Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities

Georgios Paliouras[1], Vangelis Karkaletsis[1],
Christos Papatheodorou[2], Constantine D. Spyropoulos[1]

[1]Institute of Informatics and Telecommunications, [2]Division of Applied Technologies,
National Centre for Scientific Research (NCSR) "Demokritos", 15310, Aghia Paraskevi Attikis, Greece.
E-mail:{paliourg, vangelis, costass}@iit.demokritos.gr, papatheodor@lib.demokritos.gr

Abstract. In this paper we examine the acquisition of user stereotypes and communities automatically from users' data. Stereotypes are built using supervised learning (C4.5) on personal data extracted from a set of questionnaires answered by the users of a news filtering system. Particular emphasis is given to the characteristic features of the task of learning stereotypes and, in this context, the new notion of community stereotype is introduced. On the other hand, the communities are built using unsupervised learning (COBWEB) on data containing users' interests on the news categories covered by the news filtering system. Our main concern is whether meaningful communities can be constructed and for this purpose we specify a metric to decide on the representative news categories for each community. The encouraging results presented in this paper, suggest that established machine learning methods can be particularly useful for the acquisition of stereotypes and communities.

1 Introduction

User modeling technology 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, adaptive user interfaces, tutoring systems. In this paper we examine the exploitation of machine learning techniques in user modeling technology for news filtering services. More specifically, we examine the organisation of the users of a news filtering system into groups with common characteristics (*stereotypes*) and groups with common interests (*communities*). The choice of the appropriate learning techniques, the use of stereotypes or communities, as well as the construction of meaningful communities are some of the important issues examined in this paper.

Stereotypes have been widely used in user modeling, but their construction has been almost exclusively manual (Brajnik & Tasso, 1994 and Kay, 1995). Attempts to automate the acquisition of stereotypes have been limited to the adaptive refinement of numeric parameters, rather than the construction of the stereotype (Rich, 1983). The manual construction process usually involves the classification of users by an expert and/or the analysis of data relating to the interests of individual users. Acquiring the stereotypes in this way is a difficult task. Similar difficulties have been encountered in other classification tasks and one solution that has yielded positive results is the automatic acquisition of knowledge, using machine learning. The work presented here focuses on the characteristics of the task of learning stereotypes and introduces the new notion of community stereotype. The method examined (*C4.5*) performs *supervised learning* from personal data extracted from questionnaires answered by the users of a news filtering system. These are data about the company the user is working in (type, department, location, size, market) and his interests on specific news categories.

The lack of sufficient personal data about the users of the news filtering system led us to the use of community modeling. Communities are built from data containing only the users' interests on news categories. These interests are determined by the users themselves. In this paper we examine *unsupervised learning* (*COBWEB*) for the acquisition of user communities. The resulting communities can be used to improve the services provided by the news filtering system. However, this can be done effectively only when the communities are meaningful, that is if they associate users with a limited set of common interests. For this reason we use a metric to decide which news categories are most representative for each community.

The work presented in this paper has been performed in the context of the research project ECRAN project (Language Engineering 2110, Telematics Applications Programme) which focuses on the adaptation of information extraction systems to new domains and users. Section 2 of the paper explains how machine learning techniques can be exploited for the acquisition of user stereotypes and communities and describes the learning algorithms that were applied in this work. Sections 3 and 4 present the setting of the two experiments for stereotype and community acquisition respectively and discuss the experimental results. Finally, section 5 describes ongoing work and introduces our plans for future work.

2 Learning User Stereotypes and Communities

Machine learning methods have been applied to user modeling problems mainly for acquiring models of individual users interacting with an information system (Bloedorn et al., 1997, Chiu, 1997 and Raskutti & Beitz, 1996). In such situations, the use of the system by an individual is monitored and the collected data are used to construct the *user's model*, i.e., his

individual requirements. We are concerned with a higher level of generalisation of the users' interests: the construction of user stereotypes and communities. This task requires the application of learning techniques to user models, which are assumed to have been constructed by a separate process, either manual or automatic.

