Modelling Ontology-based Multilayered Communities of Interest for Hybrid Recommendations

Iván Cantador, Pablo Castells, Alejandro Bellogín

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

Abstract. This paper describes a strategy that automatically identifies Communities of Interest (CoI) from the tastes and preferences expressed by users in personal ontology-based profiles, and presents early experiments that evaluate how these CoI can be applied to recommend annotated items combining several content-based collaborative recommendation techniques. Specifically, we have experimented with a set of synthetic profiles generated from data of the well-known IMDb and MovieLens repositories. The obtained results show the feasibility of our CoI identification and recommendation approaches.

Keywords: community of interest, ontology, user profile, content-based collaborative filtering.

1 Introduction

Communities of Interest (CoI) are groups of people who share common interests or passions. However, it is very often the case that the membership to a community is unknown or unconscious. In many social applications, a person describes his interests in a profile to find people with similar ones, but he is not aware of other related interests that might be useful to find those people. In these cases, a strategy to automatically identify CoI might be very beneficial [1].

The issue of finding hidden links between users based on the similarity of their preferences is not a new idea. In fact, it is the essence of the well-known collaborative recommender systems [6], where items are recommended to a user based on his shared interests with other users, or according to ratings of items given by similar users. In typical approaches, the comparison between users is done globally, in such a way that partial, but useful similarities may be missed. For instance, two people may have a highly coincident taste in cinema, but a very divergent one in sports.

We propose a novel approach towards building multilayered CoI by analyzing the individual preferences of users, described in ontology-based profiles, and broken into potentially different areas of personal interest. Like in previous approaches [5], our method builds profiles of user interests for specific concepts in order to find similarities among users. But in contrast to prior work, we divide the profiles into clusters of cohesive interests, and based on this, several layers of CoI are found.

Depending on the current context, only a specific subset of the segments (layers) of a profile should be considered to establish his similarities with other people, enabling more accurate and context-sensitive results in recommendation processes. Thus, as an applicative development of our clustering and CoI building methods, here we evaluate empirically several content-based collaborative filtering models that retrieve annotated items according to a number of synthetic user profiles generated with data from MovieLens[1] and Internet Movie Database[2] (IMDb) repositories.

The rest of the paper has the following structure. Section 2 describes the ontology-based knowledge representation, upon which our personalised content retrieval processes are built. The proposed clustering technique to build multilayer CoI is presented in Section 3. The exploitation of the CoI to perform content-based collaborative filtering is explained in Section 4. Section 5 describes the experiments conducted to evaluate the proposals, and Section 6 includes some conclusions.

2 Personalised Ontology-based Content Retrieval

Our approach uses explicit user profiles. Working within an ontology-based personalisation framework [3], preferences are represented as vectors $\mathbf{u}_m = \left(u_{m,1}, u_{m,2}, ..., u_{m,K}\right)$ where $u_{m,k} \in [0,1]$

[1] GroupLens Research, http://www.grouplens.org/
[2] Internet Movie Database, IMDb, http://imdb.com/

measures the intensity of the interest of user $u_m \in \mathcal{U}$ for concept $c_k \in \mathcal{O}$ (a class or an instance) in a domain ontology $\mathcal{O}$. Similarly, the items $d_n \in \mathcal{D}$ in the retrieval space are assumed to be annotated by vectors $\mathbf{d}_n = (d_{n,1}, d_{n,2}, ..., d_{n,K})$ of concept weights, in the same vector-space as user preferences.

With the above knowledge representation, we use a retrieval model that works in two phases. In the first one, a formal ontology-based query is issued by some form of query interface (e.g. NLP-based) formalising a user information need. The query is processed, outputting a set of ontology concepts that satisfy it. From this point, the second phase is based on an adaptation of the classic vector-space IR model [2], where the axes of the space are the concepts of $\mathcal{O}$, instead of text keywords. The query and each item are thus represented by vectors $\mathbf{q}$ and $\mathbf{d}$, so that the satisfaction of the query is computed by the cosine measure $sim(d,q) = \cos(\mathbf{d},\mathbf{q})$.

The above model is then adapted to include a matching algorithm that provides a personal relevance measure $pref(d,u)$ of an item d for a user u. This measure is set according to the semantic preferences of the user and the semantic annotations of the item based again on a cosine-based vector similarity $sim(d,u) = \cos(\mathbf{d},\mathbf{u})$. In order to bias the result of a search (the ranking) to the preferences of the user, this measure has to be combined with the query-based score without personalisation $sim(d,q)$ defined previously, to produce a combined ranking [3].

