A Study on the Granularity of User Modeling for Tag Prediction

E. Frías-Martinez M. Cebrián A. Jaimes
Data Mining and User Modeling Group, Telefonica Research
Emilio Vargas 6, 28043, Madrid, Spain
{efm, manuelc, ajaimes}@tid.es

Abstract

One of the characteristics of tag prediction mechanisms is that, typically, all user models are constructed with the same granularity. In this paper we hypothesize and empirically demonstrate that in order to increase tag prediction accuracy, the granularity of each user model has to be adapted to the level of usage of each particular user. We have constructed user models for tag prediction using association rules in Bibsonomy, a popular social bookmark and publication sharing system, at three granularity levels: (1) canonical, (2) stereotypical and (3) individual. Our experiments show that prediction accuracy improves if the level of granularity matches the level of participation of the user in the community (i.e., amount of tagging in Bibsonomy).

1. Introduction

A user model (UM) has been traditionally defined as a set of information structures designed to represent user preferences [1]. Inherently, the definition of user model does not imply a given granularity of the model (i.e., how much information is used to construct the model and how it is represented in the model). In terms of granularity, three main types of user models can be differentiated [2]: (1) a *canonical user model*, where the model is the same for all users; (2) *stereotypical*, which classifies users into clusters and creates a model for all users within each cluster, and (3) *individual*, in which a model is constructed per user. The *individual* UM has the highest granularity, while the *canonical* has the lowest.

Traditionally, adaptive (i.e. automatic) UMs are created using the same granularity for all users. This approach, extensively used in the literature [3], does not capture the fact that the more that a user has utilized a system, the better service the user expects [4]. A possible solution is to create UMs whose granularity is adapted to the level of usage of the system. In Bibsonomy[5], the social bookmark and publication sharing system used in this paper, the previous idea implies that the more the user has participated in tagging resources, the better the prediction rate should be.

In this paper we empirically demonstrate that a match between the level of usage and the granularity of the user model increases prediction rates (if there is a match the accuracy increases, and if there is a mismatch accuracy decreases).

Related Work. The problem of UM granularity has been raised in the area of information retrieval, and the general agreement is that different tasks require user models of different granularity [6]. However, tag prediction mechanisms similar to the one used in this paper [7][8], have not specifically studied the impact of matching usage and UM granularity levels on tag prediction accuracy. Studies in web interaction have focused on implementing specialized services according to different levels of usage [9], but in most cases the levels were explicitly stated by experts. In contrast, in our approach, the level of usage is given by the amount of information available for a particular user, i.e., the process of determining the level of usage is also adaptive.

2. Bibsonomy Tagging Behavior

Bibsonomy [5] is a social bookmarking system in which users describe the resources added to their shared personal library using tags. The data considered for this study is freely available at [10].

We use the TAS file from [10] which contains 816,197 entries. Each entry consists of a user ID, one tag, and a resource (bookmark or publication) tagged by that user. The file was manually filtered in order to include only non-spammers.

We define the *level of usage* as the total number of tags introduced by a particular user, and a *tagging session* as the set of tags that a particular user has used to describe a given resource.

The dataset of 816,197 entries contains 2,467 unique users who use 69,902 unique tags on 268,692 resources. As shown in Figure 1, the distribution of tags per user follows a power law distribution, with

$$\Pr(\text{tags per user} > k) \sim k^{-1.0458} \quad (1)$$

i.e., a small number of users tag very frequently while the majority of users participate in the tagging community with a reduced number of tags. Similar statistical behavior has been shown in other tagging communities [11][12].

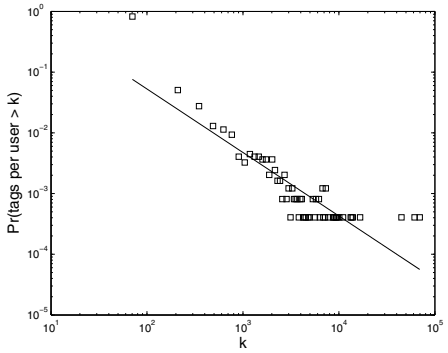

Figure 1. Distribution of tags per user in TAS.