The choice of learning method depends largely on the type of training data that are available. The main distinction in machine learning research is between *supervised* and *unsupervised* learning. *Supervised learning requires the training data to be preclassified.* This means that each training item (*example*) is assigned a unique label, signifying the class in which the item belongs. In our case, this would mean that each user model must be associated with a class label out of a set of possible classes that have been defined beforehand. Given these data, the learning algorithm builds a characteristic description for each class, covering the examples of this class, i.e., the users belonging to the class, and only them, i.e., none of the users of other classes. The important feature of this approach is that the class descriptions are built conditional to the preclassification of the examples in the training set. *In contrast, unsupervised learning methods do not require preclassification of the training examples*. These methods form clusters of examples, which share common characteristics. When the cohesion of a cluster is high, i.e., the examples in it are similar, it defines a new class.

2.1 Learning User Stereotypes

The problem of learning user stereotypes cannot easily be categorised as a supervised or an unsupervised learning task. This is due to the nature of user models, i.e., the training data. Typically, user models contain two types of information for the user: personal characteristics and the user's requirements from the system. In a news-filtering system, the personal characteristics for the user could be age, education and occupation, while the user's requirements are news categories that the user is interested in. The former type of information is *system-independent*, while the latter is *system-dependent*.

The peculiarity of the task of learning user stereotypes stems from the fact that each user model is not labelled by a user class. In Fig.1 the class label is shown in a shaded box, signifying that the label is latent information, not provided in the data. However, this information acts as a link between the system-independent and the system-dependent information. The system-dependent information can be used to build a description of a user community, while the system-independent information to associate a stereotypical behaviour with that community. If the class was known for each user, then the task of learning the stereotypes could be split into two supervised learning subtasks: learning the description of a user class and learning the stereotypical behaviour of a class. Lacking the information about the class label, this two-stage approach requires the combination of an unsupervised with a supervised learning method. For instance, using an unsupervised learning method to form communities and a supervised learning method to learn the stereotypical behaviour of each community. Henceforth, this approach will be referred to as *community stereotype learning*. The symmetric approach, i.e., performing unsupervised learning on the system-independent information, is less interesting, because it contradicts the goal of maximising the similarity of user requirements within each user class.

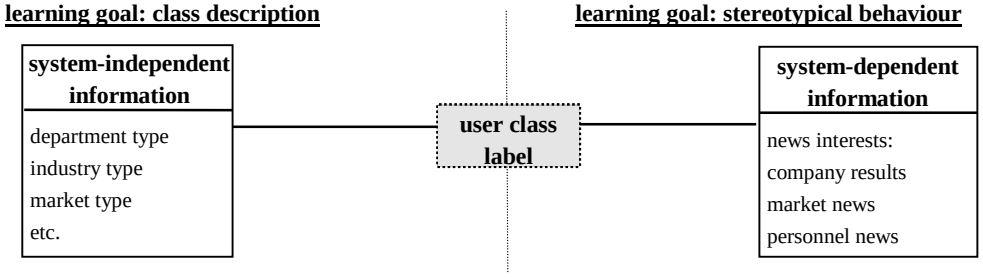

Figure 1. Information available in the training data and the role of the latent user category.

An alternative approach that is adopted here is to ignore the latent class label and use the system-dependent information as the classification data. Thus, *instead of classifying users into communities, the goal is to learn the description of the user classes associated with each of the system-dependent pieces of information*. For instance, in the news-filtering task the goal of the learning method is to construct a user class for each news category. The class description is in terms of the system-independent information, e.g. "people who work in the marketing department of financial companies are interested in company results". Such a class would ideally cover all people who are interested in company results and none of the others. This approach is a valid alternative to community stereotype learning. Its main advantage is that the construction of user classes is driven directly by the system-dependent information. There is no need for the construction of intermediate concepts, such as the user communities. In this manner, more cohesive user classes can be constructed.

The machine learning method used in this study is called *C4.5* (Quinlan, 1993) and performs induction of *decision trees*, i.e., it constructs decision trees from training data. In the case of stereotype learning, each decision tree corresponds to the stereotype for one system-dependent variable, e.g. a news category. Figure 2 shows a decision tree focusing on users' interest in company results.