Additionally, we perform a semantic preference spreading mechanism, which expands the initial set of preferences stored in user profiles through explicit relations with other concepts in the ontology. Based on Constrained Spreading Activation (CSA) strategies [4], the expansion is self-controlled by applying a decay factor to the intensity of preference each time a relation is traversed, and applying constraints (threshold weights) during the spreading.

3 Multilayered Communities of Interest

It is commonly accepted that people who are known to share a specific interest are likely to have additional connected interests. We assume this hypothesis here as well, in order to cluster the concept space in groups of preferences shared by several users.

We propose 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 considering the preference weights of the user profiles we shall cluster the domain concept space generating groups of interests shared by several users. Thus, those users who share interests of a specific concept cluster will be connected in the network, and their preference weights will measure their degree of membership to each cluster. Specifically, a vector $\mathbf{c}_k = (c_{k,1}, c_{k,2}, ..., c_{k,M})$ is assigned to each concept vector c_k present in the preferences of at least one user, where $c_{k,m} = u_{m,k}$ is the weight of concept c_k in the profile of user u_m. Based on these vectors a classic clustering strategy is applied. The obtained clusters represent the groups of preferences (topics of interests) in the concept-user vector space shared by a significant number of users, and each user can be assigned to a specific cluster. The similarity between a user's preferences $\mathbf{u}_m = (u_{m,1}, u_{m,2}, ..., u_{m,K})$ and a cluster C_q is computed by:

$$sim(u_m, C_q) = \frac{\sum_{c_k \in C_q} u_{m,k}}{|C_q|} \quad \textbf{(1)}$$

where c_k represents the concept that corresponds to the $u_{m,k}$ component of the user preference vector. The clusters with highest similarities are then assigned to the users, thus creating groups of users with shared interests.

In this scenario, the concept and user clusters can be used to find emergent, focused semantic CoI. Taking into account the concept clusters, user profiles are partitioned into semantic segments. Each of these segments corresponds to a cluster, and represents a subset of the user's interests that is shared by the users who contributed to the clustering. By thus introducing further structure in profiles, it is possible to define relations among users at different levels, obtaining a multilayered network of users. The resulting semantic CoI have many potential applications. For example, they can be exploited to the benefit of content-based collaborative filtering recommendations, 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.

4 Hybrid recommendations

Collaborative filtering (CF) 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. We believe that exploiting the relations of the underlying CoI which emerge from the users' interests, and combining them with semantic item information can have an important benefit in CF recommendation. Using our multilayered CoI proposal, we present two hybrid recommendation models that generate ranked lists of items. The first model (labelled UP) is based on the profile of the user to whom the list is delivered. The second model (labelled NUP) outputs lists disregarding the profile. This can be applied in situations where a new user does not have a profile yet, or when the preferences in a profile are too generic for a specific context. 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 clusters, and, b) one that provides a ranking for each cluster. We thus study four different strategies, UP (profile-based), UP-*q* (profile-based, considering a specific cluster C_q), NUP (no profile), and NUP-*q* (no profile, considering cluster C_q).

In the following, for a user profile $\mathbf{u}_m$, an item vector $\mathbf{d}_n$, and a cluster C_q, we denote by $\mathbf{u}_m^q$ and $\mathbf{d}_n^q$ the projections of the vectors onto cluster C_q, i.e. the k-th component of $\mathbf{u}_m^q$ and $\mathbf{d}_n^q$ is $u_{m,k}$ and $d_{n,k}$ respectively if $c_k \in C_q$, and 0 otherwise.

Model UP. The profile of a user u_m is used to return a unique list. The score of an item d_n is computed as a weighted sum of the indirect preferences based on similarities with other users in each cluster. The sum is weighted by the similarities with the clusters.

