Combining Stereotypes for Robust Information Prioritization

Zoë Lock[1,2] and Daniel Kudenko[2]

[1] QinetiQ, Malvern Technology Centre, St Andrew's Road, Malvern, WR14 3PS, UK
zplock@qinetiq.com
[2] Department of Computer Science, University of York, Heslington, York, YO10 5DD, UK
{kudenko, zlock}@cs.york.ac.uk

Abstract. In agile team settings, such as military command, flexible user modeling is required to respond to major shifts in user requirements triggered by changes in team membership or role assignment. Single-component user models are not robust enough for such situations. In this paper, we describe and evaluate a modular user modeling approach, more appropriate to use in agile team settings. We show that this approach performs as well as a single-component approach in terms of average precision while being more portable, to new circumstances, for many users. We also show how the new approach can address the new user problem for many users.

1 Introduction

In this paper, we describe an investigation into the robustness of a modular user modelling approach for textual information prioritization[1]. Our work is driven by the need for an effective information prioritization tool for use in agile military environments in which a user's team membership, role assignment and task allocation, all of which dictate the user's information requirements, are subject to wide-reaching and abrupt changes. We believe that flexibility is an important attribute of user models alongside mobility and pervasiveness in that the user models used to prioritize information in a personalised manner remain useful despite such changes.

Most user modeling approaches represent the attributes of an individual user within a single, monolithic structure. Such a model is brittle in response to sudden and significant changes in the circumstances of the user. In addition, this single-component approach suffers from the new user problem [1]: when a new user joins a team, there are no examples of how he rates items. The user will then be required to provide examples with which to train a new model from scratch, even though his role and team membership might already provide useful information that could serve to reduce the training overhead.

To overcome the above limitations of the monolithic approach, a modular approach is required. Rich introduced *stereotypes* as components that could be combined to

[1] This work was carried out as part of the UK Ministry of Defence Corporate Research Programme.

construct modular user models [2]. A stereotype is itself a model of the common attributes or interests of a group of users. Once a user has been tagged with a stereotype, some of his attributes can be inferred with minimal information from the user. Users can be tagged with multiple stereotypes and if these differ in their knowledge of a particular trait of a user then this is resolved automatically using the confidence the stereotypes have in their knowledge.

Our investigations focus on the robustness of modular user models for text prioritization, based on stereotypes, to changing requirements and new users. In our work, stereotypes correspond to teams and roles, which are subject to sudden changes, rather than personality traits such as *feminist* or *sporty*, which are normally more persistent and may change gradually over long periods of time.

The modular approach has the advantage that when a perspective changes, the corresponding stereotype can be replaced with another that is more appropriate to the new situation. All other stereotypes, assumed to be unaffected by the perspective change, remain in the overall user model unchanged. The updated user model can then be used to rate new items without the need to retrain a new user model from scratch. Perspective changes are discussed further in section 5.2.

Another advantage is that existing stereotypes can be used to construct an initial user model for a new user. In this way, the new user problem can be alleviated using a modular approach as long as the stereotypes of the new user are known. Our approach to the new user problem is described in more detail in section 5.3.

The paper is structured as follows. Section 2 is a short overview of relevant modular modeling approaches. Section 3 describes our approach to learning modular user models. Section 4 details the two data sets used for our investigation. Section 5 outlines the results of the experiments. Finally, the conclusions and an outlook to future work are presented in section 6.

2 Related Work

Within the field of user modeling, some modular user models have been developed in which the interests of a user is divided into multiple topics of interest (such as football or local news), each of which can be represented by a separate model component. In most cases, only one component is used at any one time and so the relevant component must be selected to suit the current query or interest [3,4]. In our approach, a set of components will all contribute towards the rating of a new item. Baudisch and Brueckner describe a system in which the outputs of multiple queries are fused, however, these queries are not shared between multiple users so do not constitute stereotypes [5]. In another TV program recommendation system, household interest models are represented within a modular system but each stereotype contributes to the same extent to the overall interest of all users [6] so the system performs poorly for household members who do not share many interests with their housemates. More details about these modular approaches can be found in [7].

More recently, group modeling approaches that capture the attributes of multiple users have been subject to investigation. Many of these are used to make classification or ranking decisions for a group rather than the individual user. In

effect, they construct stereotypes that perform well for a group of users. For example, Masthoff has developed a mechanism for constructing a TV program schedule for a group of users from their individual ratings of the programs [8].

