Building Emergent Social Networks and Group Profiles by Semantic User Preference Clustering

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 presents a novel approach to automatic semantic social network construction based on semantic user preference clustering. Considering a number of users, each of them with an associated ontology-based profile, we propose a strategy that clusters the concepts of the reference ontology according to user preferences of these concepts, and then determines which clusters are more appropriate to the users. The resultant user clusters can be merged into individual group profiles, automatically defining a semantic social network suitable for use in collaborative and recommendation environments.

1 Introduction

The swift development, spread, and convergence of information and communication technologies and support infrastructures, 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, and collective interaction are indeed starting to proliferate and grow in increasingly sophisticated ways, opening new opportunities for research 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.

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 [2,10,15]). 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, or totally different professional interests. 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 the global similarity between the users is low.

In this paper we propose a multi-layered approach to dynamic social networking. Like in previous approaches [1,13,14], 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, in our approach user profiles are divided into clusters of cohesive interests, and based on this, several layers of networks are found. This provides a richer, finer-grained model of interpersonal links, which better represents the way people find common interests in real life, which typically takes place on different, partial planes of each other's 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.

Multi-layered 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 addition to these possibilities, we have experimented with the proposed two-way space clustering mechanism. Finding clusters of users based on those clusters of concepts that represent common topics of interest, we obtain a reinforced partition of the user space that can be exploited to build group profiles for sets of related users. These group profiles 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 organization. Section 2 describes our ontology-based user profile representation and gives an overview of the personalized content retrieval system in which it is being used. Section 3 explains our proposal to automatic construction of multi-layered social networks based on semantic user preference clustering. In section 4 several strategies for modeling group profiles are experimentally investigated. Finally, some conclusions and future research lines are given in section 5.

2 User Profile Representation

Our research builds upon an ontology-based personalization framework. In this section we provide an overview of the basic principles of this framework, with a special focus on user profile representation, and the exploitation of the profiles for personalized content retrieval. Further details can be found in [16].

In contrast with 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, in which the domain of interest is described through semantic concepts corresponding to the different classes and instances of a domain ontology, user preferences are represented as vectors

$U_m = (w_{m1},...,w_{mN})$, where $m = 1,...,M$, M is the number of existent user profiles, and $w_{mn} \in [0,1]$ is the weight that measure the intensity of the interest of user *m* for concept c_n 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 $\vec{c}_n = (c_{n1},...,c_{nM})$ of concept weights, in the same vector-space as user preferences. Comparing the metadata of content items, and the preferred concepts in a user profile, the system finds how the user may like each element. Based on her preference weights, measures of user interest for content units can be computed, with which it is possible to prioritize, filter and rank contents (a collection, a catalog section, a search result) in a personal way.

Ontology-based representations [12] are richer, more precise, less ambiguous than keyword-based or item-based models. They provide an adequate grounding for the representation of coarse to fine-grained user interests (e.g. interest for individual items such as a sports team, an actor, a stock value) in a hierarchical way, 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 in the system to further 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*, *snakes*, and *chameleons* can be inferred to be interested in *reptiles* with a certain confidence. Also, a user keen of *Madrid* can be assumed to like *Spain*, through the *locatedIn* relation.

3 Emergent Semantic Social Networks

As explained above, our ontology-based personalization framework makes use of explicit user profiles. The users preferences are represented as vectors $U_m = (w_{m1},...,w_{mN})$, where the weights $w_{mn} \in [0,1]$ measure the intensity of the *m*-th user interest for each of the N concepts in the domain ontology. The weights thus represent a way of connecting the concept and the user preferences spaces (top left picture of Figure 1).

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.

The next subsections explain in more detail the steps followed in the clustering process, which is shown in Figure 1. An example will be given afterwards to illustrate our proposal.

