How social relationships affect user similarities

Alan Said
DAI-Labor
Technische Universität Berlin
alan.said@dai-labor.de

Ernesto W. De Luca
DAI-Labor
Technische Universität Berlin
ernesto.deluca@dai-labor.de

Sahin Albayrak
DAI-Labor
Technische Universität Berlin
sahin.albayrak@dai-labor.de

ABSTRACT
In this paper we present an analysis of the social movie recommendation community *Filmtipset*. Filmtipset is Sweden's largest movie recommendation community with more than 80,000 users. The website offers movie recommendations based on usage data, but also consists of a social network where users are able to befriend one another. All content is user-generated and there is a multitude of features that make the dataset stand out among other movie recommendation datasets. We evaluate the social graphs' impact on users similarities in taste in movies, and show that utilizing this relation could be used to improve movie recommendation quality.

ACM Classification Keywords
E.0 Data: General; H.3.3 Information Systems Applications: Information Search and Retrieval; H.4.m Information System Applications: General

General Terms
Analysis, Data

Author Keywords
machine learning, movie recommendations, social networks, collaborative filtering

INTRODUCTION
Movie recommendation websites have been an integral part of the Web for almost as long as the Web has been around, one of the first being The Internet Movie Database[1], actually predating the first web browser[6]. In the last couple years movie recommenders have experienced a renaissance with a multitude of new services trying to establish themselves as the best one available, e.g. *Jinni*[2] and *Moviepilot*[3]. In this paper we focus on a veteran in the field, namely Filmtipset[4],

---

[1] http://www.imdb.com
[2] http://www.jinni.com
[3] http://www.moviepilot.com
[4] http://www.filmtipset.se

which is a Swedish online community that has been offering movie recommendation to its users for almost ten years.

One of the reasons movie recommenders have been a hot topic throughout the last couple of years has been the public availability of real-world data. Thanks to services such as *MovieLens*[5], which makes data available via the *GroupLens*[6] research team, and *Netflix*[7], who have made their datasets available as part of the *Netflix Prize*[8]. The amount of research conducted in this domain has been increasing steadily. There is one drawback though, the openly available datasets have all had the same, or a very similar, structure. Most of these datasets provide researchers with two basic relations – a rated user-movie relation consisting of a users' rating for a particular movie, and a movie-genre relation stating which genre, or genres, a movie belongs to. GroupLens, being one of the exceptions, provides the GroupLens 10M100K dataset[9] which also contains personal tags assigned by users to movies. Filmtipset, on the other hand, not only provides relations well known from other datasets, it also introduces several other features, such as actor ratings, review ratings, and a social graph connecting its users.

The focus of this paper is to evaluate and understand the implications of the user-user relationships found in the social graph in a recommendation scenario. Where most user similarities are derived from the user-movie relations, we use user-user relations instead. For this we use a snapshot of the social graph of the movie recommendation community Filmtipset in order to find how user similarities in movie taste correlate to the social relations between users.
We show that friendship does, in fact, affect users taste in movies and that this friendship increases user similarities by nearly 100%.

RELATED WORK
As mentioned in the previous section, there is a vast amount of related work in the field of movie recommendation. The winners of the Netflix Prize, for instance, published a quite provocative paper [8] stating, and showing, that metadata is of little value when it comes to predicting movie ratings. In [2], Amatriain et al., pose that re-rating movies is of signif-

---

[5] http://www.movielens.org
[6] http://www.grouplens.org
[7] http://www.netflix.com
[8] http://www.netflixprize.com
[9] http://www.grouplens.org/system/files/README_10M100K.html

icantly higher value than rating new ones. They show how the amount of time that has passed since the original rating affects the users' new rating, and thus the quality of the recommendations. Other research, not concerning the Netflix Prize, has explored other aspects of movies, such as in [1] where the authors use cultural metadata crawled from comments and reviews in order to boost recommendation results. Ono et al. [7] have instead tried to find the present context of the user in order to give recommendations suiting the users' current situation.

Guy et al. [5] create a system for recommending items based on a users' aggregated *familiarity* network, where relations are extracted from different sources such as co-authorship on wiki pages. The results show that the familiarity network produces better recommendations than classical similarity based approaches. A similar approach is presented in [3] by Bonhard and Sasse.

Golbeck and Hendler [4] are among the very few ones touching on a concept in movie recommendation similar to the one described in this paper. Their approach is based on explicitly defined trust gathered through the *FilmTrust*[10] movie recommendation website. FilmTrust asks its users to explicitly assign trust values to their peers, thus stating whose taste to follow and whose not to follow. They conclude that trust does add to the quality of the recommendations. We consider the concept of trust to be related to the explicitly stated friaendship relation available in Filmtipset.

