Investigating the Decision Making Process of Users based on the *PoliMovie* Dataset

Mona Naseri, Mehdi Elahi, and Paolo Cremonesi

Politecnico di Milano, Via Ponzio 34/5 20133, Milano, Italy
{mona.naseri,mehdi.elahi,paolo.cremonesi}@polimi.it
http://www.polimi.it

Abstract. Making decision on which movie to watch is nowadays not an easy process for majority of people. Many people may decide to watch a movie based on the genre attribute of a movie, while for the others, the director can be the attribute that drives them to watch a movie. Hence, people may have different features, taken into account, when deciding which movie to watch.
Recommender Systems can help people in making decision by allowing them to enter the attribute(s), that is the most important to them, and filter the movie catalog accordingly.
In this paper we try to investigate the process that results in choosing a movie to watch by people. Hence we present an ongoing work that will ultimately lead to building a dataset (called *PoliMovie*) that will contain the preferences of users not only on movies but also on attributes of the movies, such as genre, director, and cast., that users selected as the most important attributes when choosing a movie to watch.
We report some preliminary results based on the preferences collected from about 400 users, which confirm the difference and complexity of decision making process for different users.

Keywords: Recommender systems, attributes, decision-making, user-profile

1 Introduction

Nowadays, choosing the right movie to watch is challenging for people due to huge variety of the movies. By investigating the way people choose the movies to watch, one may notice that there is no unique way that people may follow in making decision. While some people may decide based on the genre of the movies, others may prefer movies in which their favorite actor plays.

Recommender Systems try to support this process by finding movies that can match users' need and interest. They analyze set of features (attributes) of the items and create user profile for a user, that indicates the preferences and interests of her for those features. Indeed, recommendations are generated by matching up the features of the user profile (i.e., a structured representation of her interests) against the features of the item. In order to do this, the content-based recommender systems build a Vector Space Model (VSM), where each item is represented by an n-dimensional vector. Each dimension in this model

represents an attribute from the overall set of attributes used to describe the item. Using this model, the system computes a relevance score that represents the user's degree of interest toward that item [3]. This also allows the recommender systems to produce explanations to recommendations and to naturally solve the new item problem [1].

These systems typically include also this side information describing the attributes of movies (e.g., in the movie domain categories like genre, director, cast) and build user models as prediction of users' preference on features [6,5]. Incorporating such information is beneficial since it implicitly helps the system to understand important attributes that users may base their reasoning in order to choose the right movie to watch.

Moreover, traditionally, such user models are tested using their ability to provide relevant recommendations of movies. Hence, one of the publicly available datasets such as Movielens is used as a ground truth for the preferences of users on movies. However, as of our knowledge, none of the publicly available datasets contain the explicit preferences of users on movie attributes, and hence, the user models can not be evaluated for the true prediction of users preferences on attributes, that they may choose movies to watch based on them.

In this paper, we have analyzed our dataset, called *PoliMovie* [4], and obtained some preliminary results that can show clear differences between the way different users may form their reasoning process when making decision on what to watch. These results confirm our initial intuition that traditional user models based on implicit user preferences on attributes do not match well with explicit user models in which users are explicitly elicited to provide their opinions on attributes.

We show that, in many cases, the favorite movies selected by a user have attributes (e.g., actors, directors, genre) totally different from the attributes selected as favorites by the same user. Hence, a user may like "The Dark Knight" movie without choosing "Action, Crime, Drama" as favorite genre, "Christopher Nolan" as favorite director, and "Christian Bale" as favorite actor. We show that a big group of users have chosen attributes that they choose movies based on them, that have not even contained in the movies they actually choose as their favorite. These results are convincing us to even continue the data collection procedure as well as a more extensive analysis of the data.

2 Preliminary results

In this section we report some preliminary results of analysis we have conducted on this initial collection of data. Up to the date of writing this paper, PoliMovie dataset contains approximately 1600 movies, 300 casts, 200 directors and all genres (23) have been selected as favorite at least by one user.

First of all, we have understood that the most important attributes, the users take into account when making decision on which movie to watch are "Genre" and the "Cast" of the movies, respectively. However, the "Director" and the "Year of Production" play the least important role in decision making for choosing which movie to watch.

