Multi-Layered Ontology-Based User Profiles and Semantic Social Networks for Recommender Systems

Iván Cantador and Pablo Castells

Escuela Politécnica Superior, Universidad Autónoma de Madrid
Campus de Cantoblanco, 28049 Madrid, Spain
{ivan.cantador, pablo.castells}@uam.es

Abstract. This paper describes a strategy that automatically clusters ontology-based user profiles taking into account their common interests for domain concepts. The obtained semantic clusters are used to identify similarities among individuals at multiple semantic preference layers, and to define emergent, layered social networks that can be applied in collaborative and recommender systems. As an applicative development of our method, we have experimented with building a personalized information retrieval model that provides ranked item lists based on the existing concept clusters and multi-layered user networks.

1 Introduction

The rapid development, spread, and convergence of information and communication technologies are leading to new ways of inter-personal connection, communication, and collaboration. Virtual communities and computer-supported social networks [5,6] are starting to proliferate in increasingly sophisticated ways, opening new research opportunities on social group analysis, modeling, and exploitation. Finding hidden links between users based on the similarity of their preferences or historic behavior is not a new idea. In fact, this is the essence of the well-known collaborative recommender systems (e.g. see the survey given in [7]). However, in typical approaches, the comparison between users is done globally, in such a way that partial, but strong and useful similarities may be missed. For instance, two people may have a highly coincident taste in cinema, but a very divergent one in sports. The opinions of these people on movies could be highly valuable for each other, but risk to be ignored by many collaborative recommender systems, because global similarity between the users is low.

In this paper we propose a multi-layered approach to social networking. Like in previous approaches, our method builds and compares profiles of user interests for semantic topics and specific concepts, in order to find similarities among users. But in contrast to prior work, we divide the user profiles into clusters of cohesive interests, and based on this, several layers of social networks are found. This provides a richer model of interpersonal links, which better represents the way people find common interests in real life.

Our approach is based on an ontological representation of the domain of discourse where user interests are defined. The ontological space takes the shape of a semantic network of interrelated domain concepts. Taking advantage of the relations between concepts, and the (weighted) preferences of users for the concepts, our system clusters the semantic space based on the correlation of concepts appearing in the preferences of individual users. After this, user profiles are partitioned by projecting the concept clusters into the set of preferences of each user. Then, users can be compared on the basis of the resulting subsets of interests, in such a way that several, rather than just one, (weighted) links can be found between two users.

The rest of the paper is organized as follows. Section 2 describes the underlying semantics representation framework upon which our social network models are built. The proposed clustering techniques to build the multi-level relations between users are presented in Section 3. The exploitation of the derived networks to enhance collaborative filtering is described in Section 4. A small experiment where the techniques are tested is described in Section 5, and conclusions are given in Section 6.

2 Semantic User Preference Space

Our research builds upon an ontology-based personalization framework [1] that makes use of explicit user profiles. User preferences are represented as vectors $u_i = (u_{i,1}, u_{i,2}, ..., u_{i,N})$ where the weight $u_{i,j} \in [0,1]$ measures the intensity of the interest of user i for concept c_j in the domain ontology, and N is the total number of concepts in the ontology. Similarly, the objects d_k in the retrieval space are assumed to be described (annotated) by vectors $(d_{k,1}, d_{ik2}, ..., d_{k,N})$ of concept weights, in the same vector-space as user preferences. Based on this common logical representation, measures of user interest for content items can be computed by comparing preference and annotation vectors, and these measures can be used to prioritize, filter and rank contents in a personal way.

The ontology-based representation is richer and less ambiguous than a keyword-based or item-based model. It provides an adequate grounding for the representation of coarse to fine-grained user interests, and can be a key enabler to deal with the subtleties of user preferences. An ontology provides further formal, computer-processable meaning on the concepts (who is coaching a team, an actor’s filmography), and makes it available for the personalization system to take advantage of. Moreover, ontology standards support inference mechanisms that can be used to enhance personalization, so that, for instance, a user interested in animals (superclass of *cat*) is also recommended items about cats. Inversely, a user interested in *lizards* and *snakes* can be inferred to be interested in *reptiles*. Also, a user keen of *Dublin* can be assumed to like *Ireland*, through the *locatedIn* relation.