3. UM Granularity for Tag Prediction

This section describes the implementation of three prediction mechanisms, one for each granularity level considered. All three prediction mechanisms have been tested using 3-fold cross validation (67%-33% for training and testing).

3.1. Tag Prediction Architecture

We use association rules (AR) to construct user models. Association Rules are obtained using the Apriori algorithm [13]. The prediction architecture is independent of the granularity of the UM and consists of an off-line and an on-line component. The off-line component generates, from a training set, a UM used for prediction. UMs are defined by the set of rules that predict a tag (consequent of the rule) given the set of the previous tags used (antecedent of the rule). Each AR has a confidence and a support value associated. The support (θ) of a rule is defined as the fraction of strings in the set of transactions of where the rule successfully applies and the confidence (σ) of a rule is defined as the fraction of times for which if the antecedent *X* is satisfied, the consequent *Y* is also true. In the terminology of a tagging community, the set of "items" is the set of tags used, and the set of "transactions" is the set of tagging sessions. ARs were obtained using the Association Rule Induction Tool [14].

The on-line component of the architecture, for each one of the tagging sessions of the testing set, considers each individual tag, and uses that tag to fire the association rules. From all the ARs fired, the consequents of the three rules with the highest confidence form the predicted set of tags. If one of those predicted tags is included in the rest of the tagging session the prediction is considered correct. Figure 2 presents this generic architecture.

The results of the evaluation of the prediction mechanisms are expressed in terms of CP (number of Correct Tags predicted), RF (total number of rules fired), and T (total number of tags of the testing set), where CP/RF expresses the percentage of the total number of correctly predicted tags with respect to the total number of rules fired, CP/T the percentage of the correct predictions per tag (similar to recall, but here we divide by the total number of tags in order to evaluate the impact of the prediction system in the final user), and RF/T the number of rules fired per tag, which indicates the computational load.

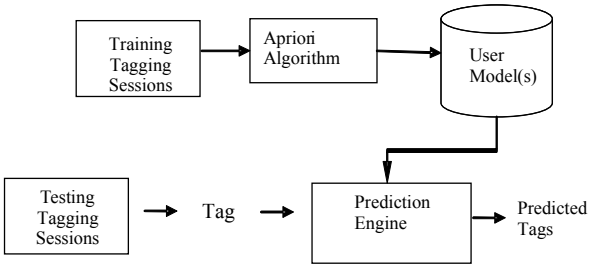

Figure 2. Prediction Engine Architecture

3.2. Canonical User Modeling

We generated four different canonical UMs for different values of support (θ=1%, 0.1%) and confidence (σ=50%, σ=80%). The results obtained from 3-fold cross-validation are presented in Table 1.

Canonical UMs can be used for any user of the tagging community but only capture knowledge that is common to the entire community (i.e., all of the users whose tags are included in the training set, with the rules defined by θ and σ). Our hypothesis is that canonical UMs perform well at predicting tags for users with a low level of participation in the tagging community (new users or users with a low level of usage), but higher granularity models (i.e.,

stereotypical, individual) perform better for users that have participated more in tagging activities.

Table 1. Evaluation of the canonical user model.

θ- σ	CP/RF	CP/T	RF/T
0.1-50	22%	2%	0.08
1-50	24%	2%	0.08
0.1-80	26%	3%	0.09
1-80	27%	3%	0.09

It is important to note that in the canonical user model few rules fire per tag—this is to be expected because, since the majority of users participate with a small number of tags, relatively "few" of the tags will lead to rules and those ARs constructed are not likely to fire.

3.3. Stereotypical User Modeling

This level of granularity constructs UMs from clusters of users. Ideally, such clusters should correspond to groups of users that share common tagging interests. First, we identify the tagging interests of every user by building a "tagging interest model". Next, we find clusters of users with similar interests, and finally, we build UMs for each of the clusters found.