None of the user modeling approaches use teams and roles to construct stereotypes and modular user models.

Within the field of machine learning, the use of committees or ensembles of models is an active research topic. The central motivation for this is that models formed by integrating multiple models can be more accurate than the individual component models. There are now many different approaches to learning and integrating multiple models including stacking [9] and delegation [10]. Many of these approaches either involve training models using different classification algorithms or training models of the same type with different parts of the training set. The outputs of these models are then combined or fused in some way to output a single classification or ranking decision. Rather than accuracy, our central motivation for combining models is to increase the robustness and flexibility of user models in the face of changing requirements.

3 Our Modular User Modeling Approach

This section outlines our approaches to constructing both single-component and modular user models. User models and stereotypes are automatically derived from a set of text items, binary relevance feedback on those items from users and the team and role assignments of those users using text analysis and machine learning.

Firstly, we perform feature selection in order to reduce the size of the data set and to construct initial single-component user models and stereotypes. In our text domain, features correspond to words appearing in the data set documents. In the case of a single-component user model, feedback of an individual is used to train the model. In the case of a stereotype, feedback from multiple users is used[2]. This task involves three steps:

1. standard stop words are removed from each item;
2. a statistical metric is applied to the set to score each remaining word according to its indication of relevance in a training data set[3];
3. the top n scoring words are used to seed the stereotype[4] and their scores are normalized to sum to 1.

For a fair comparison between modular and single-component user models, each type is trained to contain the same total number of features. If the number of features selected for each stereotype in a user's modular user model is n then the number of features selected for a single-component user model is n * the number of stereotypes that apply to the user. All stereotypes and single-component user models in the

[2] Handling disagreements between stereotype members in an important issue but is not be discussed here during to space constraints.

[3] We have used the χ^2 statistical metric (numerous metrics were compared in [11])

[4] Unless otherwise stated, n = 10 as preliminary experiments indicated that the value gave most consistent performance.

experiments are learnt using same technique to remove confounding systemic differences (an important issue noted in [12]).

To rate an item according to its relevance to a trained single-component user model or single stereotype, the weights of any of the *n* words contained within the model or stereotype and appearing in the item are summed together to obtain a single relevance rating between 0 and 1.

A modular user model consists of a set of stereotypes and their associated personalized *inter-component weights*. The inter-component weights (between ±1.0) indicate the relative contributions the stereotypes make to overall relevance of items to a user. The user's relevance feedback is used to train the inter-component weights: the weights are all initialized to 0 then a simple gradient descent algorithm [13] is used to train their initial values so that the inter-component weights of stereotypes that contribute most to accurate overall relevance ratings are higher than others. A new, unseen text item is rated by a modular user model by computing the linear weighted sum of the ratings of its stereotypes.

We claim that the modular models:

1. are comparable in prioritization performance to single-component user models;
2. are robust to changes in user perspectives;
3. can be used to alleviate the new user problem.

4 Evaluation Data Sets

The few publicly available relevance feedback data sets on the WWW that can be used by researchers to evaluate their information prioritization systems do not involve to team settings so they do not declare the teams and roles of users so are not directly applicable to our approach. In order to stimulate experimentation, other relevance feedback data was sought. We obtained two sets of data; one from a military source and one from a non-military source. The data sources are described below. Users were asked to rate an item as relevant if he would make use of any of the information contained within it when performing his declared roles and irrelevant otherwise.

The military data came from an experimental planning and execution task for two UK Joint Force Component Headquarters teams: Land and Air. Both teams contained five team members with the following roles: Chief-of-Staff; Intelligence officer; Plans/Operations Officer; Logistics Officer; Liaison Officer. Each role was assumed by exactly one participant and each participant belonged to one of the two teams. An ex-military officer provided a set of 133 realistic text documents for the experiment. Explicit, binary feedback on the text items was obtained from the 10 users.

A research group at QinetiQ has its own Intranet environment in which pages may be added, edited and viewed by group members. The environment is used to store and share useful information concerning research projects and other technical matters. 84 pages about diverse projects and topics were taken from the Intranet and used as information items. Group members work in several different project teams and adopt different roles within those teams. Explicit, binary relevance feedback was obtained from 13 users.