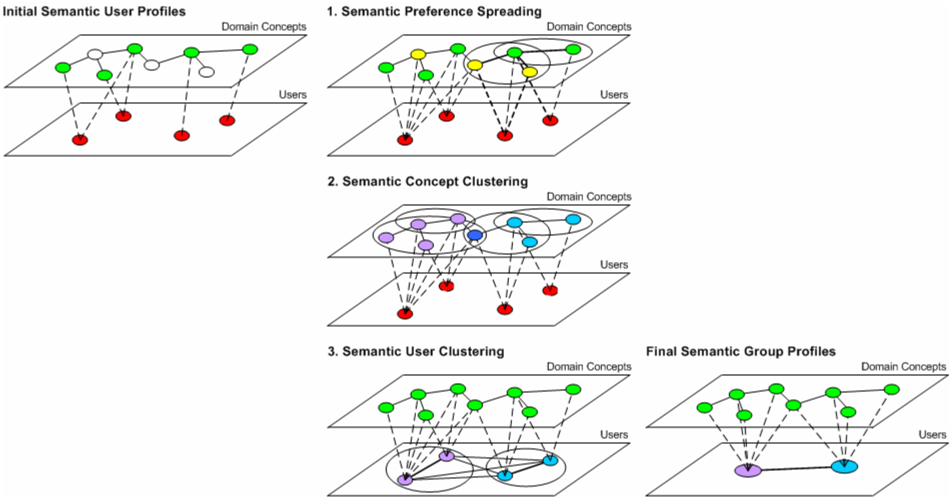
Initial Semantic User Profiles
Domain Concepts
Users
1. Semantic Preference Spreading
Domain Concepts
Users
2. Semantic Concept Clustering
Domain Concepts
Users
3. Semantic User Clustering
Domain Concepts
Users
Final Semantic Group Profiles
Domain Concepts
Users

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

3.1 Semantic Preference Extension

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 numbered 1 in Figure 1). Our approach is based on the Constrained Spreading Activation (CSA) strategy [1,4,5]. The expansion is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed. For example, if an initial profile has a preference about *animals* with a weight of 0.7, the semantic CSA might add to the profile concepts such as *mammals* or *dog*, both of them with associated weights less than 0.7.

We have conducted 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.

The enhancements achieved by the automatic preference extension mechanism show the benefits of an ontology-based representation of user profiles, in contrast to traditional, less expressive ones based on keywords and/or thematic categories.

3.2 Semantic Concept Clustering

In social communities it is fairly accepted that people who are known to share a specific interest are likely to have additional connected interests [9]. For instance, it is easy to understand that, in general, people who like *climbing*, also like topics related to *mountains* or topics related to other *adventure sports*. In fact, this assumption is the basis of most of the existing recommender and collaborative filtering recommender systems [2,10,15]. Here we take into account this hypothesis in order to cluster the concept space in groups of preferences shared by a number of users.

Specifically, for each concept c_n present in at least one of the *M* considered user profiles a vector $\vec{c}_n = (c_{n1},...,c_{nM})$ is assigned, where the component c_{nm} is the weight of concept c_n in the semantic profile of the *m*-th user or 0 if the concept does not appear on it. With these vectors a classical hierarchical clustering strategy [6] is applied. The obtained clusters (picture numbered 2 in Figure 1) thus represent in the concept-user vector space those groups of preferences (topics of interests) shared by the users.

Of course, several issues need to be addressed for the clustering algorithm, such as the distance measure between concepts and clusters, or the appropriate number of final clusters. These will be refined in future work. Here, as we shall explain in section 3.4, we have experimented with a simple example in which the number of clusters is known, and where we have used the Euclidean distance to measure the distances between concepts and an average linkage to measure the distances between clusters [6].

