Multilayered Semantic Social Network Modeling by Ontology-Based User Profiles Clustering: Application to Collaborative Filtering

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. We propose a multilayered semantic social network model that offers different views of common interests underlying a community of people. The applicability of the proposed model to a collaborative filtering system is empirically studied. Starting from a number of ontology-based user profiles and taking into account their common preferences, we automatically cluster the domain concept space. With the obtained semantic clusters, similarities among individuals are identified at multiple semantic preference layers, and emergent, layered social networks are defined, suitable to be used in collaborative environments and content recommenders.

1 Introduction

The swift development, spread, and convergence of information and communication technologies and support infrastructures in the last decade, which is reaching all aspects of businesses and homes in our everyday lives, is giving rise to new and unforeseen ways of inter-personal connection, communication, and collaboration. Virtual communities, computer-supported social networks [1,8,10,11], and collective interaction applications are indeed starting to proliferate in increasingly sophisticated ways, opening new research opportunities on social group analysis, modeling, and exploitation. In this paper we propose a novel approach towards building emerging social networks by analyzing the individual motivations and preferences of users, broken into potentially different areas of personal interest.

The issue of 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 [2,9,13], where items are recommended to a certain user concerning those of her interests shared with other users or according to opinions, comparatives and ratings of items given by similar users. However, in typical approaches, the comparison between users and items 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 might be low.

Here 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 and the user profiles are initially described as weighted lists measuring the user interests for those 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.

Multilayered social networks are potentially useful for many purposes. For instance, users may share preferences, items, knowledge, and benefit from each other's experience in focused or specialized conceptual areas, even if they have very different profiles as a whole. Such semantic subareas need not be defined manually, as they emerge automatically with our proposed method. Users may be recommended items or direct contacts with other users for different aspects of day-to-day life.

In recommendation environments there is an underlying need to distinguish different layers within the interests and preferences of the users. Depending on the current context, only a specific subset of the segments (layers) of a user profile should be considered in order to establish her similarities with other people when a recommendation has to be performed. We believe models of social networks partitioned at different common semantic layers could be very useful in the recommender processes offering more accurate and context-sensitive results. Thus, as an applicative development of our automatic semantic clustering and social network building methods, we present and empirically study in this paper several collaborative filtering models that retrieve information items according to a number of real user profiles and within different contexts.

In addition to these possibilities, our two-way space clustering, which finds clusters of users based on the clusters of concepts found in a first pass, offers a reinforced partition of the user space that could be exploited to build group profiles for sets of related users. These group profiles might enable an efficient strategy for collaborative recommendation in real-time, by using the merged profiles as representatives of classes of users.

The rest of the paper has the following structure. Section 2 describes the 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. Section 5 describes a simple example where the techniques are tested. An early experiment with real subjects and user profiles is presented in Section 6, and conclusions are given in Section 7.

2 Ontology-based User Profiles and Preference Spreading

In contrast to other approaches in personalized content retrieval, our approach makes use of explicit user profiles (as opposed to e.g. sets of preferred documents). Working within an ontology-based personalization framework [16], 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 (a class or an instance) in the domain ontology, N being 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 ($d_{k,1}$, $d_{i\mathrm{k}2}$, ..., $d_{k,N}$) 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 (a collection, a catalog, a search result) 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 (e.g. interest for items such as a sports team, an actor, a stock value), 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, financial data on a stock), and makes it available for the personalization system to take advantage of. Furthermore, ontology standards, such as RDF and OWL, 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 *Czech Republic* can be assumed to like *Prague*, through the *locatedIn* transitive relation. These characteristics will be exploited in our personalized retrieval model.

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 [4,5]. The expansion is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed.

Thus, the system outputs ranked lists of content items taking into account not only the preferences of the current user, but also a semantic spreading mechanism through the user profile and the domain ontology. In fact, previous experiments were done without the semantic spreading process and very poor results were obtained. The profiles were very simple and the matching between the preferences of different users was low. This observation shows a better performance when using ontology-based profiles, instead of classical keyword-based preferences representations.

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.