DATASET AND EXPERIMENTS

Filmtipset is Sweden's largest online movie community and has been available to its users since 2000. The service has grown in size and number of features since it started almost ten years ago. At the time of the writing of this paper Filmtipset had more than 80,000 users, 70,000 movies, almost 19 million ratings and more than 10,000 daily visitors. However, its data has never been analyzed and worked on for reasons other than the recommendation service on the website. Therefore we will describe some of the features and attributes of the community not explicitly used in our experiments.

Figure 1 shows a screenshot of the Filmtipset website, similar to most other movie recommendation websites it contains mainly movie related news and links taking the user further into the website.

Figure 2 shows the full set of existing entities and relations in the complete Filmtipset dataset. These features include entities such as user-generated lists, reviews, review ratings, actor comments, etc[11]. The focus of this paper is the social graph contained in Filmtipset, thus we do not use the full set of entities and relations. The ones used are indicated by the bolder font and darker color used in Figure 2.

The dataset

The dataset used in in this paper is a snapshot of a subset of the features available on the website. We focus on the social user-user relations and the user-movie ratings for our experiments and conclusions. This snapshot consists of all ratings
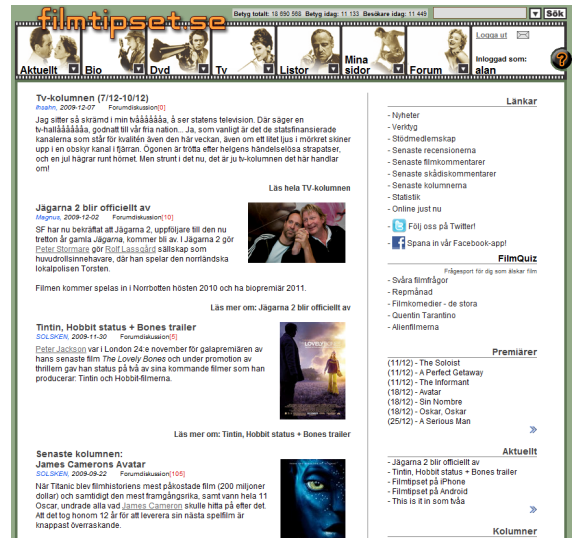
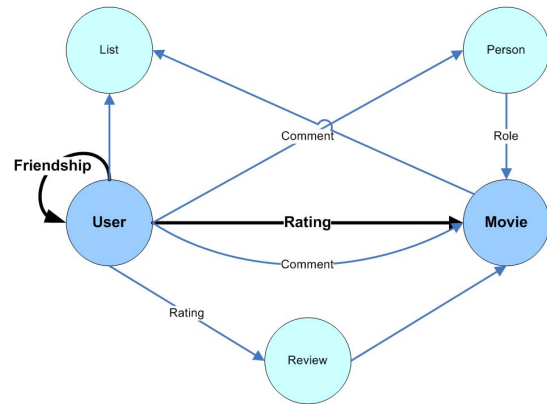**Figure 1. A screenshot of the Filmtipset website.**

Figure 2. The full relationship diagram of the Filmtipset dataset. Relations and entities used are indicated by bold fonts and darker color.

posted to the system between April 18, 2000 and September 14, 2009 and all existing friendship relations at the latter date. The exact numbers of entities and relations are shown in Table 1 and Table 2.

Entity	No.	Entity	No.
Movies	67,684	Actors	137,548
Users	76,505	Directors	28,077
Comments	1,205,160	Writers	41,830
Reviews	6,157	Genres	39

Table 1. The entities provided in the dataset.

One of the key features of this website, if not the main, is the social network – users have the possibility to befriend one another in an asymmetric fashion. Asymmetric friendship relations are similar to the follower/following relation on Twitter[12], meaning that if user u_a choses to befriend user

Relation	No.	Relation	No.
$u_a \rightarrow u_b$	29,443	Ratings	18,074,899
$u_a \leftrightarrow u_b$	53,041	Genre assignments	171,850
		Production roles[a]	553,981

[a] i.e. director, actor, writer

Table 2. The relations provided in the dataset.

u_b, the latter users' social graph remains unchanged (in fact, u_b is not even notified of the event), whereas u_a will have an explicitly stated friendship relation to u_b.