We defined the tagging interests of each user as the set of tags that the user has assigned at least 20 times. This eliminates from consideration any users that have not used any tags 20 times or more, reducing the candidate users to 355 (out of the original 2,467 unique users), each one with a different number of tags in their user model.

The next step consists in finding clusters of users using the tagging interest model (i.e., clustering the 355 users based on the tags they have used more than 19 times). The first option is to represent the tagging interest model as a vector in which each dimension corresponds to one of the tags used more than 19 times. Given the large number of tags, however, this approach produces a sparse matrix which implies, from a clustering algorithm perspective, the curse of dimensionality problem. In order to avoid this problem, each user is represented by a vector containing the number of tags that the user has in common with each one of the remaining 354 users. Users were clustered using agglomerative hierarchical clustering using Euclidean distance. Based on the clustering obtained (Figure 3), we manually identified 4 clusters for the construction of stereotypical UMs.

It is interesting to note that the clusters correspond not only to similar interests but also coincide with the level of usage (see Table 2): while cluster 2 groups 314 users where the average number of tags introduced in the system amounts to 1,843, cluster 4 groups only 7 users but where each one is responsible, in average, for introducing more than 44,000 tags. The fact that there is a kernel of users that are responsible for the majority of the interaction was already shown in Figure 1.

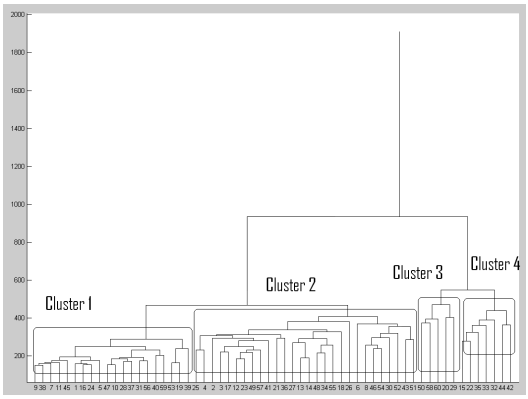

Figure 3. Clusters identified by hierarchical clustering.

Table 2. Cluster Characteristics.

Cluster	# of users	Average # of tags introduced
C1	26	7,680
C2	314	1,843
C3	5	12,823
C4	7	44,032

A stereotypical UM has been generated for each one of the 4 clusters. We have generated association rules for θ=1% and σ=80%. Table 3 presents the CP/RF, CP/T and RF/T rates for each one of the set of association rules generated for each cluster (C1, C2, C3, C4).

Table 3. Evaluation of the set of stereotypical UMs.

Cluster	CP/RF	CP/T	RF/T
C1	43%	12%	0.12
C2	30%	5%	0.15
C3	46%	20%	0.25
C4	47%	22%	0.25

Table 3 shows that if the level of usage matches the granularity of the user model there is an increase in the correct prediction rate. Clusters C1, C3 and C4 group users that, when compared to C2, have a higher level of usage which causes an increase in the CP/RF and CP/T ratios.

3.3. Individual User Modeling

Individual UMs can be constructed for any user of the tagging system. However, for new users or users with very little activity, there will usually not be enough information to construct a useful UM (the cold start problem). Individual UMs can be very powerful, however, because in tagging communities tags tend to

be very individual and unique for each user [11], especially for individuals with a high level of participation. Because C4 users have the highest participation in the community they are the best candidates to implement individual UMs. We selected, from C4, the first five users with the highest participation in the community, and for each one of these users an individual UM was constructed (θ=1% and σ=80%). Table 4 presents the results for each one of the users selected. All five users have CP/RF ratios higher than 52% and CP/T values of at least 19%.

Table 4. Evaluation of the set of individual UMs.