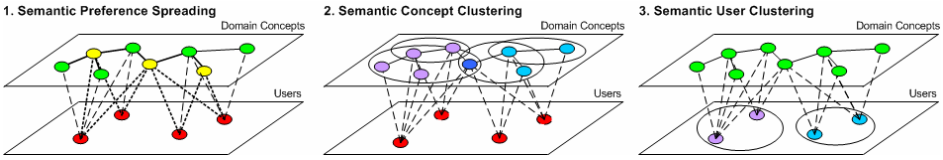

Fig. 1. Overall sequence of our proposed approach, comprising three steps: 1) semantic user preferences are spread, extending the initial sets of individual interests, 2) semantic domain concepts are clustered into concept groups, based on the vector space of user preferences, and 3) users are clustered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

3 Multilayered Semantic 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 [8]. For instance, people who share interests in traveling might be also keen on topics related in photography, gastronomy or languages. In fact, this assumption is the basis of most recommender system technologies [3,7,12,14]. We assume this hypothesis here as well, in order to cluster the concept space in groups of preferences shared by several users.

We propose here to exploit the links between users and concepts to extract relations among users and derive semantic social networks according to common interests. Analyzing the structure of the domain ontology and taking into account the semantic preference weights of the user profiles we shall cluster the domain concept space generating groups of interests shared by certain users. Thus, those users who share interests of a specific concept cluster will be connected in the network, and their preference weights will measure the degree of membership to each cluster. 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 [6,15] is applied. The clusters obtained (picture 2 in Figure 1) represent the groups of preferences (topics of interests) in the concept-user vector space 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 cluster C_r is computed by:

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

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. 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 multilayered 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.

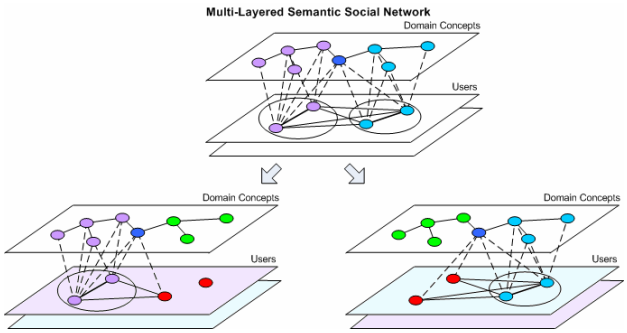

Fig. 2. Multilayered semantic social network built from the clusters of concepts and users

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 semantic contexts for different information needs. The design of two information retrieval models in this direction is explored in next section.

4 Multilayered Models for Collaborative Filtering

Collaborative filtering applications adapt to groups of people who interact with the system, in a way that single users benefit from the experience of other users with which they have certain traits or interests in common. User groups may be quite heterogeneous, and it might be very difficult to define the mechanisms for which the system adapts itself to the groups of users, in such a way that each individual enjoys or even benefits from the results. Furthermore, once the user association rules are defined, an efficient search

for neighbors among a large user population of potential neighbors has to be addressed. This is the great bottleneck in conventional user-based collaborative filtering algorithms [12]. Item-based algorithms [3,7,14] attempt to avoid these difficulties by exploring the relations among items, rather than the relations among users. However, the item neighborhood is fairly static and do not allow to easily apply personalized recommendations or inference mechanisms to discover potential hidden user interests.

We believe that exploiting the relations of the underlying social network which emerges from the users' interests, and combining them with semantic item preference information can have an important benefit in collaborative filtering and recommendation. Using our semantic multilayered social network proposal explained in previous sections, we present here two recommender models that generate ranked lists of items in different scenarios taking into account the links between users in the generated social networks. 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:

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

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

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

are the single and normalized similarities at layer r between user profiles u_i and u_l, 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.

The idea behind this first model is to compare the current user interests with those of the others users, and, taking into account the similarities among them, weight all their complacencies about the different items. The comparisons are done for each concept cluster measuring the similarities between the items and the clusters. We thus attempt to recommend an item in a double way. First, according to the item characteristics, and second, according to the connections among user interests, in both cases at different semantic layers.