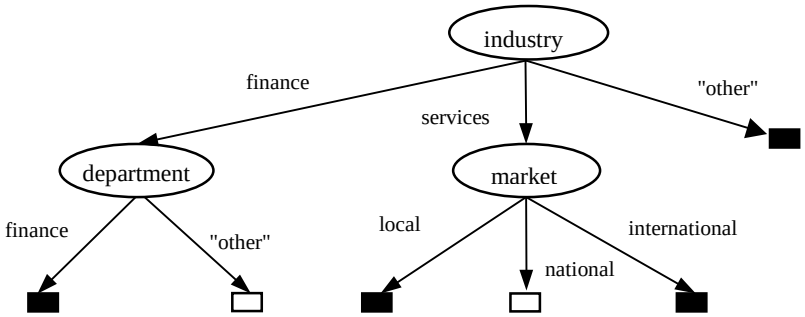

Figure 2. Part of the decision tree representing stereotypes for the company results category (clear boxes indicate interest in the category while filled boxes represent lack of interest)

The decision tree of Fig. 2 could be interpreted into the following simple stereotype rule for the company results category:

IF (industry = finance AND department ≠ finance) OR (industry = services AND market = national)
THEN AND ONLY THEN the user is interested in company results.

Similar decision trees can be constructed for other news categories. The resulting set of trees is the set of stereotypes, which can be used to determine the interests of a user, based on his personal characteristics. The construction of multiple decision trees is atypical of work on decision tree induction. However, this approach is necessary here, since the goal is not to discriminate between the various news categories, but between people being interested in each category and those who are not.

2.2 Learning User Communities

User communities can be constructed automatically using an unsupervised learning method. Unsupervised learning tasks have been approached by a variety of methods, ranging from statistical clustering techniques to neural networks and symbolic machine learning. In this work, we have opted for the symbolic learning methods, because we are interested in the comprehensibility of the results. The branch of symbolic machine learning that deals with unsupervised learning is called *conceptual clustering* and a popular representative of this approach is the algorithm COBWEB (Fisher, 1987). Conceptual clustering is a type of learning by observation that is particularly suitable for summarising and explaining data. Summarisation is achieved through the discovery of appropriate clusters, which involves determining useful subsets of an object set. In unsupervised learning, the object set is the set of training examples, i.e., each example is an object. Explanation involves concept characterisation, i.e., determining a useful concept description for each cluster.

COBWEB is an incremental algorithm that uses hill-climbing search to obtain a concept (cluster) hierarchy, partitioning the object space. The term *incremental* means that objects are incorporated into the concept structure as they are observed. An object is a vector of feature-value pairs. In our case, objects are user models and features are the news categories taking the values true and false for each user. Each concept in the hierarchy produced by COBWEB, is a probabilistic structure that summarises the objects classified under that concept.

In order to construct the clusters, COBWEB uses *category utility* (Gluck & Corter, 1985), which is a probabilistic measure of the usefulness of a cluster. COBWEB incorporates objects into the concept hierarchy using four clustering operators: placing the object in an existing cluster, creating a new cluster, combining two clusters into a new one (merging) and dividing a cluster (splitting). Given a new object, the algorithm applies each of the previous operators and selects the hierarchy that maximises category utility.

3 Case Study I: Learning User Stereotypes

3.1 Experimental Setting

In the framework of the ECRAN project, we developed a prototype user modeling module (UMIE, *http://www.iit.demokritos.gr/UMIE*) that filters, according to the user's interests, the facts extracted by the ECRAN information extraction system (Benaki et al., 1997). The interests of an individual user are stored in his user model. In addition, the system uses a set of stereotypes, in order to anticipate the interests of a new user.

At a first stage, the stereotypes were built manually by analysing questionnaires answered by the users. The questionnaires contained information about the company the user is working in and the news categories that the user is interested in. Thirty-one questionnaires were analysed and it was decided that only one feature was informative of the interests of the user: the department in which he/she works. Nine stereotypes were constructed, corresponding to nine different company department types, and each was associated with a list of news categories, covering the interests of all the users in the stereotype. This manual acquisition process was difficult and the generated stereotypes did not prove particularly useful. The first

difficulty was the interpretation of the questionnaire data. The number of features which could be extracted from the data was very small and concerned only company information. Moreover, there was a lot of missing - or hard to interpret - information, even regarding the interests of the user. The main problem was the extent of overlap between the stereotypes. There were five news categories and each stereotype contained in average 3.3 of these. Thus, it is clear that this classification is hardly more useful than the base rule, which says that everybody is interested about everything. This is an indication that the stereotypes are oversimplified.