$$pref\left(d_n, u_m\right) = \sum_q nsim\left(d_n, C_q\right) \sum_i nsim_q\left(u_m, u_i\right) \cdot sim_q(d_n, u_i) \quad (2)$$

where:

$$sim\left(d_n, C_q\right) = \frac{\sum_{c_k \in C_q} d_{n,k}}{\left\|\mathbf{d}_n\right\| \sqrt{\left|C_q\right|}}, \; nsim\left(d_n, C_q\right) = \frac{sim\left(d_n, C_q\right)}{\sum_i sim\left(d_n, C_i\right)}$$

are the single and normalized similarities between the item d_n and the cluster C_q,

$$sim_q\left(u_m, u_i\right) = \cos\left(\mathbf{u}_m^q, \mathbf{u}_i^q\right) = \frac{\mathbf{u}_m^q \cdot \mathbf{u}_i^q}{\left\|\mathbf{u}_m^q\right\| \cdot \left\|\mathbf{u}_i^q\right\|}, \; nsim_q\left(u_m, u_i\right) = \frac{sim_q\left(u_m, u_i\right)}{\sum_j sim_q\left(u_m, u_j\right)}$$

are the single and normalized similarities at layer q between users u_m and u_i, and:

$$sim_q\left(d_n, u_i\right) = \cos\left(\mathbf{d}_n^q, \mathbf{u}_i^q\right) = \frac{\mathbf{d}_n^q \cdot \mathbf{u}_i^q}{\left\|\mathbf{d}_n^q\right\| \cdot \left\|\mathbf{u}_i^q\right\|}$$

is the similarity at layer q between item d_n and user u_i.

Model UP-*q*. The user's preferences are used to return a 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 is selected.

$$pref_q\left(d_n, u_m\right) = \sum_i nsim_q\left(u_m, u_i\right) \cdot sim_q\left(d_n, u_i\right) \quad (3)$$

where q maximizes $sim\left(u_m, C_q\right)$.

Model NUP. The profile of the user is ignored. The ranking of an item d_n is determined by its similarities with the clusters and the profiles of users within each cluster.

$$pref\left(d_n, u_m\right) = \frac{1}{M-1} \sum_q nsim\left(d_n, C_q\right) \sum_{i \neq m} sim_q\left(d_n, u_i\right) \quad (4)$$

Model NUP-*q*. The user's preferences are ignored, and a ranking per cluster is delivered. The ranking that corresponds to the cluster the user is most close to is selected.

$$pref_q\left(d_n, u_m\right) = \frac{1}{M-1} \sum_{i \neq m} sim_q(d_n, u_i) \quad (5)$$

5 Experiments

The MovieLens database is one of the repositories most referenced and evaluated by the Recommender Systems research community. In its large public version, it consists of approximately 1 million ratings for 3,900 movies by 6,040 users on a 1-5 rating scale. It is in turn based on the Internet Movie Database (IMDb) that contains a catalogue of every pertinent detail about a movie, such as the cast, director, shooting locations, languages, soundtracks, etc. In our experiments, we have explored the combination of both sources of data. Specifically, we exploit some of the IMDb information to produce ontology-driven, content-based user profiles (described in Section 2) from the MovieLens ratings. For such purpose, we have defined a domain ontology describing the fundamental concepts involved in IMDb, including classes such as movies, actors, directors, genres, languages, countries, keywords, etc., and relations among them. Then we have parsed the IMDb content (as publicly available in text form), and converted it to an OWL KB, based on the aforementioned movie ontology. Semantic user preferences are then built from the MovieLens ratings by means of a number of transformations exploiting the generated IMDb KB.

5.1 Generating User Profiles from MovieLens ratings and IMDb data

Let $i_{m,1}, i_{m,2}, ..., i_{m,N_m}$ be the N_m items (movies) rated by user u_m and let $r_{m,1}, r_{m,2}, ..., r_{m,N_m} \in [1,5]$ be the corresponding ratings. We define the weight of movie i_n for user u_m as:

$$w_{m,n} = \frac{r_{m,n}}{5} \in (0,1]$$

For each user u_m we measure the relevance of the different movie features by summing the weights of the movies in which these features appear.

$$w_{m,f} = \frac{1}{N_m} \sum_{n:\, f \in features(i_n)} w_{m,n}$$

Taking into account all the movies rated by a user, the feature weights obtained with the previous formulas could be taken as initial semantic user preferences. However, we noticed that we had to filter an appropriate proportion of the features to be included in the final profiles as follows. After we expanded the features, we found that some of them appeared in the user profiles with too many instances, while others with very few. For instance, we observed that in general the initial user profiles contained lots of keywords and very few directors. Furthermore, we obtained a lot of weights with values very close to 0, too low to be considered significant or reliable. According to the cumulative distributions, for each feature, we selected the number of instances that covers approximately 90% of the feature values distribution. By applying this criterion, the resulting user preferences included the 8 top-weighted genres, 3 countries, 15 actors and 3 directors per movie.