3.3 Semantic User Clustering

Co-clustering based recommender systems involve simultaneous clustering of users and items and generating predictions based on the average ratings of the generated co-clusters [3,8]. Following this idea, once the semantic concept clusters are created, we assign each user to a specific cluster. The similarities between a certain user profile $U_m = (w_{m1},...,w_{mN})$ and the different clusters C_k are computed by the following expression:

$$similarity(U_m, C_k) = \frac{\sum_{n:c_{nm} \in C_k} w_{nm}}{|C_k|} \qquad \textbf{(1)}$$

where c_{mn} represents the concept associated to the *n*-th component of the user profile U_m, and $|C_k|$ is the number of concepts included in cluster C_k. The clusters with highest similarities are then assigned to the users, thus creating groups of users with shared interests (picture numbered 3 in Figure 1).

The obtained user clusters can be used to define underlying 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 provide mechanisms to describe the relations between two distinct types of social items: individuals and groups of individuals. As an applicative development of the obtained user semantic relations, in section 4 we give a first contribution investigating strategies for merging user profiles with common preferences to generate semantic group profiles (picture on the bottom right side of Figure 1). But before that, in the next subsection we describe an artificial experiment that shows an example of evolution and results generated from the presented clustering proposal.

3.4 A simple experiment

In order to check the feasibility of the explained clustering strategy an artificial problem has been set up for this work. The scenario of the problem is the following. A set of twenty user profiles are considered. Each profile is manually defined taking into account six possible topics: *motor*, *construction*, *family*, *animals*, *beach* and *vegetation*. The degree of interest each user has for the different topics are shown in Table 1.

Table 1. Degrees of interest of each user about the six considered topics, and expected user clusters to be obtained with our semantic preference clustering strategy

	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*

For a certain user and a certain topic, a *high* degree of interest means that the user semantic profile has weights close to 1 in some of the concepts corresponding to the topic, a *medium* degree of interest represents weights close to 0.5, and finally a *low* degree of interest indicates weights close to 0.

As it can be seen from table 1, the six first users (1 to 6) have *medium* or *high* degrees of interests in *motor* and *construction* topics. 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* topics. For them a new cluster is expected, named cluster 2. The same situation happens with the next six users (13 to 18). In this case their common preferences are the topics *beach* and *vegetation*. Their expected cluster is 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* topic and User20 should be assigned to cluster 2 due to her high interests in *family* topic.

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)

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.

We have tested our method on this simple ontology and the twenty defined user profiles. In the first step, new concepts are added to the profiles by the Constrained Spreading Activation strategy, 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, $N – 1$ clustering levels are computed, N being the total number of concepts. A stop criterion to set an appropriate number of clusters would be needed, but since in our case the number of expected clusters is three, the stop criterion was not necessary. Table 3 summarizes the users assigned to each final cluster and their similarities values.

It can be seen that the results are totally coincident 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 membership to each cluster. Hence the first two users of each cluster (those with high degrees of interest in their preferred topics) have the highest similarity values inside their clusters, and users 18 and 19, who had the 'noisiest' profiles, present the lowest ones. Regarding user 12, it has to be noted that her exceedingly low similarity value is due to the low preference weights in her profile. Although Table 1 show that this user has *medium* degrees of interest for the *family* and *animals* topics, we assigned her weights close to but always below 0.5.

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

Finally, we show in Table 4 the final concepts obtained and grouped in the semantic Constrained Spreading Activation and concept clustering phases. Although most of them do not appear in the initial user profiles, they help in the construction of the clusters. Our plans for future work include studying in depth the influence of the CSA phase in realistic empirical experiments.

Table 4. Concepts assigned to the different obtained user clusters

Cluster	Concepts
1	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, Fortress, Garage, Road, Speedway, Racing-Circuit, Street, Wind-Tunnel, Pier, Lighthouse, Beach-Hut, Mountain-Hut, Mountain-Shelter, Mountain-Villa, Short-Oval
2	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 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	Water, Sand, Sky Vegetation, 'Tree' (instance of Vegetation), 'Plant' (instance of Vegetation), 'Flower' (instance of Vegetation)

4 Semantic Group Profile Modeling

As an applicative development of our method, we have experimented with building focused group profiles. After computing a multi-layered user network, and finding clusters of users with similar interests, following our previously described approach, the profiles of such users are merged. The group profiles can be built off-line, enabling an efficient strategy for collaborative recommendation in real-time, by using the merged profiles as representatives of classes of users, whereby newcomers can be assigned to a class by comparing their profiles with the joint profile, and then be recommended items based on the group profile.