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.

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

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

Analogously to the previous model, this one makes use of the relations among the user interests, and the user satisfactions with the items. The difference here is that recommendations are done separately for each layer. If the current semantic cluster is well identified for a certain item, we expect to achieve better precision/recall results than those obtained with the overall model.

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.

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

Designed for situations in which the current user profile has not yet been defined, this model uniformly gathers all the user complacencies about the items at different semantic layers. Although it would provide worse precision/recall results than the models UP and UP-r, this one might be fairly suitable as a first approach to recommendations previous to manual or automatic user profile constructions.

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.

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

This last model is the most simple of all the proposals. It only measures the users' complacencies with the items at the layers that best fit them, representing thus a kind of item-based collaborative filtering system.

5 An example

For testing the proposed strategies and models a simple experiment has been set up. A set of 20 user profiles are considered. Each profile is manually defined considering 6 possible topics: *animals*, *beach, construction*, *family*, *motor* 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

As it can be seen from the table, the six first users (1 to 6) have *medium* or *high* degrees of interests in *motor* and *construction*. For them it is expected to obtain a common cluster, named cluster 1 in the table. The next six users (7 to 12) share again two topics in their preferences. They like concepts associated with *family* and *animals*. For them a new cluster is expected, named cluster 2. The same situation happens with the next six users (13 to 18); their common topics are *beach* and *vegetation*, an expected cluster named cluster 3. Finally, the last two users have noisy profiles, in the sense that they do not have preferences easily assigned to one of the previous clusters. However, it is comprehensible that User19 should be assigned to cluster 1 because of her high interests in *construction* and User20 should be assigned to cluster 2 due to her high interests in *family*.

Table 2 shows the correspondence of concepts to topics. Note that user profiles do not necessarily include all the concepts of a topic. As mentioned before, in real world

applications it is unrealistic to assume profiles are complete, since they typically include only a subset of all the actual user preferences.

Table 2. Initial concepts for each of the six considered topics

Topic	Concepts
Motor	Vehicle, Motorcycle, Bicycle, Helicopter, Boat
Construction	Construction, Fortress, Road, Street
Family	Family, Wife, Husband, Daughter , Son, Mother, Father, Sister, Brother
Animals	Animal, Dog, Cat, Bird, Dove, Eagle, Fish, Horse, Rabbit, Reptile, Snake, Turtle
Beach	Water , Sand, Sky
Vegetation	Vegetation, Tree (instance of Vegetation), Plant (instance of Vegetation), Flower (instance of Vegetation)

We have tested our method with this set of 20 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, *N–1* (with *N* the total number of concepts) clustering levels were obtained, and a stop criterion to choose an appropriate number of clusters would be needed. In our case the number of expected clusters is three so the stop criterion was not necessary. Table 3 summarizes the assignment of users to clusters, showing their corresponding similarities values. It can be shown that the obtained results completely coincide with the expected values presented in Table 1. All the users are assigned to their corresponding clusters. Furthermore, the users' similarities values reflect their degrees of belonging to each cluster.

Table 3. 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 20 users of the example, assigning scores between 1 (totally irrelevant) and 5 (very relevant) to each picture, for each user. We show in Table 4 the final concepts obtained and grouped in the semantic Constrained Spreading Activation and concept clustering phases. Although most of the final concepts do not appear in the initial user profiles, they are very important in further steps because they

help in the construction of the clusters. Our plans for future work include studying in depth the influence of the CSA in realistic empirical experiments.