5.2 Evaluating the hybrid recommendation models

Conventional recommender algorithms are modelled as ratings estimators. They receive a set of existent user ratings as input and predict new ratings for unseen items. In this context, it is easy to measure the effectiveness of the models if we use evaluations based on the Mean Absolute Error (MAE), i.e., the mean of the absolute differences between the ratings $r_{m,n}$ and their predicted values $p_{m,n}$:

$$MAE = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N_m} \sum_{n=1}^{N_m} \left| r_{m,n} - p_{m,n} \right| \quad (\mathbf{6})$$

However, since our recommender models have been defined under a personalised content retrieval framework that generates rankings with values in [0,1], and aiming to make comparisons with MovieLens ratings, we saw the need to convert our recommendations into 1-5 scale ratings. To tackle this issue, we used again the cumulative distributions. In Figure 1 we show the cumulative distributions F and G of the real MovieLens ratings and the values obtained with our recommenders. To normalize each predicted value $p_{m,n}$ we first map its cumulative probability $G(p_{m,n})$ into the equivalent cumulative probability $F(r_{m,n})$ in the rating value distribution. Then, we calculate its inverse value $F^{-1}\left(G\left(p_{m,n}\right)\right)$ to extract the corresponding rating $r_{m,n}$:

$$r_{m,n} = F^{-1}\left(G\left(p_{m,n}\right)\right)$$

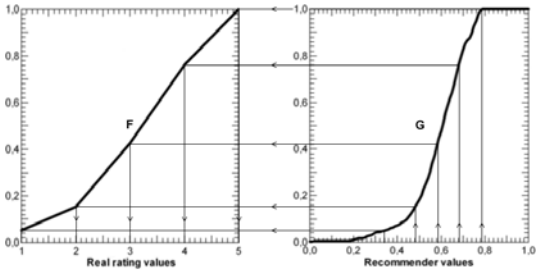

Fig. 1. Cumulative distributions mappings of our recommender values into MovieLens ratings

Once the rating transformations are defined, we are able to evaluate our recommenders by measuring their MAE. To this end, we built the models with 100 users and considering 10% to 90% of their ratings. The rest of their ratings were used for testing. The results are shown in Figure 2.

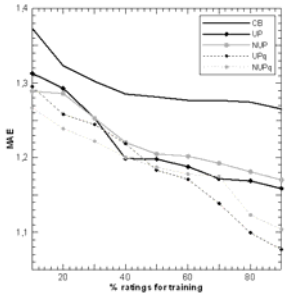

Fig. 2. MAE for our content-based (CB), UP, UP-*q*, NUP and NUP-*q* hybrid recommenders

It can be seen that the user profile-based cluster-oriented UP-*q* model appears to be an appropriate hybrid recommender strategy, outperforming the base line established by our content-based recommender.

6 Conclusions and Future Work

In this work, we have presented an approach to automatically identify Communities of Interest from 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 sub-profiles to the clusters, and the similarities among them, are used to define links that can be exploited by collaborative filtering techniques. Early experiments have been done applying the emergent CoI to a variety of novel semantic content-based collaborative filtering strategies showing the feasibility of our clustering strategy. However, more sophisticated experiments need to be performed in order to properly evaluate the recommendation models.

Acknowledgments. This research was supported by the European Commission (FP6-027685 - MESH) and the Spanish Ministry of Science and Education (TIN2005-06885 - S5T). The expressed content is the view of the authors but not necessarily the view of the MESH or S5T projects as a whole.

References

1. Alani, H., O'Hara, K., and Shadbolt, N. *ONTOCOPI: Methods and Tools for Identifying Communities of Practice*. Proc. of the 17th IFIP World Computer Congress. Montreal,2002
2. Baeza-Yates, and R., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.
3. Castells, P., Fernández, M., Vallet, D., Mylonas, P., and Avrithis, Y. *Self-Tuning Personalised Information Retrieval in an Ontology-based Framework.* 1st IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics. Agia Napa, Cyprus, 2005.
4. Crestani, F., and Lee, P. L. *Searching the Web by Constrained Spreading Activation*. Information Processing and Management 36 (4), 2000.
5. Liu, H., Maes, P., and Davenport, G. *Unravelling the Taste Fabric of Social Networks*. International Journal on Semantic Web and Information Systems 2 (1), 2006.
6. Sarwar, B. M., Karypis, G., Konstan, J., and Riedl, J. *Item-Based Collaborative Filtering Recommendation Algorithms*. Proc. of the 10th Int. WWW Conference. Hong Kong, 2001.