In order to combine the preferences of groups of users, a number of group modeling strategies based on social choice theory, i.e. deciding what is best for a group given the opinions of individuals, have been applied in a personalized multimedia content retrieval system. The strategies, that have been adapted to consider the semantic (weighted) preferences of our user profile representation, have been empirically tested against real subject opinions about which should be the optimal retrieved multimedia item rankings for a certain set of items and a certain group of users.

In this section, we study the feasibility of applying strategies, based on social choice theory [11], for combining multiple individual semantic profiles in our knowledge-based multimedia retrieval system. Several authors have tackled the problem combining, comparing, or merging content-item based preferences from different members of a group. Here we propose to exploit the expressive power and inference capabilities [1,12] supported by ontology-based technologies.

Combining several semantic profiles with the considered group modeling strategies we pursuit to establish how humans set an optimal multimedia items ranked list for a group, and how they measure the satisfaction of a given item list. The theoretical and empirical experiments performed will demonstrate the benefits of using semantic user preferences representations and exhibit which semantic user profiles combination strategies could be appropriate for a collaborative environment.

In [11] Judith Masthoff discusses several strategies for combining individual user models to adapt to groups. Considering a list of TV programs and a group of viewers, she investigates how humans select a sequence of items for the group to watch. Here we reproduce some of her experiments considering our personalized retrieval system and its semantic user profile representations. In this scenario, because of we have explored the combination of ontology-based user profiles, instead of rating lists, we had to slightly modify the original strategies. For instance, due to items preference weights have to belong to the range [0,1], the weights obtained for a group profile must be normalized.

The following are brief descriptions of the selected strategies.

- **Additive Utilitarian Strategy**. Preference weights from the users of the group are added, and the larger the sum the more influential the preference is for the group.
- **Multiplicative Utilitarian Strategy**. Instead of adding the preference weights, they are multiplied, and the larger the product the more influential the preference is for the group. This could be self-defeating: in a small group the opinion of each individual will have too much large impact on the product. Moreover, in our case it is advisable not to have null weights because we would lose valued preferences. Hence if this situation happens, we set the weights to very small values (e.g. 10^{-3}).
- **Borda Count**. Scores are assigned to the preferences according to their weights in a user profile: those with the lowest weight get zero scores, the next one up one point, and so on. When an individual has multiple preferences with the same weight, the averaged sum of their hypothetical scores are equally distributed to the involved preferences.

- **Copeland Rule**. Being a form of majority voting, this strategy sorts the preferences according to their *Copeland index*: the difference between the number of times a preference beats (has higher weights) the rest of the preferences and the number of times it loses to them.
- **Approval Voting**. A threshold is considered for the preferences weights: only those weights values greater or equal than the threshold value are taking into account for the profile combination. A preference receives a vote for each user profile that has its weight surpassing the establish threshold. The larger the number of votes the more influential the preference is for the group. In the experiments the threshold will be set to 0.5.
- **Least Misery Strategy**. The weight of a preference in the group profile is the minimum of its weights in the user profiles. The lower weight the less influential the preference is for the group. Thus, a group is as satisfied as its least satisfied member. Note that a minority of the group could dictate the opinion of the group: although many members like a certain item, if one member really hates it, the preferences associated to it will not appear in the group profile.
- **Most Pleasure Strategy**. It works as the Least Misery Strategy, but instead of considering for a preference the smallest weights of the users, it selects the greatest ones. The higher weight the more influential the preference is for the group.
- **Average Without Misery Strategy**. As the Additive Utilitarian Strategy, this one assigns a preference the average of the weights in the individual profiles. The difference here is that those preferences which have a weight under a certain threshold (we used 0.25) will not be considered.
- **Fairness Strategy**. The top preferences from all the users of the group are considered. We select only the *N*/2 best ones, where *N* is the number of preferences not assigned to the group profile yet. From them, the preference that least misery causes to the group (that from the worst alternatives that has the highest weight) is chosen for the group profile with a weight equal to 1. The process continues in the same way considering the remaining *N*-1, *N*-2, etc. preferences and uniformly diminishing to 0 the further assigned weights.
- **Plurality Voting**. This method follows the same idea of the Fairness Strategy, but instead of selecting from the *N*/2 top preferences the one that least misery causes to the group, it chooses the alternative which most votes have obtained.