In real scenarios, user profiles tend to be very scattered, especially in those applications where user profiles have to be manually defined. Users are usually not willing to spend time describing their detailed preferences to the system, even less to assign weights to them, especially if they do not have a clear understanding of the effects and results of this input. On the other hand, applications where an automatic preference learning algorithm is applied tend to recognize the main characteristics of user preferences, thus yielding profiles that may entail a lack of expressivity. To overcome this problem, we propose a semantic preference spreading mechanism, which expands the initial set of preferences stored in user profiles through explicit semantic relations with other concepts in the ontology (see picture 1 in Figure 1). Our approach is based on the Constrained Spreading Activation (CSA) strategy [2,3]. The expansion is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed.

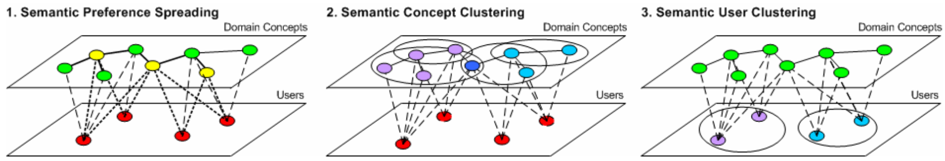

Fig. 1. Overall sequence of our proposed approach, comprising three steps: 1) semantic user preferences are spread, extending the initial sets of individual interests, 2) semantic domain concepts are clustered into concept groups, based on the vector space of user preferences, and 3) users are clustered in order to identify the closest class to each user

We have conducted several experiments showing that the performance of the personalization system is considerably poorer when the spreading mechanism is not enabled. Typically, the basic user profiles without expansion are too simple. They provide a good representative sample of user preferences, but do not reflect the real extent of user interests, which results in low overlaps between the preferences of different users. Therefore, the extension is not only important for the performance of individual personalization, but is essential for the clustering strategy described in the following sections, and it shows the advantages of a rich and precise ontology-based representation.

3 Semantic Multi-Layered Social Networks

In social communities, it is commonly accepted that people who are known to share a specific interest are likely to have additional connected interests [5]. In fact, this assumption is the basis of most recommender system technologies [7]. We assume this hypothesis here as well, in order to cluster the concept space in groups of preferences shared by several users. Specifically, a vector $c_j = (c_{j,1}, c_{j,2}, ..., c_{j,M})$ is assigned to each concept vector c_j present in the preferences of at least one user, where $c_{j,i} = u_{i,j}$ is the weight of concept $c_{j,}$ in the semantic profile of user i. Based on these vectors a classic hierarchical clustering strategy [4] is applied. The clusters thus obtained (picture 2 in Figure 1) represent the groups of preferences (topics of interests) in the concept-user vector space that are shared by a significant number of users. Once the concept clusters are created, each user is assigned to a specific cluster. The similarity between a user's preferences $u_i = (u_{i,1}, u_{i,2}, ..., u_{i,N})$ and a concept cluster C_r is computed by:

$$sim(u_i, C_r) = \frac{\sum_{c_j \in C_r} u_{i,j}}{|C_r|} \quad \textbf{(1)}$$

where c_j represents the concept that corresponds to the $u_{i,j}$ component of the user preference vector, and $|C_r|$ is the number of concepts included in the cluster. The clusters with highest similarities are then assigned to the users, thus creating groups of users with shared interests (picture 3 in Figure 1).

The concept and user clusters are then used to find emergent, focused semantic social networks. On the one hand, the preference weights of user profiles, the degrees of membership of the users to each cluster and the similarity measures between clusters are used to find relations between two distinct types of social items: individuals and groups of individuals. On the other hand, using the concept clusters user profiles are partitioned into semantic segments. Each of these segments corresponds to a concept cluster and represents a subset of the user interests that is shared by the users who contributed to the clustering process. By thus introducing further structure in user profiles, it is now possible to define relations among users at different levels, obtaining a multi-layered network of users. Figure 2 illustrates this idea. The top image represents a situation where two user clusters are obtained. Based on them (images below), user profiles are partitioned in two semantic layers. On each layer, weighted relations among users are derived, building up different social networks.