We take a closer look at the characteristics of the two kinds of friendship relations, i.e. $u_a \rightarrow u_b$ and $u_a \leftrightarrow u_b$, asymmetric and symmetric friendships. The asymmetric relations $u_a \rightarrow u_b$ can be expressed as; u_a is a *fan* of u_b. u_b's relation graph remains unchanged as the user has not taken a similar action. In the remainder of this paper we will call the relation from u_b to u_a *friend*, as in u_b is a friend of u_a, the opposite relation will thus be called *fan*. Furthermore it should be noted that the asymmetric relation $u_a \rightarrow u_b$ is identical to $u_b \leftarrow u_a$. In the symmetric case, $u_a \leftrightarrow u_b$, u_a and u_b are both users who have chosen to befriend one another.
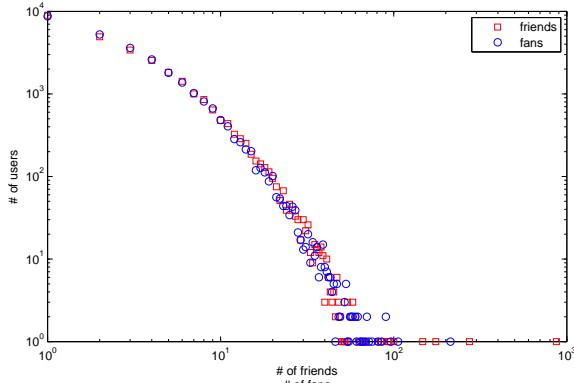

Figure 3. The degree distributions of the numbers of friends and fans following a typical power-law distribution.

The degree distribution of the friend and fan relations is shown in Figure 3. We see that both relations follow a typical power-law distributions and are almost identical. This graph contains the symmetric relations as well as the asymmetric ones, as a symmetric relation simply is two asymmetric ones.
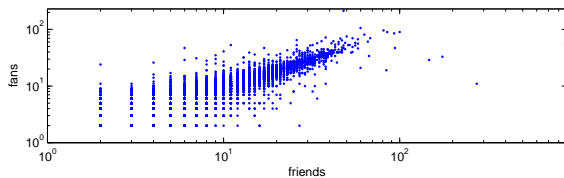
Figure 4. Fans versus friends.

In Figure 4 we show the ratios of friends versus fans. We see that the majority of the users that are part of the social graph have less than 50 friends, the average being 4.6 friends and 4.4 fans.
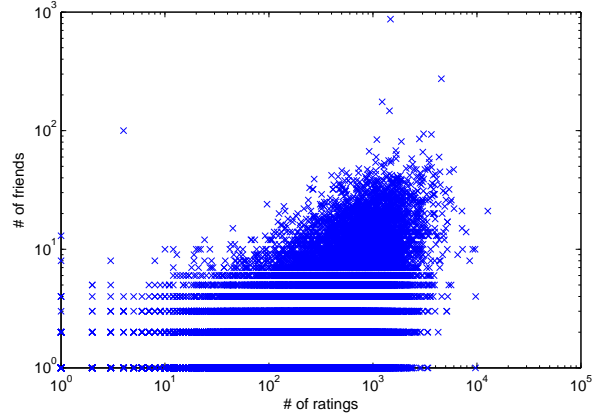
Figure 5. Distribution of number of friends versus number of ratings.

Figure 5 shows the distribution of number of friends versus the number of ratings, we see that if a user has a large number of friends, it is an indication that the same user will often have rated a large number of movies. The opposite rule does not apply though, i.e. a large number of ratings does not imply a large number of social relations.

Experimental setup

We evaluate the significance of user-user relations on the overall taste of users, i.e. the number of similar movie every related user pair has seen. In order to do this we convert the ratings in the user-movie matrix from the $\{1, 2, 3, 4, 5\}$ scale that Filmtipset uses to a binary representation. This creates a user-movie relation where users have either seen (1) or not seen (0) a movie. We call this resulting matrix C_{all}. We then create two additional versions of this matrix; one where all users part of the matrix have at least 5 friends (C_{um5}), and one where all movies and users in the matrix appear at least 5 times (C_{r5}). These three matrices are then used to calculate the average *Jaccard similarity coefficient* for all related user pairs for the user-user relation types declared in Table 2. The Jaccard similarity coefficient is defined as

$$J(A|B) = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

where A and B are the two sets to be compared. We calculate the average Jaccard similarities, $\overline{J}$, for all related user pairs in the three matrices according to

$$\overline{J} = \frac{2}{n(n-1)}^{-1} \sum_{i<j} \frac{|A_i \cap B_j|}{|A_i \cup B_j|} \qquad (2)$$

where n is the number of users, and i and j two related users. One additional setting is added, namely the pairwise comparison of users who are not part of the social graph, however, this is only done for the original C_{all} matrix.

The results of these calculations show us the relation-specific average Jaccard similarities between users, e.g. how similar symmetric friends, asymmetric friends and users outside of the social sphere are. The higher this similarity value is, the more movies the users have in common.