Table 4. Concepts assigned to the obtained user clusters classified by semantic topic

Cluster	Concepts
1	**MOTOR**: Vehicle, Racing-Car, Tractor, Ambulance, Motorcycle, Bicycle, Helicopter, Boat, Sailing-Boat, Water-Motor, Canoe, Surf, Windsurf, Lift, Chair-Lift, Toboggan, Cable-Car, Sleigh, Snow-Cat **CONSTRUCTION**: Construction, Fortress, Garage, Road, Speedway, Racing-Circuit, Short-Oval, Street, Wind-Tunnel, Pier, Lighthouse, Beach-Hut, Mountain-Hut, Mountain-Shelter, Mountain-Villa
2	**FAMILY**: Family, Wife, Husband, Daughter , Son, Mother-In-Law, Father-In-Law, Nephew, Parent, 'Fred' (instance of Parent), Grandmother, Grandfather, Mother, Father, Sister, 'Christina' (instance of Sister), Brother, 'Peter' (instance of Brother), Cousin , Widow **ANIMALS**: Animal, Vertebrates, Invertebrates, Terrestrial, Mammals, Dog, 'Tobby' (instance of Dog), Cat, Bird, Parrot, Pigeon, Dove, Parrot, Eagle, Butterfly, Fish, Horse, Rabbit, Reptile, Snake, Turtle, Tortoise, Crab
3	**BEACH**: Water, Sand, Sky **VEGETATION**: Vegetation, 'Tree' (instance of Vegetation), 'Plant' (instance of Vegetation), 'Flower' (instance of Vegetation)

The four different models are finally 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.

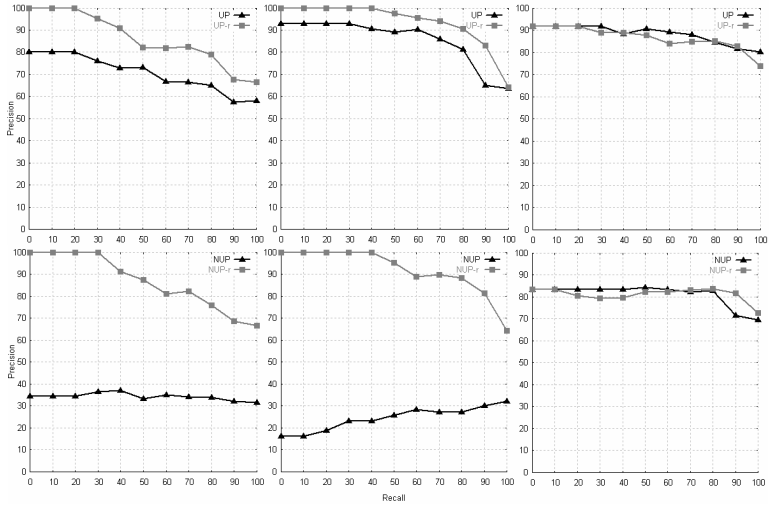

Fig. 3. Average precision vs. recall curves for users assigned to cluster 1 (left), cluster 2 (center) and cluster 3 (right). The graphics on top show the performance of the UP and UP-r models. The ones below correspond to the NUP and NUP-r models

6 Early Experiments

We have performed an experiment with real subjects in order to evaluate the effectiveness of our proposed recommendation models. Following the ideas exposed in the simple example of the previous section, the experiment was setup as follows.

The set of 24 pictures used in the example was again considered as the retrieval space. As mentioned before, each picture was annotated with semantic metadata describing what the image depicts, using a domain ontology including six certain topics: *animals*, *beach*, *construction*, *family*, *motor* and *vegetation*. A weight in [0,1] was assigned to each annotation, reflecting the relative importance of the concept in the picture. 20 graduate students of our department participated in the experiment. They were asked to independently define their weighted preferences about a list of concepts related to the above topics and existing in the pictures semantic annotations. No restriction was imposed on the number of topics and concepts to be selected by each of the students. Indeed, the generated user profiles showed very different characteristics, observable not only in their joint interests, but also in their complexity. Some students defined their profiles very thoroughly, while others only annotated a few concepts of interest. This fact was obviously very appropriate for the experiment done. In a real scenario where an automatic preference learning algorithm will have to be used, the obtained user profiles would include noisy and incomplete components that will hinder the clustering and recommendation mechanisms.

Once the 20 user profiles were created, we run our method. After the execution of the semantic preference spreading procedure, the domain concept space was clustered according to similar user interests. In this phase, because our strategy is based on a hierarchical clustering method, various clustering levels (representable by the corresponding dendrogram) were found, expressing different compromises between complexity, described in terms of number of concept clusters, and compactness, defined by the number of concepts per cluster or the minimum distance between clusters. In Figure 4 we graph the minimum inter-cluster distance against the number of concept clusters.