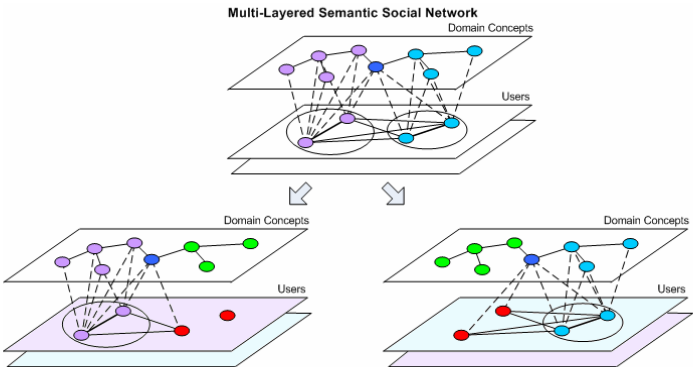

Fig. 2. Semantic multi-layered social network built from the clusters of concepts and users

The resulting networks have many potential applications. For one, they can be exploited to the benefit of collaborative filtering and recommendation, not only because they establish similarities between users, but also because they provide powerful means to focus on different contexts. The design of two information retrieval models in this direction is explored in next subsection.

4 Multi-Layered Models for Collaborative Filtering

As an applicative development of our multi-layered social network approach, we propose two recommender models that generate ranked lists of items in different scenarios. The first model (that we shall label as UP) is based on the semantic profile of the user to whom the ranked list is delivered. This model represents the situation where the interests of a user are compared to other interests in a social network. The second model (labeled NUP) outputs ranked lists disregarding the user profile. This can be applied in situations where a new user does not have a profile yet, or when the general preferences in a user's profile are too generic for a specific context, and do not help to guide the user towards a very particular, context-specific need. Additionally, we consider two versions for each model: a) one that generates a unique ranked list based on the similarities between the items and all the existing semantic clusters, and, b) one that provides a ranking for each semantic cluster. Thus, we consider four retrieval strategies, UP (profile-based), UP-r (profile-based, considering a specific cluster C_r), NUP (no profile), and NUP-r (no profile, considering a specific cluster C_r). The four strategies are formalized next. In the following, for a user profile u_i, an information object vector d_k, and a cluster C_r, we denote by u_i^r and d_k^r the projection of the corresponding concept vectors onto cluster C_r, i.e. the j-th component of u_i^r and d_k^r is $u_{i,j}$ and $d_{k,j}$ respectively, if $c_j \in C_r$, and 0 otherwise.

Model UP. The semantic profile of a user u_i is used by the system to return a unique ranked list. The preference score of an item d_k is computed as a weighted sum of the indirect preference values based on similarities with other users in each cluster, where the sum is weighted by the similarities with the clusters, as follows:

$$pref\left(d_k, u_i\right) = \sum_r nsim\left(d_k, C_r\right) \sum_l nsim_r\left(u_i, u_l\right) \cdot sim_r\left(d_k, u_l\right) \quad \textbf{(2)}$$

where:

$$sim\left(d_k, C_r\right) = \frac{\sum_{c_j \in C_r} d_{k,j}}{\left\| d_k \right\| \sqrt{\left| C_r \right|}}, \quad nsim\left(d_k, C_r\right) = \frac{sim\left(d_k, C_r\right)}{\sum_l sim\left(d_k, C_l\right)}$$

are the single and normalized similarities between the item d_k and the cluster C_r,

$$sim_r\left(u_i, u_l\right) = \cos\left(u_i^r, u_l^r\right) = \frac{u_i^r \cdot u_l^r}{\| u_i^r \| \cdot \| u_l^r \|}, \quad nsim_r\left(u_i, u_l\right) = \frac{sim_r\left(u_i, u_l\right)}{\sum_t sim_r\left(u_i, u_t\right)}$$

are the single and normalized similarities at layer r between user profiles U and U_m, and

$$sim_r\left(d_k, u_i\right) = \cos\left(d_k^r, u_i^r\right) = \frac{d_k^r \cdot u_i^r}{\left\| d_k^r \right\| \cdot \left\| u_i^r \right\|}$$

is the similarity at layer r between item d_k and user u_i.

Model UP-r. The preferences of the user are used by the system to return one ranked list per cluster, obtained from the similarities between users and items at each cluster layer. The ranking that corresponds to the cluster for which the user has the highest membership value is selected. The expression is analogous to equation (2), but does not include the term that connects the item with each cluster C_r.

$$pref_r\left(d_k, u_i\right) = \sum_l nsim_r\left(u_i, u_l\right) \cdot sim_r\left(d_k, u_l\right) \quad \textbf{(3)}$$

where r maximizes $sim(u_i,C_r)$.

