reweight the terms used in the initial query (query reformulation). In vector processing methods, the most commonly used is the Rocchio formula presented in Salton and Buckley (1990), where the new query vector is the vector sum of the old query vector plus the vectors of the relevant and non relevant documents. An extension of this formula proposed by Salton and Buckley (1990), called Ide, eliminates the normalization of the number of relevant and nonrelevant documents and allows limited negative feedback from only the top-ranked nonrelevant document.

In the case of the ontology-based user models, the learning process needs to exploit the deeper ontological knowledge about the underlying domain, thus allowing the personalization systems to handle heterogeneous and complex objects based on their properties and relationships and to automatically explain or reason about the user models or user recommendations. During the learning process, the concepts represented in the ontology are rated according to user-specified preferences such as semantic relevance, syntactic relevance, and categorical match (Kerschberg, Kim, & Scime, 2001).

The learning methods in the ontology-based user models are often enhanced with one of the most expressive and human readable representations for learned hypotheses, which is to use sets of if-then rules. One important special case involves learning sets of rules containing variables. First order rules are more expressive than propositional rules. In general, in many cases it is useful to learn the target function represented as a set of if-then rules that jointly define the function. One way to learn sets of rules is to first learn a decision tree, then translate the decision tree into an equivalent set of rules; one rule for each leaf node in the tree (Mobasher, Dai, Luo, & Nakagawa, 2001). Another method is to use a genetic algorithm that encodes each rule set as a bit string and uses genetic search operators to explore this hypothesis space. There are also a variety of algorithms that directly learn rule sets (Mitchell, 1997).

In all the above-mentioned learning methods, content filtering agents attempt to alleviate information overload by identifying which items a user will find worthwhile. Content filtering focuses on the analysis of item content and the development of a personal user interest profile.

However, to overcome the problem of handling end users solely as units and missing possible information and trends beyond those within the scope of user's history, *collaborative filtering* learning methods are introduced. Collaborative approaches find and recommend information sources for an individual user that have been rated highly by other users who have a pattern of ratings similar to that of the user (Pazzani, 1999). Each technique has advantages and limitations. Current methods for collaborative filtering can be divided into two categories: memory-based, which use all of the available data when making recommendations, and model-based methods which, at some point, learn a statistical model from data and use that model for predicting user interests. Collaborative filtering has a number of advantages over Content-based filtering methods. The quality of memory-based collaborative filtering algorithms typically increases with the size of the user population, and Collaborative filtering recommendations benefit from improved diversity when compared to Content-based filtering recommendations (Claypool et al., 1999). For a start, memory based algorithms are not suitable for recommending new items or one-off content items because these techniques can only recommend items already rated by other users. On the other hand, the model based methods can use smoothing methods to give prior probabilities to items without any ratings. Collaborative filtering matches people with similar interests and then make recommendations on this basis. Collaborative filtering is based on a statistical analysis of patterns and analogies of ratings obtained explicitly or implicitly from user system usage. Typically, for each user a set of nearest neighbors is defined using the correlation between past ratings. Collaborative filtering techniques can be classified to two categories according to the source of information; the user-based and item-based collaborative filtering. The main deficiency of user-based collaborative filtering systems is that they usually make recommendations from very thinly scattered data. In item-based filtering, there are techniques for computing item-item similarities and for obtaining recommendations from them. Linden, Smith, and York (2003, p. 76) follow this approach in Amazon's recommendation system. The main advantage of the item-based approach over the user-based one is its scalability. A combination of the two approaches can be seen in the work of Renda and Straccia (2005).

A further extension of the collaborative filtering approaches is to exploit the *social user context*, which is mainly composed of the user's relationships with other users. Social Information filtering exploits similarities between the tastes of different users to recommend (or advise against) items. It relies on the fact that people's tastes are not randomly distributed: there are general trends and patterns within the taste of a person as well as between groups of people. The basic idea is that the system maintains a user profile, a record of the user's interests (positive or negative) in specific items. It compares this profile to the profiles of other users, and weighs each profile for its degree of similarity with the user's profile. Mika (2005) presents the advances in exploiting the opportunity of semantically-enriched network data.