Some of the above strategies, e.g. the *Multiplicative* and the *Least Misery* ones, apply penalties to those preferences that involve dislikes from few users. As mentioned before, this fact can be dangerous, as the opinion of a minority would lead the opinion of the group. If we assume users have common preferences, the effect of this disadvantage will be obviously weaker. For this reason, we shall define the individual profiles with preferences shared by the users in more or less degree.

The mechanism to apply the above strategies in the retrieval process is shown in Figure 2. Combining the semantic user profiles we shall generate a unique semantic group profile that will establish the final multimedia ranking. In the experiments we try to find the group modeling strategy that better fits the human way of selecting items when the personal and collective interests of the group have to be considered.

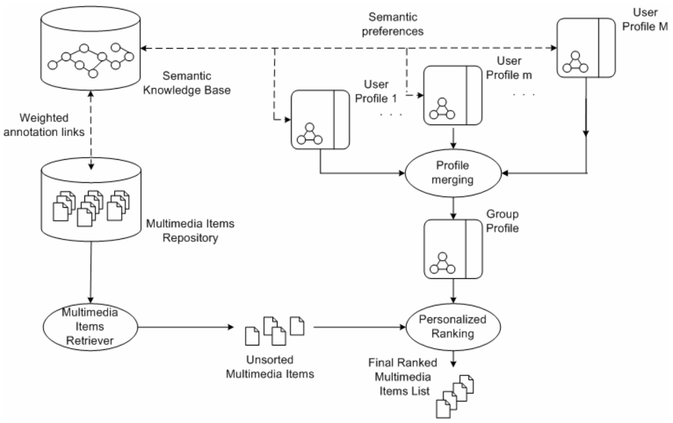
Semantic preferences
User Profile M
Semantic Knowledge Base
User Profile m
User Profile 1
Weighted annotation links
Profile merging
Group Profile
Multimedia Items Repository
Multimedia Items Retriever
Personalized Ranking
Unsorted Multimedia Items
Final Ranked Multimedia Items List

Fig. 2. User profile combination mechanism

The scenario of the experiments was the following. A set of twenty four pictures was considered. For each picture several semantic-annotations were taken, describing its topics (at least one of *beach*, *construction*, *family*, *vegetation*, and *motor*) and the degrees (real numbers in [0,1]) of appearance these topics have on the picture. Ten subjects participated in the experiments. They were Computer Science Ph.D. students of our department. They were asked to assume a group of three users with different interests. In decreasing order of preference: a) User$_1$ liked *beach*, *vegetation*, *motor*, *construction* and *family*, b) User$_2$ liked *construction*, *family*, *motor*, *vegetation* and *beach*, and c) User$_3$ liked *motor*, *construction*, *vegetation*, *family* and *beach*.

To determine which group modeling strategies give ranked lists closest to those empirically obtained from the subjects we have defined a distance that measures the existing difference between two given ranked multimedia item lists. In typical information retrieval systems, where many items are retrieved for a specific query, a user usually takes into account only the first top ranked items. In general, she will not browse the entire list of results, but stop at some top *k* in the ranking. We propose to more consider those items that appear before the *k*-th position of the strategy ranking and after the *k*-th position of the subject ranking, in order to penalize those of the top *k* items in the strategy ranked list that are not relevant for the subject.