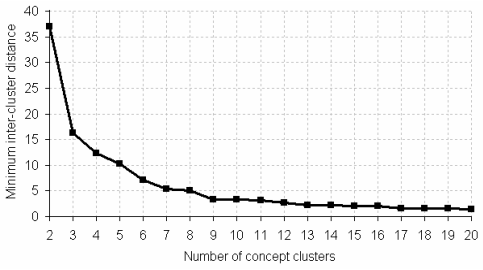

Fig. 4. Minimum inter-cluster distance at different concept clustering levels

A stop criterion has then to be applied in order to determine what number of clusters should be chosen. In this case, we shall use a rule based on the *elbow criterion*, which says you should be choose a number of clusters so that adding another cluster does not add sufficient information. We are interested in a clustering level with a relative small number of clusters and which does not vary excessively the inter-cluster distance with respect to previous levels. Therefore, attending to the figure, we will

focus on clustering levels with R = 4, 5, 6 clusters, corresponding to the angle (elbow) in the graph. Table 5 shows the users that most contributed to the definition of the different concept cluster, and their corresponding similarities values.

Table 5. User clusters and associated similarity values between users and clusters obtained at concept clustering levels R = 4, 5, 6

R	**Cluster**	**Users**								
4	*1*	*User01* 0.388	*User02* 0.370	*User05* 0.457	*User06* 0.689	*User19* 0.393				
	2									
	3	*User03* 0.521	*User04* 0.646	*User07* 0.618	*User09* 0.209	*User12* 0.536	*User15* 0.697	*User16* 0.730	*User18* 0.461	
	4	*User08* 0.900	*User10* 0.089	*User11* 0.810	*User13* 0.591	*User14* 0.833	*User17* 0.630	*User20* 0.777		
5	*1*	*User03* 0.818	*User07* 0.635							
	2									
	3	*User04* 0.646	*User09* 0.209	*User12* 0.536	*User16* 0.730	*User18* 0.461				
	4	*User01* 0.395	*User02* 0.554	*User05* 0.554	*User06* 0.720	*User15* 0.712	*User19* 0.399			
	5	*User08* 0.900	*User10* 0.089	*User11* 0.810	*User13* 0.591	*User14* 0.833	*User17* 0.630	*User20* 0.777		
6	*1*	*User6* 0.818								
	2									
	3	*User18* 0.481								
	4	*User02* 0.554	*User05* 0.554	*User06* 0.720	*User19* 0.399					
	5	*User08* 0.900	*User13* 0.591	*User11* 0.810	*User17* 0.630	*User20* 0.777				
	6	*User01* 0.786	*User04* 0.800	*User07* 0.771	*User09* 0.600	*User10* 0.214	*User12* 0.671	*User14* 0.857	*User15* 0.829	*User16* 0.814

It has to be noted that not all the concept clusters have assigned user profiles. However, there are semantic relations between users within a certain concept cluster, independently of being associated to other clusters or the number of users assigned to the cluster. For instance, at clustering level R = 4, we obtained the weighted semantic relations plotted in Figure 5. Representing the semantic social networks of the users, the diagrams of the figure describe the similarity terms $sim_r(u_i, u_l)$, $i,l \in \{1,20\}$ (see equations 2 and 3). The color of each cell depicts the similarity values between two given users: the dark and light gray cells indicate respectively similarity values greater and lower than 0.5, while the white ones mean no existent relation. Note that a relation between two certain users with a high weight does not necessary implicate a high interest of both for the concepts on the current cluster. What it means is that they interests agree at this layer. They could really like it or they might hate its topics.