At a second stage, we used a machine learning method for stereotype acquisition. Replacing the manual process by a machine learning method cannot solve the problem of the quality of information that can be extracted from the questionnaire data. However, it may provide a solution to the problem of oversimplification. A set of experiments, using C4.5, was done on the news-filtering data. The training data were extracted from the set of the 31 questionnaires. In this study there are five system-independent variables:

- The **department** in which the user is working: personnel, development, planning, finance, marketing, information support, consulting, public relations, sales.
- The type of **industry**: construction, manufacturing, financial, wholesale, retail, public services, public administration, research and education, services.
- The **size** of the company: small, medium and large.
- The **location** of the company: local, national and multinational.
- The location of the **market** for the company: local, national and international.

In many cases, the value of some of the above variables could not be decided from the questionnaires. The amount of missing information increased moving down the list of the five variables, i.e., the department type was available in all cases, while there were many cases, in which the company's market was not obvious. C4.5 handles missing information in a probabilistic manner. It fills the missing value by all possible values, attaching a weight to each one. This weight depends on the proportion of cases having this value in the training set. The news categories used in this study were the following: business development, product news, market news, company results and personnel news.

3.2 Results

Two initial experiments were run using C4.5. In the first experiment, all variables except the department, were ignored. The aim of this experiment was to see the association of news categories with the department type and compare these results with the manually constructed stereotypes of the ECRAN project. The learning task involved a simple calculation of the number of cases in which a news category was selected for each department type. These numbers are provided in Table 1. The shading of the table cells corresponds to the presence or absence of the news category in the manually constructed stereotypes. Shaded cells represent combinations appearing in the manual stereotypes. When a cell is empty it means that no cases were found in which the corresponding department type was associated with the respective news category.

Table 1. Associations between department type and news categories in the stereotypes.

department	business development	product news	market news	Company Results	personnel news
Personnel					
Development		1.0			
Planning			1.0	1.0	
Finance			1.0		
Marketing	0.55	0.45	0.92	0.92	
Information support	1.0	1.0			
Consulting	0.66	0.33	0.66	0.66	0.11
Public relations	0.33		1.0	1.0	
Sales		1.0	1.0	1.0	

The first important difference between the manually constructed stereotypes and the ones generated by C4.5 was in terms of generality. Excluding the personnel information, which appeared only once in the data, the manually constructed stereotypes defined 26 out of the 32 possible associations. Only 20 of these are verified in the training data and many of these are very weak. Thus, the automatically constructed stereotypes are more specialised and therefore seem more helpful in deciding the information that one is interested in. Furthermore, the association weights given in Table 1 could be used to prioritise the interests of the user. A similar approach was adopted in (Rich, 1983). These results should be interpreted with caution, due to the small size of the training set. A larger training set could provide more interesting and robust stereotypes.

In the second experiment, all five system-independent variables were used. Out of these the ones selected most often by C4.5 were the department and industry type, suggesting that these two variables have a larger classification power. In general, relatively small decision trees were constructed. The tree of Fig. 2 is one of the five trees that were generated.

The important problem with the induced decision trees was their low accuracy in discriminating between people who are interested in a news category and those who are not. Table 2 presents the accuracy of each of the four trees (excluding the personnel news category). For the sake of comparison, the accuracy of the default rule, classifying everybody in the majority

class is presented. The majority class is usually those not interested in the news category. The exception to this rule is the market news category.

Table 2. Performance of the induced trees.

category	accuracy of stereotype	default accuracy
business development	78.6%	50.0%
products news	82.1%	60.7%
market news	92.9%	78.6%
company results	85.7%	71.4%

Despite the significant improvement over the default rule, there is still much to be desired by the decision trees. Moreover, it should be noted that these results are calculated on the training set and may give a distorted picture of the performance of the stereotypes. More thorough testing is required, involving a standard testing method, such as the n-fold cross-validation (*leave-one-out*), which is the most appropriate for small datasets.

4 Case Study II: Learning User Communities

4.1 Experimental Setting

We applied COBWEB on the task of constructing user communities for the news-filtering system (Paliouras et al., 1998). The news articles are organised into 24 news categories, e.g. economic indicators, computers, etc. During his/her registration to the news filtering system, each user specifies a subset of these news categories, which correspond to his/her personal interests. This personal list of news categories constitutes the user model, which determines what news he receives. The user model can be modified by the user, reflecting changes in his interests. The dataset for the experiment contained 1078 user models, with an average of 5.4 news categories specified in each model. These user models formed a set of training examples for the learning algorithm. Each example was represented as a binary feature vector, specifying which news categories the user was interested in. Given these training examples, COBWEB constructed classes of users with common interests, i.e., communities.