5. Results

This section details the results of our experiments to assess the three claims in section 3. The primary evaluation measure for our investigation is average uninterpolated precision or average precision (AP). This metric assesses how good a user model is at pushing relevant items to the top of the ranking above irrelevant ones (prioritization):

$$AP = \frac{1}{N}\sum_{i=1}^{N} P_i(r)$$

where $P_i(r)$ is the precision at relevant retrieved document i and N is the number of relevant documents retrieved by the ranking algorithm (true positives). A score of 1.0 means that all relevant items are ranked above all irrelevant items.

Stratified ten-fold cross validation [14] was used for all the experiments described in this paper. Paired t-tests have been used to analyze the differences between the performances of different models ($p < 0.05$ is deemed significant).

5.1 Comparison Between Single-Component and Modular User Models

The first experiment was run to assess claim 1: that modular user models can provide the same levels of performance than that of single-component ones. Table 1 and Table 2 show the AP results for the Intranet and military users for the single-component approach and the modular approach. For both data sets the performances of the trained modular models are comparable to that of the single-component models.

Table 1. Performances of single-component (SC) and modular approaches (MOD) for the military data domain

User	SC	MOD
Air – Intelligence Officer (A2)	0.97	0.98
Air – Plans/Operations Officer (A3)	0.51	0.43
Air – Logistics Officer (A4)	0.78	0.83
Air – Chief-of-Staff (AC)	0.40	0.41
Air – Liaison Officer (AL)	0.72	0.66
Land – Intelligence Officer (G2)	0.96	0.92
Land – Plans/Operations Officer (G3)	0.41	0.50
Land – Logistics Officer (G4)	0.65	0.52
Land – Chief-of-Staff (GC)	0.45	0.37
Land – Liaison Officer (GL)	0.82	0.77
Average	0.67	0.64

Table 2. Average precision values of the single-component (SC) and modular user models (MOD) approaches for the Intranet data domain

User	SC	MOD
1	0.71	0.76

2	0.75	0.73
3	0.79	0.81
4	0.81	0.82
5	0.63	0.45
6	0.80	0.79
7	0.45	0.58
8	0.92	0.97
9	0.37	0.45
10	0.74	0.79
11	0.54	0.41
12	0.58	0.54
13	0.65	0.69
Average	0.67	0.68

5.2 Robustness to Perspective Changes

In section 3, we claimed that our modular user models are robust to major shifts in a user's information requirements caused by changes to team membership or role assignment. The advantage of this feature is a more rapid response to a new situation as the need for training is alleviated. In this section, we present results to empirically support the claim. Ideally, to test the robustness of modular user models, relevance feedback should be collected from user experiments in which users do indeed change team or role whilst rating items. The military and Intranet experiments did not involve such changes so we have used the existing data to simulate the situation in which information requirements change abruptly.

Consider two users, U1 and U2, where U1's model contains stereotypes C1, C2 and C3 and U2's model contains stereotypes C1, C2 and C4. U1 and U2 have stereotypes C1 and C2 in common but each has one stereotype the other does not (C3 and C4). If C3 and C4 represent team stereotypes and U1 and U2 swap teams, then the new model for U1 should contain C1, C2 and C4 stereotypes and the new model for U2 should contain C1, C2 and C3 – making U1's new model the same as U2's old model and vice versa. If it is assumed that a user's requirements depends only on his team and role assignment (which is reasonable given the rating criteria used by the users given in section 4), then after stereotype swapping, U1's new requirements are now equivalent to U2's old requirements and vice versa. Based on this assumption, U1's new model can be tested on U2's test feedback, which was obtained before team swapping, and vice versa. If U1's new model performs well against U2's feedback then the modular user model is robust to changes in perspective under the assumption made.

In general, for each pair of users who share all but two stereotypes of the same type (role or team), the two stereotypes that are not shared are swapped (keeping the inter-component weights static) to create two new user models. The results of the robustness experiments are given in Table 3 and Table 4. Each pair of users who swap stereotypes over is represented as A→B where B's test feedback is used to evaluate the performance of A's new user model.