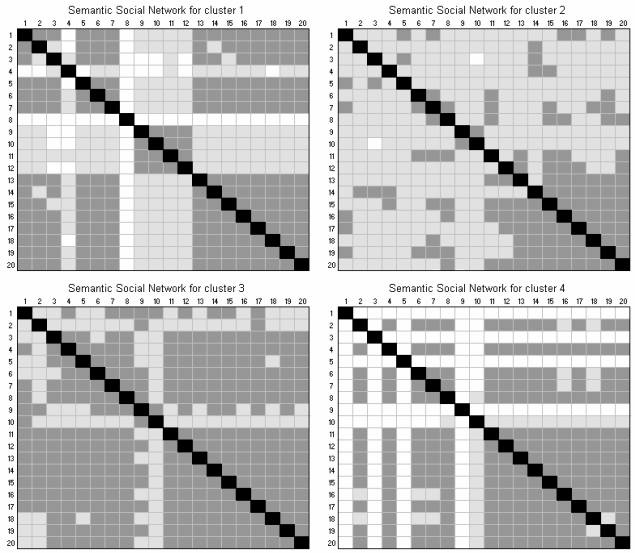

Fig. 5. Symmetric user similarity matrices at layers 1, 2, 3 and 4 between user profiles u_i and u_l ($i, l \in \{1, 20\}$) obtained at clustering level R=4. Dark and light gray cells represent respectively similarity values greater and lower than 0.5. White cells mean no relation between users

Table 6 shows the concept clusters obtained at clustering level $R = 4$. We have underlined those general concepts that initially did not appear in the profiles and were in the upper levels of the domain ontology. Inferred from our preference spreading strategy, these concepts do not necessary define the specific semantics of the clusters, but help to build the latter during the clustering processes.

Table 6. Concept clusters obtained at clustering level R=4

Cluster	Concepts
1	**ANIMALS**: Rabbit **CONSTRUCTION**: Construction, Speedway, Racing-Circuit, Short-Oval, Garage, Lighthouse, Pier, Beach-Hut, Mountain-Shelter, Mountain-Villa, Mountain-Hut, **MOTOR**: Vehicle, Ambulance, Racing-Car, Tractor, Canoe, Surf, Windsurf, Water-Motor, Sleigh, Snow-Cat, Lift, Chair-Lift, Toboggan, Cable-Car
2	**ANIMALS**: Organism, Agentive-Physical-Object, Reptile, Snake, Tortoise, Sheep, Dove, Fish, Mountain-Goat, Reindeer **CONSTRUCTION**: Non-Agentive-Physical-Object, Geological-Object, Ground, Artifact, Fortress, Road, Street **FAMILY**: Civil-Status, Wife, Husband **MOTOR**: Conveyance, Bicycle, Motorcycle, Helicopter, Boat, Sailing-Boat
3	**ANIMALS**: Animal, Vertebrates, Invertebrates, Terrestrial, Mammals, Dog, 'Tobby' (instance of Dog), Cat, Horse, Bird, Eagle, Parrot, Pigeon, Butterfly, Crab **BEACH**: Water, Sand, Sky **VEGETATION**: Vegetation, 'Tree' (instance of Vegetation), 'Plant' (instance of Vegetation), 'Flower' (instance of Vegetation)
4	**FAMILY**: Family, Grandmother, Grandfather, Parent, Mother, Father, Sister, Brother, Daughter, Son, Mother-In-Law, Father-In-Law, Cousin, Nephew, Widow, 'Fred' (instance of Parent), 'Christina' (instance of Sister), 'Peter' (instance of Brother)

Some conclusions can be drawn from this experiment. Cluster 1 contains the majority of the most specific concepts related to *construction* and *motor*, showing a significative correlation between these two topics of interest. Checking the profiles of the users associated to the cluster, we observed they overall have medium-high weights on the concepts of these topics. Cluster 2 is the one with more different topics and general concepts. In fact, it is the cluster that does not have assigned users in Table 6 and does have the most weakness relations between users in Figure 5. It is also notorious that the concepts 'wife' and 'husband' appear in this cluster. This is due to these concepts were not be annotated in the profiles by the subjects, who were students, not married at the moment. Cluster 3 is the one that gathers all the concepts about *beach* and *vegetation*. The subjects who liked vegetation items also seemed to be interested in beach items. It also has many of the concepts belonging to the topic of *animals*, but in contrast to cluster 2, the annotations were for more common and domestic animals. Finally, cluster 4 collects the majority of the *family* concepts. It can be observed from the user profiles that a number of subjects only defined their preferences in this topic.