The question is whether there is any meaning in the generated communities. Since there is no personal information available about the users, the construction of stereotypical descriptions for the communities is not possible. The natural way to construct meaningful communities is by trying to identify those sets of interests on news categories that are representative for the participating users. To achieve this, we specified a metric FI_c that measures the increase in the frequency of a category "c" within a community, as compared to the default frequency of the category in the whole data set (Paliouras et al., 1998). When FI_c is negative there is a decrease in frequency and the corresponding category is not representative for the community. The definition of a representative news category for a community, is just that $FI_c > \alpha$, where α is the required extent of frequency increase. In order to see the impact on the characterisation of the communities, we varied α and measured the following two properties of the generated community descriptions:

- *Coverage:* proportion of news categories covered by the community descriptions. Some of the categories will not be covered, because their frequency will not have increased sufficiently.
- *Overlap:* amount of overlap between the constructed community descriptions. This is measured as the ratio between the total number of categories in the description and the number of distinct categories that are covered.

4.2 Results

Using COBWEB on the data set generated a concept hierarchy of which the first three layers are presented in Fig. 3. An important property of the hierarchy is the balanced split of objects in different branches. Therefore the underlying concepts are of similar strength.

Ideally, we would like to acquire descriptions that are maximally distinct, i.e., minimise the overlap, and increase coverage. We examined two different partitions of the objects, corresponding to the second and the third level of the concept hierarchy (for details see Paliouras et al., 1998). The coverage on the second level of the hierarchy is consistently lower than that on the third level. This is because the clusters on the third level are more specialised than those on the second. On the other hand, the larger the number of the communities, the larger the overlap between their descriptions. Thus, there is a trade-off between coverage and overlap, as the number of communities increases. In the case of the 15 different clusters, the extent of the overlap is significantly higher than in the 5-cluster case which gives the best results in terms of the distinctness in the descriptions. A set of very concise and meaningful concept descriptions were acquired for the 5-cluster case. Table 3 lists the news categories for the 5 clusters, together with their FI_c values.

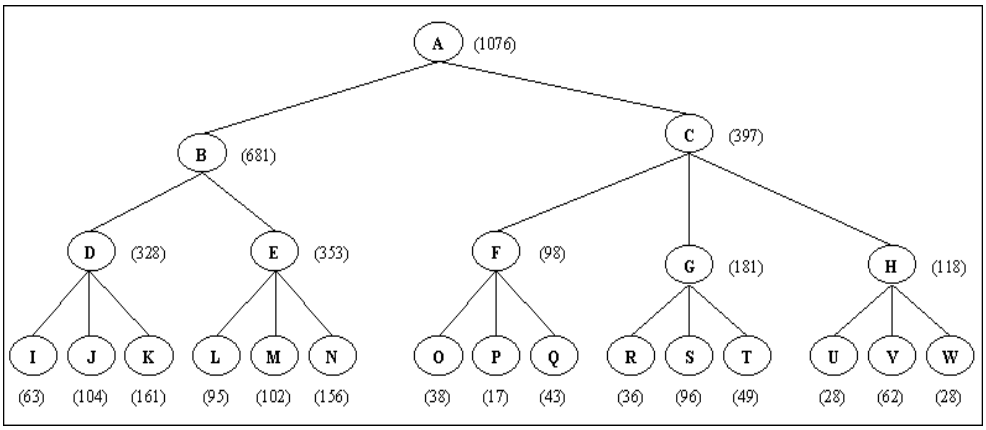

Figure 3. Hierarchy generated by COBWEB (top three levels). The numbers in brackets correspond to the size of the corresponding subset of objects.

Communities E, G and H are well-separated, corresponding to a group of people interested in the internet, economics and computers respectively. Community F, consists of people interested mainly in economics and finance, but also in computers. Some interest in computers is to be expected from the users of a system on the internet. Finally, community D serves as a "miscellaneous" cluster that is not homogeneous. The common feature of people in D, is that they have very specific interests, leading to sparse user models.