Model NUP. The semantic profile of the user is ignored. The ranking of an item d_k is determined by its similarity with the clusters, and the similarity of the item and the profiles of the users within each cluster. Since the user does not have connections to other users, the influence of each profile is averaged by the number of users M.

$$pref\left(d_k, u_i\right) = \frac{1}{M} \sum_r nsim\left(d_k, C_r\right) \sum_l sim_r\left(d_k, u_l\right) \quad \textbf{(4)}$$

Model NUP-r. The preferences of the user are ignored, and one ranked list per cluster is delivered. As in the UP-r model, the ranking that corresponds to the cluster the user is most close to is selected. The expression is analogous to equation (4), but does not include the term that connects the item with each cluster C_r.

$$pref_r\left(d_k, u_i\right) = \frac{1}{M}\sum_l sim_r\left(d_k, u_l\right) \tag{5}$$

5 A simple experiment

To test the proposed strategies and models, a simple experiment has been set up, as follows. A set of 20 user profiles are considered. Each profile is manually defined considering 6 possible topics: *motor*, *construction*, *family*, *animals*, *beach* and *vegetation*. The degree of interest of the users for each topic is shown in Table 1, ranging over *high*, *medium*, and *low* interest, corresponding to preference weights close to 1, 0.5, and 0.

Table 1. Degrees of interest of users for each topic, and expected user clusters to be obtained

	Motor	***Construction***	***Family***	***Animals***	***Beach***	***Vegetation***	**Expected Cluster**
User1	High	High	Low	Low	Low	Low	1
User2	High	High	Low	Medium	Low	Low	1
User3	High	Medium	Low	Low	Medium	Low	1
User4	High	Medium	Low	Medium	Low	Low	1
User5	Medium	High	Medium	Low	Low	Low	1
User6	Medium	Medium	Low	Low	Low	Low	1
User7	Low	Low	High	High	Low	Medium	2
User8	Low	Medium	High	High	Low	Low	2
User9	Low	Low	High	Medium	Medium	Low	2
User10	Low	Low	High	Medium	Low	Medium	2
User11	Low	Low	Medium	High	Low	Low	2
User12	Low	Low	Medium	Medium	Low	Low	2
User13	Low	Low	Low	Low	High	High	3
User14	Medium	Low	Low	Low	High	High	3
User15	Low	Low	Medium	Low	High	Medium	3
User16	Low	Medium	Low	Low	High	Medium	3
User17	Low	Low	Low	Medium	Medium	High	3
User18	Low	Low	Low	Low	Medium	Medium	3
User19	Low	High	Low	Low	Medium	Low	1
User20	Low	Medium	High	Low	Low	Low	2

We have tested our method with this set of twenty user profiles, as explained next. First, new concepts are added to the profiles by the CSA strategy mentioned in Section 2, enhancing the concept and user clustering that follows. The applied clustering strategy is a hierarchical procedure based on the Euclidean distance to measure the similarities between concepts, and the average linkage method to measure the similarities between clusters. During the execution, a stop criterion should be defined based on an appropriate number of clusters, but since in our case the number of expected clusters is three, the stop criterion was not necessary. Table 2 summarizes the assignment of users to clusters cluster, showing their corresponding similarities values. It can be seen that the results are totally coincident with the expected values presented in Table 1.

Table 2. User clusters and associated similarity values between users and clusters. The maximum and minimum similarity values are shown in bold and italics respectively

Cluster	Users						
1	*User1*	*User2*	*User3*	*User4*	*User5*	*User6*	*User19*
	0.522	**0.562**	0.402	0.468	0.356	0.218	*0.194*
2	*User7*	*User8*	*User9*	*User10*	*User11*	*User12*	*User20*
	0.430	**0.389**	0.374	0.257	0.367	*0.169*	*0.212*
3	*User13*	*User14*	*User15*	*User16*	*User17*	*User18*	
	0.776	**0.714**	0.463	0.437	0.527	*0.217*	

Once the concept clusters have been automatically identified and each user has been assigned to a certain cluster, we apply the information retrieval models presented in the previous section. A set of 24 pictures was considered as the retrieval space. Each picture was annotated with (weighted) semantic metadata describing what the image depicts using a domain ontology. Observing the weighted annotations, an expert rated the relevance of the pictures for the twenty users of the example, assigning scores between 1 (totally irrelevant) and 5 (very relevant) to each picture, for each user.