Finally, as we did in the example of section 5, we evaluate the proposed retrieval models computing their average precision/recall curves for the users of each of the existing clusters. In this case we calculate the curves at different clustering levels (R = 4, 5, 6), and we only consider the models UP and UP-r because they make use of the relations among users in the social networks, and offer significant improvements with respect to those that do not take into consideration similarities between user profiles. Figure 6 exposes the results.

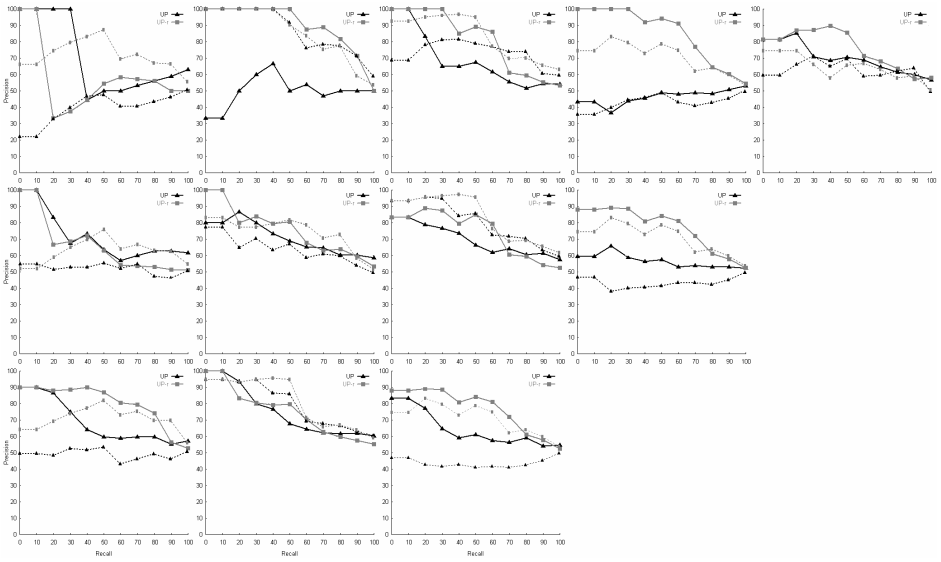

Fig. 6. Average precision vs. recall curves for users assigned to the user clusters obtained with the UP (black lines) and UP-r (gray lines) models at levels R=6 (graphics on the top), R=5 (graphics in the middle), and R=4 (graphics on the bottom) concept clusters. For both models, the dotted lines represent the results achieved without semantic preference spreading

Again, the version UP-r, which returns ranked lists according to specific clusters, outperforms the version UP, which generates a unique list assembling the contributions of the users in all the clusters. Obviously, the more clusters we have, the better performance is achieved. The clusters tend to have assigned fewer users and seem more similar to the individual profiles. However, it can be seen that very good results are obtained with only three clusters. Additionally, for both models, we have plotted with dotted lines the curves achieved without spreading the user semantic preferences. Although more statistically significative experiments have to be done in order to make founded conclusions, it can be pointed out that our clustering strategy performs better when it is combined with the CSA algorithm, especially in the UP-r model. This fact let give us preliminary evidences of the importance of spreading the user profiles before the clustering processes.

7 Conclusions and further work

In this 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 we cluster the ontology concept space, obtaining common topics of interest. With these topics, preferences are partitioned into different layers.

The degree of membership of the obtained subprofiles to the clusters, and the similarities among them, are used to define social links that can be exploited by collaborative filtering systems. Early experiments with real subjects have been done applying the emergent social networks to a variety of collaborative filtering models showing the feasibility of our clustering strategy. However, more sophisticated and statistically significative experiments need to be performed in order to properly evaluate the models. We have planned to implement a web-based recommender agent that will allow users to easily define their profiles, see their semantic relations with other people, and evaluate the existing items and recommendations given by the system. Thus, we expect to enlarge the repositories of items and user profiles, and improve our empirical studies.