Table 3. Descriptions for the 5 clusters generated on the second level of the concept hierarchy

D	E	F	G	H
	Internet (0.55)	Economic Indicators (0.73)	Economic Indicators (0.58)	Computers (0.53)
		Economy & Finance (0.68)	Economy & Finance (0.61)	
		Computers (0.6)		
		Transport (0.53)		
		Financial Indicators (0.5)		

A problem with the descriptions in Table 3 is that a large proportion of the 24 news categories are not covered. In general, these are the categories that are chosen by either too few or too many users. In the former case the algorithm ignores them during learning and in the latter case, they correspond to such general interests, that they cannot be attributed to particular communities. Filtering out these two types of category is a positive feature of the FI_c metric. Coverage can increase, by moving selectively to lower levels of the COBWEB hierarchy. For instance, the children of node H give meaningful and concise communities, corresponding to categories that are related to computing, e.g. electronics and networks. However, this is not the case for all five communities in Table 3, e.g. the children of node E do not provide further meaningful subgroups in the community. The ability to select nodes at different levels of the concept hierarchy is an important advantage of the COBWEB algorithm.

5 Conclusions

This paper presented a pilot study on the acquisition of user stereotypes and communities from user-modeling data. The choice of the appropriate learning techniques, the use of stereotypes or communities, as well as the construction of meaningful communities were the major issues examined.

The notion of a stereotype and the problems of manual construction of stereotypes were first discussed. Particular emphasis was given to the characteristic features of the task of learning stereotypes and in this context the new notion of community stereotype was introduced. Experimental results, using C4.5 show that the task is solvable, but raise a number of issues. The most important one is the quality of the data. The data that were extracted from the questionnaires did not prove adequate for learning robust and general stereotypes. The lack of sufficient system-independent information led us to community modeling.

Communities can be used to improve the exploitation of a news-filtering system by its users. The construction of the communities was achieved using an unsupervised learning technique. We also proposed a metric to decide which are the representative news categories of a community. Our results are very encouraging, showing that meaningful community descriptions can be generated.

Concluding, this paper has shown that the application of learning methods to user-modeling data for the construction of stereotypes and communities is a promising research direction. We consider this work as a first step towards a methodology that can easily be integrated in a variety of news-filtering systems. A number of issues have arisen, which are of both theo-

retical and technical interest. Taking into account the increasing commercial interest of user modeling, especially since the advent and expansion of the WWW, we believe that this work could help considerably the construction of useful tools.

References

Benaki, E., Karkaletsis, V. and Spyropoulos, C. D. (1997). Integrating User Modeling into Information Extraction: the UMIE Prototype. In *Proceedings of the Sixth User Modeling Conference*, 55-57.

Bloedorn, E., Mani, I. and MacMillan, T. R. (1997). Machine Learning of User Profiles: Representational Issues. In *Proceedings of the National Conference on Artificial Intelligence*, 433-438.

Brajnik, G. and Tasso, C. (1994). A Shell for Developing Non-monotonic User Modeling Systems. *International Journal of Human-Computer Studies* 40:31-62.

Chiu, P. (1997). Using C4.5 as an Induction Engine for Agent Modeling: An experiment of Optimisation. In *Sixth User Modeling Conference, Workshop on Machine Learning for User Modeling.*

Fisher, D. H. (1987). Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning* 2: 139-172.

Gluck, M. A. and Corter, J. E. (1985). Information, Uncertainty and the Utility of Categories. In *Proceedings of the 7th Conference of the Cognitive Science Society*, 283-287.

Kay, J. (1995). The um Toolkit for Cooperative User Modeling. *User Modeling and User Adapted Interaction* 4:149-196.

Paliouras, G., Papatheodorou, C., Karkaletsis, V., Spyropoulos, C., and Malaveta, V. (1998). Learning User Communities for Improving the Services of Information Providers, Lecture Notes in Computer Science, 1513 : Springer-Verlag. 367-384.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Kaufmann.

Raskutti, B. and Beitz, A. (1996). Acquiring User Preferences for Information Filtering in Interactive Multi-Media Services. In *Proceedings of PRICAI*, 47-58.

Rich, E. (1983). Users are Individuals: Individualizing User Models. *International Journal of Man-Machine Studies* 18:199-214.