Let Ω be the set of multimedia items stored and retrieved by the system. Let $\tau_{sub} \in [0,1]^{|\Omega|}$ be the item ranked list for a certain subject, and $\tau_{str} \in [0,1]^{|\Omega|}$ the ranked item list for a given combination strategy. We use $\tau(x)$ to refer to the position of the multimedia item $x \in \Omega$ in the ranked list τ.

Dwork et al [7] propose the Spearman distance to measure the difference between two search result lists as the average displacement of each document across the rankings. We argue that a more significant measure of impact is whether or not a result will be seen at all by the user. Since in general the user will not browse the entire list of results, but stop at some top *k* in the ranking, there are a number of documents that the user would not see (the ones ranked after the *k*-th result) in the ranking without personalization, but would see as a result of a personalized reordering, and vice versa. If we count the rate of documents in the whole collection that cross the line for each possible value of *k*, and multiply it by the probability *P*(*k*) that the user stops at each *k*, we get a loss function ranging in [0,1] that provides a measure of the effective impact (thus, the risk) of personalization in the retrieval process:

$$d(\tau_{sub}, \tau_{str}) = \sum_{k=1}^{|\Omega|} P(k) \frac{1}{k} \sum_{x \in \Omega} \left| \tau_{sub}(x) - \tau_{str}(x) \right| \cdot \chi_k(x, \tau_{sub}, \tau_{str})$$

where *P*(*k*) is the probability of the user stops browsing the ranked list at position *k*, and

$$\chi_k(x, \tau_{sub}, \tau_{str}) = \begin{cases} 1 & \text{if } \tau_{str}(x) \le k \text{ and } \tau_{sub}(x) > k \\ 0 & \text{otherwise} \end{cases}$$

The expression basically sums the differences between the positions of each item in the subject and strategy ranked lists, as long as they are in the top *k* of the strategy list and are not in the top *k* of the subject list. Thus, the smaller the distance the more similar the ranked lists. The problem here is how to define the probability *P*(*k*). Although an approximation to the distribution function for *P*(*k*) can be taken e.g. by interpolation of data from a statistical study, we simplify the model fixing *P*(10) = 1, assuming that users are only interested in those multimedia items shown in the screen at first time after a query. Our final distance is thus defined as follows:

$$D_{10}(\tau_{sub}, \tau_{str}) = \frac{1}{10} \sum_{x \in \Omega} \left| \tau_{sub}(x) - \tau_{str}(x) \right| \cdot \chi_{10}(x, \tau_{sub}, \tau_{str}) \qquad (2)$$

Observing the twenty four pictures, and taking into account the preferences of the three users belonging to the group, the ten subjects were asked to make an ordered list of the pictures. With the obtained lists we measured the distances D_{1O} with respect to the ranked lists given by the group modeling strategies. We also measure the distances against the lists obtained using semantic user profiles. Figure 3 compares the results.

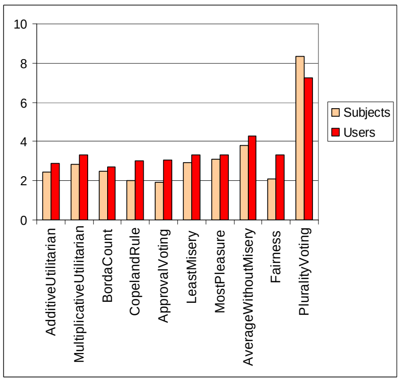

Fig. 3. Average distances D_{10} between subject lists (and user profile ranked lists) and the ranked lists obtained with the different group modeling strategies

Surprisingly, the empirical lists (those obtained from the subjects) and the theoretical (those obtained from the semantic user profiles) agree with the strategies that seem to be more adequate for modeling the group. Strategies like *Borda Count* and *Copeland Rule* give lists more similar to those manually created by the subjects, and strategies like *Average Without Misery* and *Plurality Voting* obtained the greatest distances.