Our implementation of the applied clustering strategy was 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. Of course, several aspects of the clustering algorithm have to be investigated in future work using noisy user profiles: 1) the type of clustering (hierarchical or partitional), 2) the distance measure between two concepts (Manhattan, Euclidean or Squared Euclidean distances), 3) the distance measure between two clusters (single, complete or average linkage), 4) the stop criterion that determines what number of clusters should be chosen, and, 5) the similarity measure between given clusters and user profiles; we have used a measure considering the relative size of the clusters, but we have not taken into account what proportion of the user preferences is being satisfied by the different concept clusters.

We are also aware of the need to test our approach in combination with automatic user preference learning techniques in order to investigate its robustness to imprecise user interests, and the impact of the accuracy of the ontology-based profiles on the correct performance of the clustering processes. An adequate acquisition of the concepts of interest and their further classification and annotation in the ontology-based profiles will be crucial to the correct performance of the clustering processes.

Acknowledgements

The research leading to this document has received funding from the European Community's Sixth Framework Programme (FP6-027685 – MESH), and the Spanish Ministry of Science and Education (TIN2005-06885). However, it reflects only the authors' views, and the European Community is not liable for any use that may be made of the information contained therein.

References

1. Alani, H., O'Hara, K., Shadbolt, N.: *ONTOCOPI: Methods and Tools for Identifying Communities of Practice*. Intelligent Information Processing 2002, pp. 225-236, 2002.
2. Ardissono, L., Goy, A., Petrone, G., Segnan, M., Torasso, P.: *INTRIGUE: personalized recommendation of tourist attractions for desktop and handset devices*. Applied Artificial Intelligence, Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries 17(8-9), pp. 687-714. Taylor and Francis, 2003.
3. Balabanovic, M., Shoham, Y.: *Content-Based Collaborative Recommendation*. Communications ACM, pp. 66-72, 1997.
4. Cohen, P. R. and Kjeldsen, R.: *Information Retrieval by Constrained Spreading Activation in Semantic Networks*. Information Processing and Management 23(2), pp. 255-268, 1987.
5. Crestani, F., Lee, P. L.: *Searching the web by constrained spreading activation*. Information Processing & Management 36(4), pp. 585-605, 2000.
6. Duda, R.O., Hart, P., Stork, D.G.: *Pattern Classification*. John Wiley. 2001.
7. Linden, G., Smith, B., York, J.: *Amazon.com Recommendations: Item-to-Item Collaborative Filtering*. IEEE Internet Computing, 7(1):76-80, 2003.
8. 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.
9. McCarthy, J., Anagnost, T.: *MusicFX: An arbiter of group preferences for computer supported collaborative workouts*. ACM International Conference on Computer Supported Cooperative Work (CSCW 1998). Seattle, Washington, pp. 363-372, 1998.
10. Mika, P.: *Ontologies Are Us: A Unified Model of Social Networks and Semantics*. Proceedings of the 4th International Semantic Web Conference (ISWC 2005), pp. 522-536, 2005.
11. 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.
12. Montaner, M., López, B., Lluís de la Rosa, J.: *Taxonomy of Recommender Agents on the Internet*. Artificial Intelligence Review 19, pp. 285-330, 2003.
13. O'Conner, M., Cosley, D., Konstan, J. A., Riedl, J.: *PolyLens: A recommender system for groups of users*. 7th European Conference on Computer Supported Cooperative Work (ECSCW 2001). Bonn, Germany, pp. 199-218, 2001.
14. Sarwar, B.M., et al.: *Item-Based Collaborative Filtering Recommendation Algorithms*. 10th International World Wide Web Conference, ACM Press, pp. 285-295, 2001.
15. Ungar, L., Foster, D.: *Clustering Methods for Collaborative Filtering*. Proceedings of the Workshop on Recommendation Systems at the 15th National Conference on Artificial Intelligence, AAAI Press, 1998.
16. 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.