User	CP/RF	CP/T	RF/T
#244	53%	19%	0.32
#41	60%	22%	0.3
#45	57%	22%	0.29
#467	55%	25%	0.33
#337	61%	24%	0.3

4. Comparative Analysis

The results presented verify the hypothesis presented in this paper: matching granularity in UM with level of usage is a key factor in providing good predictions. For comparison purposes, Table 5 presents the results for the canonical, stereotypical and individual models, where the canonical model presents the results for θ-σ=1%-80%, and the stereotypical and individual models the average values for the cases considered in Table 3 and Table 4. By observing the CP/RF ratio, the results show that the higher the level of usage the higher the level of granularity should be.

Table 5. Comparison of the 3 levels of granularity.

	CP/RF	CP/T	RF/T
Canonical	27%	3%	0.09
Stereotype	41%	14%	0.19
Individual	57%	22%	0.3

For a new user, or for a user that has a low participation in the tagging community, the canonical UM can be used. For users that have an intermediate level of usage, a stereotypical UM should be used. In this case, if a canonical model is used the mismatch will imply a reduction in the correct prediction rate. Finally, for users with a high level of usage, an individual UM should be constructed.

5. Conclusions

The hypothesis that this paper tested was that the matching between the level of usage with the correct granularity of the UM improves the prediction rate, i.e., that the level of granularity should not be the same for all users but has to be adapted to the level of usage of each individual user.

The hypothesis was tested with a tagging prediction mechanism based on Association Rules. The results showed that in order to increase the prediction rate, the higher the level of usage the higher the level of granularity of the UM should be.

Future work includes investigating the following: (1) how to identify the scope of information used in the construction of the models (i.e., size and shape of clusters in the stereotypical case), and (2) how and when UMs evolve from one granularity to the next. Also it would be interesting to study the relation between level of usage and level of expertise in order to extend the conclusions of the study.

6. References

[1] A. Kobsa, "Generic User Modeling Systems", *User Modeling and User-Adapted Interaction* 11, 2001, pp. 49-63.

[2] D. Hartmut, U. Malinowski, T. Kühme, M. Schneider-Hufschmidt, "*State of the Art in Adaptive User Interfaces*", Siemens Corporate Research and Development.

[3] E. Frias-Martinez, S. Chen, and S. Liu, "Survey of Data Mining Approaches to User Modeling for Adaptive Hypermedia", *IEEE Tran. SMC-C.* 36(6), 2006, pp. 734-749.

[4] E. Frias-Martinez, S. Chen, S. Liu, R. Macreadie, "The Role of Human Factors in Stereotyping Behaviour and Perception of Digital Library Users: A Robust Clustering Approach", *UMUAI* 17(3), 2006, 305-337.

[5] R. Jäschke, A. Hotho, C. Schmitz and G. Stumme, "Analysis of the Publication Sharing Behaviour in BibSonomy", *Proc. Conceptual Structures: Knowledge Architectures for Smart Applications*, Springer, 2006.

[6] J. Lu, J. Callan, "User modeling for full-text federated search in peer-to-peer networks", SIGIR 2006, pp. 332–339.

[7] G. Mishne, "A Collaborative Approach to Automated tag Assignment for Weblog Posts", *WWW 2006*, pp. 953-954

[8] P. Heyman, D. Ramage, H. Garcia-Molina, "Social Tag Prediction", SIGIR 2008, pp. 531-538

[9] A. Lazander and H. Biemans, "Differences between novice and experienced users in searching information on the World Wide Web", *JASIST 51*(6), 2000, pp. 576-581.

[10] http://www.kde.cs.uni-kassel.de/ws/rsdc08/.

[11] R. Angelova, M. Lipczak, E. Milios and P. Pralat, "Characterizing a social bookmarking and tagging network", *18th European Conf. Artificial Intelligence*, 2008.

[12] B. Sigurbjörnsson and R. van Zwol, "Flicker Tag Recommendation based on Collective Knowledge", *WWW 2008*, pp. 327-336.

[13] R. Aggrawal, T. Imielinski, A. Swami, "Mining Association Rules between Set of items in large Databases", Proc. ACM SIGMOD, 1993, pp. 207-216.

[14] C. Borgelt, Association Rule Induction Tools, http://www.borgelt.net/software.html, Ver. 1.10, 2007.