Table 3. Average precision values for a trained single-component (SC) user model and a trained modular user model (MOD) with stereotype swapped (military)

User	AP		User	AP		User	AP	
	SC	MOD		SC	MOD		SC	MOD
A2→A3	0.32	0.48	AL→A3	0.32	0.55	GC→GL	0.77	**0.83**
A3→A2	0.97	**0.89**	A4→AC	0.49	0.52	GL→GC	0.73	**0.73**
A2→A4	0.68	**0.75**	AC→A4	0.68	**0.78**	G2→G3	0.44	0.44
A4→A2	0.97	**0.98**	A4→G4	0.52	0.51	G3→G2	0.96	**0.83**
A2→AC	0.49	0.52	G4→A4	0.68	**0.78**	G2→G4	0.52	0.49
AC→A2	0.97	**0.98**	A4→AL	0.50	0.37	G4→G2	0.96	**0.89**
A2→G2	0.96	**0.91**	AL→A4	0.68	**0.82**	G2→GL	0.77	**0.84**
G2→A2	0.97	**0.97**	AC→LC	0.73	**0.71**	GL→G2	0.96	**0.89**
A2→AL	0.50	0.37	GC→AC	0.49	0.53	G3→G4	0.52	0.46
AL→A2	0.97	**0.92**	AC→AL	0.50	0.39	G4→G3	0.44	0.43
A3→A4	0.68	**0.78**	AL→AC	0.49	0.56	G3→GL	0.77	**0.82**
A4→A3	0.32	0.46	GC→G2	0.96	**0.89**	GL→G3	0.54	0.43
A3→AC	0.49	0.53	G2→GC	0.73	**0.71**	G4→GL	0.77	**0.83**
AC→A3	0.32	0.48	GC→G3	0.44	0.44	GL→G4	0.52	0.49
A3→G3	0.44	0.44	G3→GC	0.73	**0.75**	AL→GL	0.77	**0.83**
G3→A3	0.32	0.55	GC→G4	0.52	0.49	GL→AL	0.50	0.39
A3→AL	0.50	0.39	G4→GC	0.73	**0.73**	Average	0.64	0.65

Table 4. Average precision values for a trained single-component (SC) user model and a trained modular user model (MOD) withstereotype swapped (Intranet)

User pair	AP	
	SC	MOD
2→13	0.69	0.73
13→2	0.71	0.71
3→8	0.9	0.92
8→3	0.77	0.83
6→8	0.9	0.92
8→6	0.84	0.77
7→9	0.4	0.39
9→7	0.55	0.7
11→12	0.61	0.73
12→11	0.52	0.39
Average	0.69	0.71

In both domains, there is no significant difference between the performance of A's new modular model and B's single-component model. Broadly speaking, poor performance of the modular model is matched by poor performance of the single-component model, indicating that the user's information requirements are difficult to characterize by training from his feedback. For 8 out of 10 Intranet user pairs and 25 out of 50 military user pairs, the modular user model performed well ($AP > 0.7$). These results suggest that the proposed modular user modelling approach is indeed

robust to changes in user requirements as long as the user requirements can be characterized adequately by training. When a user's circumstances do change abruptly, the modular approach can be used to generate a new user model without the need for retraining, as long as his requirements can be characterized.

5.3 Rating Items for New Users Using Modular User Models

In section 3, we claimed that the proposed modular approach would help to alleviate the new user problem. The advantage of this would be that training would not be necessary for a new user before incoming information is prioritized according to his needs. In this section, we present results to empirically support this claim.

For a new user, it is assumed that no relevance feedback has been collected but that the stereotypes to which he belongs have been declared. A single-component user model cannot be constructed in the absence of training data. On-line experiments during new users arrive and provide feedback on items presented to them, the situation is simulated using the data already obtained. In each experiment, each user in turn is selected as the new user. The new user's relevance feedback is removed from the training set so the stereotypes to which he belongs are trained without it. Given no training feedback from the new user, the inter-component weights for his model cannot be learnt in the way described in section 3. Instead, three different weight assignment approaches are used:

1. All uniform - equal weights for all the new user's stereotypes (summing to 1.0)
2. Team uniform - All role stereotypes are allocated 0.0 weighting and the team stereotypes receive uniform weighting (summing to 1.0)
3. Role uniform - All team stereotypes are allocated 0.0 weighting and the role stereotypes receive uniform weighting (summing to 1.0)

The test performances of new user modular models constructed using the approach described above are given in Table 5 and Table 6.

For 6 out of 10 military users and 8 out of 13 Intranet users, at least one of the approaches provides good prioritization performance ($AP > 0.7$). In the military domain, there are no significant differences between the performances of the three approaches. Whereas, in the Intranet domain, the team uniform approach is significantly better than the other two approaches. The reason for this may be that there are more teams than roles in the Intranet domain and some users belong to more than one team. This means that teams are more likely than roles on average to be relevance indicators. With no feedback from the new user, it would be difficult to choose the best weighting approach automatically. Given the results above, the team uniform approach could be the default choice.