5 Conclusions and future work

A variety of group-based personalization functionalities can be enabled by combining, comparing, or merging preferences from different users, where the expressive power and inference capabilities supported by ontology-based technologies act as a fundamental piece towards higher levels of abstraction.

In this work, we have presented a novel approach to the automatic identification of semantic social communities according to ontology-based user profiles. Taking into account the semantic preferences of several users, the proposed mechanism clusters the ontology concept space, obtaining common topics of interest. Each of the users is assigned to a specific cluster generating groups of users with similar interests. In a further step, these groups of users can be combined in semantic group profiles, which might be used in collaborative and recommendation systems.

Early experiments with a simple artificial problem have been done showing the feasibility of the user clustering strategy. However, more sophisticated and statistically significative experiments need to be performed in order to properly evaluate the model. Several aspects of the clustering algorithm have to be investigated using noisy user profiles: the type of clustering, the distance measure between two concepts, the distance measure between two clusters, the stop criterion that determines what number of clusters should be chosen, or the similarity measure between given clusters and user profiles. Further, a formal comparison with co-clustering methods [3,8] will have to be done.

A number of other open issues have to be addressed in future work. First of all, we plan to make more realistic experiments. In real situations, preferences can not be easily clustered. User profiles usually have noisy components and do not allow to partition the concept space in a clear way. In these cases, we hope the influence of the semantic Constrained Spreading Activation phase will be beneficial for the clustering procedure. Once the user clusters are obtained, a study of the emergent semantic social networks can be done. The preference weights of user and group profiles, the degrees of belonging of the users to each cluster and the similarity measures between clusters, constitute significant mechanisms to describe the relations between two types of social items: individuals and groups of individuals. Furthermore, the user profiles might be segmented in different preference contexts. Thus, the group modeling strategies might be improved merging certain preference contexts instead of the whole individual profiles, enriching thus the obtained semantic social networks. Finally, we are aware of the need to develop an efficient and effective automatic user profile learning algorithm. The correct concepts acquisition and their further classification and annotation in the ontology-based profiles will be crucial to 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. 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 Li-braries 17(8-9), pp. 687-714. Taylor and Francis, 2003.
3. Cheng, Y., Church, G.M.: *Biclustering of expression data*. In Proc. of the 8th Intl. Conference on Intelligent Systems for Molecular Biology (ISM), pp. 93-103, 2000.
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. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: *Rank Aggregation Methods for the Web*. In Proceedings of the 10th Intl. World Wide Web Conference (WWW'01), Hong Kong, 2001.
8. George, T., Merugu, S.: *A Scalable Collaborative Filtering Framework based on Co-clustering*. In Proc. of the 5th IEEE Conference on Data Mining (ICDM), pp. 625-628, 2005.
9. Liu, H., Maes, P., Davenport, G.: *Unraveling the Taste Fabric of Social Networks*. International Journal on Semantic Web and Information Systems, Vol. 2, Issue 1, pp. 42-71. 2006.
10. 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.
11. Masthoff, J.: *Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers*. User Modeling and User-Adapted Interaction, vol.14, no.1, pp.37-85, 2004.
12. Middleton, S.E., Shadbolt, N.R., Roure, D.C.D.: *Ontological user profiling in recommender systems*. ACM Transactions on Information Systems, Vol. 22(1), pp. 54-88, 2004.
13. Mika, P.: *Ontologies Are Us: A Unified Model of Social Networks and Semantics*. Proc. of the 4th International Semantic Web Conference (ISWC 2005), pp. 522-536, 2005.
14. Mika, P.: *Flink: Semantic Web technology for the extraction and analysis of social networks*. Web Semantics: Science, Services and Agents on the WWW. Vol. 3(2-3), pp. 211-223, 2005.
15. 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, 2001, pp. 199-218, 2001.
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.