PERSONALIZATION IN COMMERCIAL APPLICATIONS

Personalization, besides its value in the research field, has also been deemed as an important part of many commercial applications due to the innovation in the services it provides. An example of automatic personalization in commercial systems is Amazon.com's personalized recommendations (Linden, Smith, & York, 2003). Google Inc. has also filed two U.S. patents on personalization technologies for Web search (Badros & Lawrence, 2005; Zamir, Korn, Fikes, & Lawrence, 2005). The Leiki concept aims at combining personalized user interfaces, communities and content targeting (Pennanen & Alatalo, 2001). The Leiki platform is applied as a personalized news service. MovieLens, http://movielens.umn.edu/login, is a free Web-based movie recommendation service provided by the GroupLens research team from the University of Minnesota. It works by matching together users with similar opinions about movies using a collaborative filtering algorithm. TiVo is a television show collaborative recommendation system (Ali & van Stam, 2004, pp. 394-401). The success and innovation of TiVo relies in their personalised television-viewing service, which recommends or automatically records programmes based on user preferences.

FUTURE TRENDS

Mobile ad-hoc networks, wireless broadcasting and open mobile applications are three prominent examples in which computation and communication intermingle with the real world changing the role of context information. Context includes user activities, goals, abilities, preferences, and surroundings.

Current personalization systems do not fully support such flexible and self-adapting models based on context. Thus, future research opportunities within the field of automatic personalization systems include the study of context-aware systems as well as seamless mobility, which is the key future trend in distributed mobile environments. These areas involve research in privacy and sharing of context information and also in the synchronization of user profile between different devices.

CONCLUSION

In this article, the state of the art on the user modeling and user profile representation was presented. More specifically, the standardization and categorization of user profiles was introduced, along with the information included in user profiles and user profile structures. Then, the emphasis is given to the automatic learning of user profiles, where different approaches are being discussed. Automatic learning of user profiles is the current trend in the academic literature and also the key requirement of the current and future commercial personalization systems.

REFERENCES

Ali, K., & van Stam, W. (2004). TiVo: Making show recommendations using a distributed collaborative filtering architecture. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, (pp. 394-401).

Azman, A., & Ounis, I. (2004). Discovery of aggregate usage profiles based on clustering information needs. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04*. New York, (pp. 470-471). ACM Press.

Badros, G. J., & Lawrence, S. R. (2005, June). *Methods and systems for personalized network searching* (U.S. Patent Application 20050131866).

Baldoni, M., Baroglio, C., & Henze, N. (2005). Personalization for the Semantic Web. *Reasoning Web*. *LNCS tutorial* (Vol. 3564, pp. 173-212). Springer-Verlag.

Billsus, D., & Pazzani, M. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction, 10*(2/3), 147-180.

Bookstein, A. (1983). Outline of a general probabilistic retrieval model. *Journal of Documentation*, *39*(2), 63-72.

Cho, Y.H., Kim, J.K., & Kim, S.H. (2002). A personalized recommender system based on Web usage mining and decision tree induction. *Expert Systems with Applications, 23*(3), 329-342.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*, Berkeley, CA.

Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology-based personalised search and browsing. *Web Intelligence and Agent System*, *1*, 219-234.

Heckmann, D., & Krüger, A. (2003). A user modeling markup language (UserML) for ubiquitous computing. In *Proceedings of the 9th International Conference on User Modeling (UM'2003).* Johnstown, PA, USA, (Vol. 2702, pp. 393-397). Springer-Verlag.