Table 5. Performance of modeling approaches for rating items for a new user

User	Inter-component weighting scheme		
	All uniform	Team uniform	Role uniform
A2	0.98	0.72	1.00
A3	0.30	0.53	0.23
A4	0.83	0.58	0.84

AC	0.47	0.73	0.44
GC	0.61	0.84	0.61
G2	0.89	0.70	0.96
G3	0.38	0.47	0.22
G4	0.31	0.16	0.40
AL	0.20	0.36	0.17
GL	0.64	0.72	0.47
Average	0.56	0.58	0.53

Table 6. Performance of modeling approaches for rating items for a new user

User	Inter-component weighting scheme		
	All uniform	Team uniform	Role uniform
1	0.75	0.75	0.61
2	0.65	0.62	0.64
3	0.86	0.84	0.98
4	0.70	0.77	0.65
5	0.36	0.43	0.43
6	0.80	0.89	0.52
7	0.59	0.71	0.32
8	0.94	0.98	0.80
9	0.29	0.44	0.19
10	0.74	0.66	0.80
11	0.29	0.44	0.03
12	0.42	0.65	0.13
13	0.61	0.83	0.44
Average	0.61	0.69	0.50

6. Conclusions and Outlook

We presented our investigation into flexible modular user modeling based on teams and roles. The experimental results support the claims made in section 3. Specifically, we have shown that the proposed modular approach has comparable accuracy to a single-component approach, but has the advantage of adjusting to changes in perspective more quickly. We have shown results that suggest that stereotype swapping could be used to bootstrap user models for new users without the need for model retraining. Our future work on flexible modular user models will investigate a number of issues:

- We have seen broadly similar results when a kNN approach was used to train the stereotypes. We plan to apply the Naïve Bayes algorithm, popular in text categorization, to determine whether our results generalize to other stereotype learning methods. We also plan to generate large artificial data sets for testing.
- Diversity and accuracy are widely cited as vital factors for achieving better performance with integrated multiple models than with individual models [12,15]. However, this applies to integrated models whose constituent models

are full classifiers of a target concept. We will assess the diversity and accuracy of the team and role stereotypes to determine any effect on model performance.

- We plan to determine whether new stereotypes can be successfully added to modular user models on-the-fly in order to adapt to new circumstances.

References

1. Salton, G. & McGill, M.: Introduction to Modern Information Retrieval. McGraw Hill. (1983).
2. Rich, E. User Modeling via Stereotypes, Cognitive Science, 3(4), 1979, 329-354. (1979)
3. Billsus, D. & Pazzani, M.: A Hybrid User Model for News Story Classification. UM'99 (1999).
4. McGowan, J., Kushmerick, N., & Smyth, B.: Who Do You Want to Be Today? Web Personae for Personalised Information Access. In Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (2002), May 29-31, Malaga, Spain.
5. Baudisch, P. & Brueckner, L.: TV Scout: Lowering the entry barrier to personalized TV program recommendation In Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (2002), May 29-31, Malaga, Spain..
6. Buczak, A., Zimmerman, J. & Kurapati, K.: Personalization: Improving Ease-of-Use, Trust and Accuracy of a TV Show Recommender. TV'02:Workshop on Personalization in TV. (2002).
7. Lock, Z & Kudenko, D.: Multi-Component User Models for Personalised Briefing Agents. Workshop on User and Group Models for Web-based Collaborative Environments. Held at the 9th International Conference on User Modeling UM'03. (2003).
8. Masthoff, J.: Selecting news to suit a group of criteria: An exploration. Proceedings of the Fourth Personalized TV workshop, associated with AH04. (2004).
9. Wolpert, D.: Stacked generalization. *Neural Networks*, 5(2):241–260, (1992).
10. Ferri, C., Flach, P & Hernández-Orallo, J.: Delegating classifiers. In Proceedings of the Twenty-First International Conference on Machine Learning ICML'04. ACM. (2004).
11. Yang, Y. & Pedersen, J. A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning. (1997).
12. Beitzel, S., Jensen, E., Chowdhury, A., Friedner, O., Grossman, D. & Goharian, N. Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies. In SAC 2003. (2003).
13. Mitchell, T.: Machine Learning. McGraw-Hill. (1997).
14. Witten, I. & Frank, E.: Data Mining. Morgan Kaufmann. (2000).
15. Kuncheva, L.: Diversity in multiple classifier systems. In Information Fusion, 6, 3-4. Elsevier. (2005).