Accordingly, we have measured the popularity of the cast based on the users explicit rating and also based on implicitly inferring from the movies they have

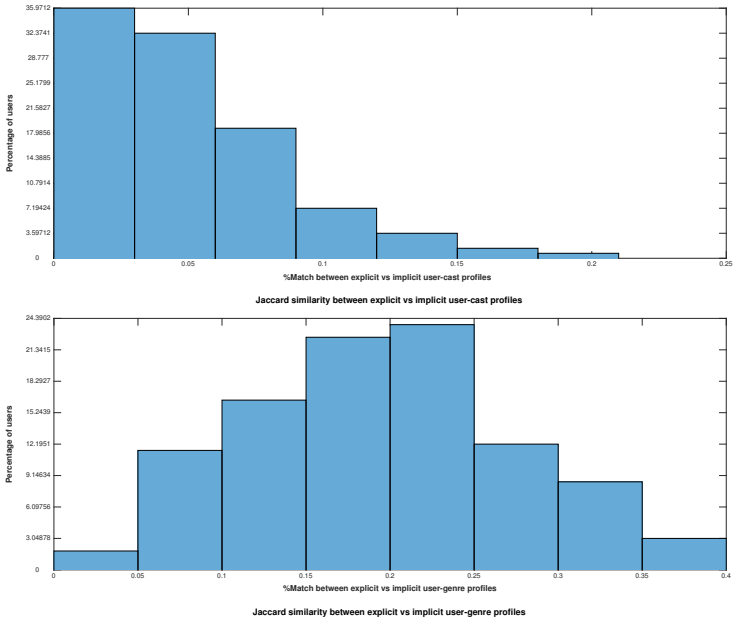

Fig. 1. Jaccard similarity implicit and explicit profile of the users based on their (top) favorite cast and (button) favorite genres

added as favorite. Comparing these two lists, we have noticed that some actors/actresses in the first list (based on explicit rating) are not actually presented in the second (based on implicit preferences) or vice versa.

Moreover, our observation shows that, based on users' favorite movies, "Drama" is the most popular genre. This is while, the most popular genre, selected by users, is totally the opposite genre, i.e. "Comedy".

Additionally, we have computed the Jaccard similarity [2] between the user profiles, based on explicit and implicit user preferences. We considered only those features that user selected as her most important feature(s). Each user could select maximum two among five movie features (genre, cast, director, rating, production-year). For instance, if user selected "Genre" and "Cast" as her most important features, we measured similarity between her explicit and implicit profiles only based on these two features.

We have observed interesting results shown in figure 1, where x axis is the percentage of match, and y axis is percentage of users with certain match between their explicit and implicit profiles. The observations show that user models based on implicit user preferences on attributes do not match well with explicit user models in which users are explicitly elicited to provide their opinions on attributes.

Finally, it worth noting that, the distribution of similarities are totally different for favorite casts and favorite genres. Indeed, the distribution of the match

between two user profiles based on cast looks exponential while the distribution based on genre interestingly looks Gaussian with the mean about 0.2. This may indicate that majority of users may have only 20% of match between the genre they like and the genre of the movies they watch. Accordingly, there are people with more match or less match between these two types of preferences. However, for the user's favorite cast, this may not be the case and most of the users may watch a lot of movies where their favorite cast do not play.

3 Conclusion

This paper presents an ongoing work of data collection that will ultimately result in a dataset called *PoliMovie*. The PoliMovie dataset is publicly available [1]. In spite of the currently available datasets, it will contain not only the preferences of users provided for movies, but also preferences of users on features (attributes) of movies. Such dataset can be very useful in investigating the difference between attributes that users consider when choosing movies to watch, as well as, their final decision on which movies they actually watch. Hence, the researchers and the practitioners in the community of Recommender Systems can use such dataset to evaluate the quality of explanations provided by some types of recommender systems as well as analyzing the complex process of decision making by users while also benchmarking their feature-based recommendation algorithms. For our future work, we are going to evaluate some of the state-of-the-art algorithms on our dataset. This would be useful to see which algorithms can better predict preference of users on features (attributes). Moreover, by having the explicit opinion of users on features, we can evaluate the quality of explanations provided by some types of recommender systems.

References

1. M. Elahi, F. Ricci, and N. Rubens. Active learning strategies for rating elicitation in collaborative filtering: a system-wide perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):13, 2013.
2. M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.
3. P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.
4. M. Nasery, M. Elahi, and P. Cremonesi. Polimovie: a feature-based dataset for recommender systems. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, 2015.
5. Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, May 2014.
6. P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Feature-weighted user model for recommender systems. In *Proceedings of the 11th International Conference on User Modeling*, UM '07, pages 97–106, Berlin, Heidelberg, 2007. Springer-Verlag.

[1] through the link: `http://recsys.deib.polimi.it/polimovie-dataset`