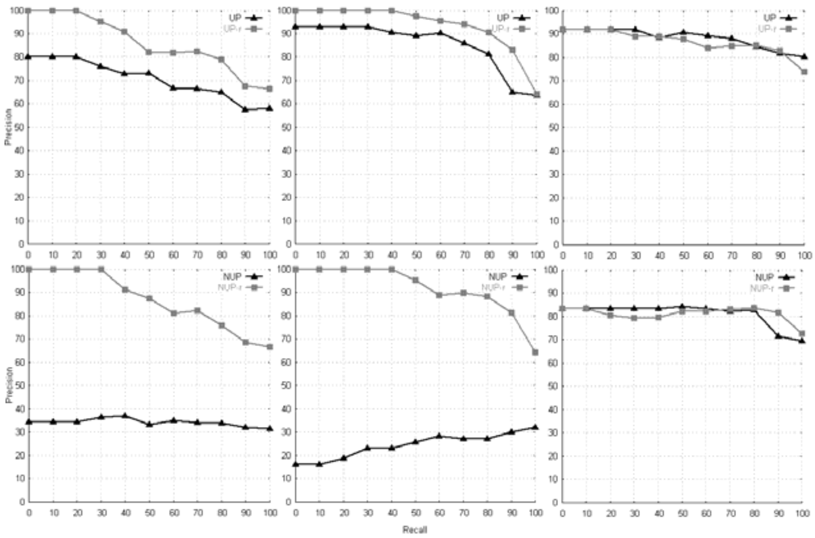

Fig. 3. Average precision vs. recall curves for users assigned to cluster 1 (left), cluster 2 (center) and cluster 3 (right). The graphics on top show the performance of the UP and UP-r models. The ones below correspond to the NUP and NUP-r models

Finally, the four different models are evaluated by computing their average precision/recall curves for the users of each of the three existing clusters. Figure 3 shows the results. Two conclusions can be inferred from the results: a) the version of the models that returns ranked lists according to specific clusters (UP-r and NUP-r) outperforms the one that generates a unique list, and, b) the models that make use of the relations among users in the social networks (UP and UP-r) result in significant improvements with respect to those that do not take into account similarities between user profiles.

6 Conclusions and future work

We have presented an approach to the automatic identification of social networks according to ontology-based user profiles. Taking into account the semantic preferences of several users, our approach clusters the ontology concept space, obtaining common topics of interest. With these topics, user preferences are partitioned into different layers. The degree of membership of the subprofiles to the clusters, and the similarities among them, are used to define social links that can be exploited by recommender systems.

Early experiments with a simple problem have been conducted, showing positive results. However, more realistic experiments need to be performed in order to properly evaluate our strategies. For instance, in real situations, user profiles usually have noisy components and do not easily allow to partition the concept space in a clear way. We are also aware of the need to test our approach in combination with automatic user preference learning techniques, to show that it is robust to imprecise user profiles, and test the impact of the accuracy of the ontology-based profiles on the correct performance of the clustering processes.

Acknowledgements

This research was supported by the European Commission (FP6-027685-MESH), and the Spanish Ministry of Science and Education (TIN2005-06885). The expressed content is the view of the authors but not necessarily the view of the MESH project as a whole.

References

1. Vallet, D., Mylonas, P., Corella, M. A., Fuentes, J. M., Castells, P., Avrithis, Y. *A Semantically-Enhanced Personalization Framework for Knowledge-Driven Media Services*. IADIS WWW/Internet Conference (ICWI 2005). Lisbon, Portugal, 2005.
2. Cohen, P. R., Kjeldsen, R.: *Information Retrieval by Constrained Spreading Activation in Semantic Networks*. Information Processing and Management 23(2), pp. 255-268, 1987.
3. Crestani, F., Lee, P. L.: *Searching the web by constrained spreading activation*. Information Processing & Management 36(4), pp. 585-605, 2000.
4. Duda, R.O., Hart, P., Stork, D.G.: *Pattern Classification*. John Wiley, 2001.
5. Liu, H., Maes, P., Davenport, G.: *Unraveling the Taste Fabric of Social Networks.* International Journal on Semantic Web and Information Systems 2(1), pp. 42-71, 2006.
6. Mika, P.: *Flink: Semantic Web technology for the extraction and analysis of social networks.* Web Semantics: Science, Services and Agents on the WWW 3(2-3), pp. 211-223, 2005.
7. Montaner, M., López, B., Lluís de la Rosa, J.: *Taxonomy of Recommender Agents on the Internet*. Artificial Intelligence Review 19, pp. 285-330, 2003.