RESULTS

The Jaccard similarities for every relation type and every dataset used are presented in Figure 6. The figure depicts the Jaccard similarities for all user pairs in C_{all} ($\forall$), asymmetric friendship pairs in C_{all} ($\leftarrow$), all user pairs where none of the users have any friends in C_{all} ($\emptyset$), all symmetric friendship pairs in C_{all} ($\leftrightarrow$), asymmetric friendship pairs in C_{um5} ($\leftarrow_{um5}$), symmetric friendship pairs in C_{um5} ($\leftrightarrow_{um5}$), asymmetric friendship pairs in C_{r5} ($\leftarrow_{r5}$) and symmetric friendship pairs in C_{r5} ($\leftrightarrow_{r5}$). We see that the three columns depicting the symmetric friendship similarities ($\leftrightarrow$, $\leftrightarrow_{um5}$ and $\leftrightarrow_{r5}$) are almost twice as high (0.18, 0.18 and 0.20) as the average similarity (0.10) of all user pairs ($\forall$).

The higher similarity values in the C_{um5} matrix are expected as we have removed unpopular movies and users who had seen few films, i.e. users having a higher probability of low similarity values to others. The highest similarity value is obtained from the C_{r5} matrix where all users with less then 5 friends have been removed. This indicates that the friendship relationship between users correlates to a similarity in movie taste.
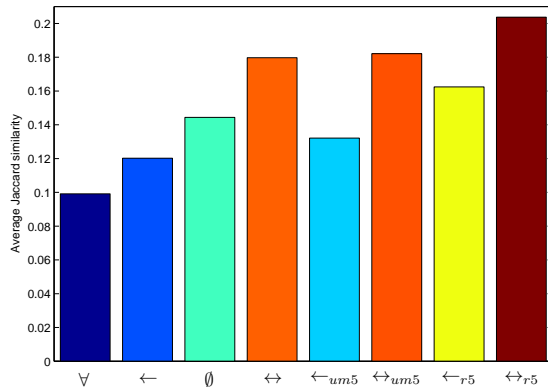

Figure 6. The Jaccard similarities for all user pairs.

CONCLUSIONS AND FUTURE WORK

We believe that the reason for the higher similarity in the reduced social graph is due to an implicit relation of this graph to the user-movie relations. The results suggest that the reduction of the social graph to a smaller core increases the density of the user-movie matrix, thus, by utilizing the user-user relations in a recommendation scenario one could improve the quality of recommendations. We believe this to be related to the findings of Pilászy and Tikk [8], where movie metadata showed to be of little importance compared to ratings. Our approach, on the other hand, uses information about the users for discovering how social relations can be utilized for recommendation purposes.

These relations provide us with a basis for future explorations of the additional features available in the Filmtipset dataset. We are currently exploring the details of these attributes in the recommendation scenario. Further planned work will focus on extending the user similarities with user-oriented features which, in combination with the obtained results, can further improve the performance of recommendation algorithms.

REFERENCES
1. Shinhyun Ahn and Chung-Kon Shi, 'Exploring movie recommendation system using cultural metadata', in *Proc. of the 2008 Intl. Conf. on Cyberworlds*, pp. 431–438. IEEE Computer Society, (2008).

2. Xavier Amatriain, Josep M. Pujol, Nava Tintarev, and Nuria Oliver, 'Rate it again: increasing recommendation accuracy by user re-rating', in *Proc. of the third ACM conf. on Recommender systems*, pp. 173–180, New York, New York, USA, (2009). ACM.

3. P. Bonhard and M. Sasse, 'Knowing me, knowing you using profiles and social networking to improve recommender systems', *BT Technology Journal*, **24**(3), 84–98, (July 2006).

4. J. Golbeck and J. Hendler, 'FilmTrust: movie recommendations using trust in web-based social networks', in *Consumer Communications and Networking Conf., 2006. CCNC 2006. 3rd IEEE*, volume 1, pp. 282–286, (2006).

5. Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman, 'Personalized recommendation of social software items based on social relations', in *Proc. of the third ACM conf. on Recommender systems*, pp. 53–60, New York, New York, USA, (2009). ACM.

6. Col Needham. IMDb history. `http://www.imdb.com/help/show_leaf?history` (retrieved on December 9, 2009).

7. Chihiro Ono, Mori Kurokawa, Yoichi Motomura, and Hideki Asoh, 'A Context-Aware movie preference model using a bayesian network for recommendation and promotion', in *User Modeling 2007*, 247–257, Springer-Verlag, (2007).

8. I. Pilászy and D. Tikk, 'Recommending new movies: even a few ratings are more valuable than metadata', in *Proc. of the third ACM conf. on Recommender systems*, pp. 93–100, New York, New York, USA, (2009). ACM.