IMS Global Learning Consortium, Inc. (2001). *IMS learner information packaging model specification*. Retrieved December 13, 2007, from http://www.imsglobal.org/profiles/lipinfo01.html

Kang, S., Lim, J., & Kim, M. (2005). Statistical inference method of user preference on broadcasting content. *LNCS*. (Vol. 3514, pp. 971-978).

Kerschberg, L., Kim, W., & Scime, A. (2001). A semantic taxonomy-based personalizable meta-search agent. In *Proceedings of the Second International Conference on Web Information Systems Engineering* (WISE'01), (Vol. 1, pp. 41).

Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., & Duri, S.S. (2001). Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery, 5*(1-2), 11-32.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing, 7*(1), 76-80.

Loeb, S. (1992). Architecting personalized delivery of multimedia information. *Communications of the ACM*, *35*(12), 39-48.

Mika, P. (2005). Social networks and the Semantic Web: The next challenge. *IEEE Intelligent Systems*, *20*(1), 80-93.

Mitchell, T.M. (1997). *Machine learning*. McGraw-Hill.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from Web usage data. In *Proceedings of the 3rd ACM Workshop on Web Information and Data Management (WIDM01),* Atlanta, Georgia.

PAPI, IEEE Computer Society. (2002). *Public and private information for learners (PAPI Learners).* Retrieved December 13, 2007, from http://edutool.com/papi/

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, *13*(5/6), 393-408.

Pennanen, P., & Alatalo, T. (2001). *Leiki—a platform for personalized content targeting*. A demo presentation at the ACM Hypertext 2001.

Perrault C. R., Allen, J. F., & Cohen, P. R. (1978). Speech acts as a basis for understanding dialogue coherence. In *Proceedings of the Theoretical Issues in Natural Language Processing-2,* Urbana-Campaign, IL, USA, (pp. 125-132).

Pohl, W., & Nick, A. (1999). Machine learning and knowledge representation in the LaboUr approach to user modeling. In *Proceedings of the 7th International Conference on User Modeling*, Banff, Canada, (pp. 179-188).

Renda, M.E., & Straccia, U. (2005). A personalized collaborative digital library environment: A model and an application. *Information Processing and Management: An International Journal*, *41*(1), 5-21.

Ricci, F., Arslan, B., Mirzadeh, N., & Venturini, A. (2002). ITR: A case-based travel advisory system. In S. Craw & A. Preece (Eds.), *6th European Conference on Case Based Reasoning, ECCBR 2002*, (pp. 613-627). Berlin: Springer-Verlag.

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*, 329-354.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*(4), 288-297.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

Terveen, L., & McDonald, D. W. (2005). Social matching: A framework and research agenda. *ACM Transactions on Computer-Human Interaction, 12*(3), 401-434.

Tuoriniemi, S., & Parkkinen, J. (2007, March). *Voucher driven on-device content personalization* (U.S. Patent Application 7191343B2).

Wei, Y. Z., Moreau, L., & Jennings, N. R. (2005). A market-based approach to recommender systems. *ACM Transactions on Information Systems, 23*(3), 227-266.

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Retrieval*, *41*(5), 334-341.

Zamir, O. E., Korn, J. L., Fikes, A. B., & Lawrence, S. R. (2005, October). Personalization of placed content ordering in search results (U.S. Patent Application 20050240580).

KEY TERMS

Acquisition and Adaptation of User Profiles: The automatic creation and adaptation of a user profile by monitoring the user interaction with the system and employing efficient machine learning algorithms.

Collaborative Filtering: The method of making automatic predictions (filtering) about the interests of a user by collecting information from other similar users.

Content-Based Filtering: The method of making recommendations to a user by matching user profile entries to content attributes.

Machine Learning: The method for processing a training input and offering support for decision based on this input.

Personalization: Delivery of content according to the individual user's needs, characteristics and preferences.

User Modeling: The process of creating a set of system assumptions about all aspects of the user, which are relevant to the adaptation of the current user interactions.

User Profile: A machine-processable description of the user model.