Putting Things in Context: Challenge on Context-Aware Movie Recommendation

Alan Said
DAI Lab
Technische Universität Berlin
Berlin, Germany
alan.said@dai-lab.de

Shlomo Berkovsky
CSIRO
Tasmanian ICT Centre
Hobart, Tasmania, Australia
shlomo.berkovsky@csiro.au

Ernesto W. De Luca
DAI Lab
Technische Universität Berlin
Berlin, Germany
ernesto.deluca@dai-lab.de

ABSTRACT

The Challenge on Context-Aware Movie Recommendation (CAMRa) was conducted as part of a join event on Context-Awareness in Recommender Systems at the 2010 ACM Recommender Systems conference. The challenge focused on three context-aware recommendation tasks: time-based, mood-based, and social recommendation. The participants were provided with anonymized datasets from two real world online movie recommendation communities and competed against each other for obtaining the highest recommendation accuracy for each task. The datasets contained contextual features, such as mood, plot annotation, social network, and comments, normally not available in movie recommendation datasets. Over 40 teams from 20 countries participated in the challenge. Their participation was summarized by 10 papers accepted to the CAMRa workshop.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Information Filtering, Selection Process

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Recommender systems, context-aware personalization, context modeling, social networks, datasets, social network analysis, user modeling

1. INTRODUCTION

The importance of context and contextualized user data for accurate recommendations has been widely recognized [1–3]. However, the vast majority of existing recommendation techniques focus on recommending the most relevant items to users and do not take the context into consideration. This results in somewhat static recommendations, which only change after a substantial user interaction with the recommender, sufficient to change their user-to-user similarity (in collaborative recommendations) or item preferences (in content-based recommendations). Hence, the recommendations a user is presented with are independent of contextual factors such as day and season, location of the user and other users, upcoming and ongoing events, or weather. Although this static approach can generate reasonably accurate recommendations, there is space for improvement through incorporating contextual factors.

The Challenge on Context-aware Movie Recommendation (CAMRa2010[1]) aimed at boosting the research of context-awareness in recommender system. Two datasets, gathered by the Moviepilot and Filmtipset online movie recommendation communities, were released exclusively for the challenge. The participants could participate in any of the following three recommendation tracks: (1) recommend movies based on the time of the year and special events, (2) recommend movies based on social relations of users, and (3) recommend movies based on a user's (implicit) mood. The participants were requested to evaluate the performance of their solutions using the Mean Average Precision (MAP), Precision@K (K=5 and K=10), and Area Under Curve (AUC) metrics [8].

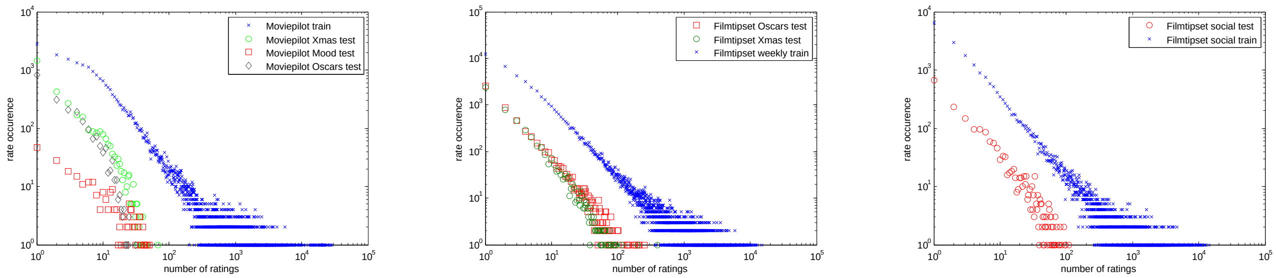
The participants were invited to submit papers summarizing the developed algorithms, obtained results, and one page summaries of their experience with the CAMRa challenge. The papers were evaluations by the Program Committee and algorithmic summaries by the Expert Panel members. Both the evaluations, as well as the results obtained by the participants, were used to identify the winners of the challenge, who were announced at the 2010 ACM Recommender Systems conference dinner.

2. THE CHALLENGE

The challenge served as a means to boost research of context-awareness in recommender systems. This was achieved through gathering researchers and practitioners working on recommender systems and letting them investigate the same challenges, while evaluating the proposed solutions using the same datasets and evaluation metrics. This semi-controlled environment would allow a just comparison of the solutions and algorithms developed by the participants. Moreover, the participants were expected to only use the provided anonymized datasets (presented in

CAMRa2010, September 30, 2010, Barcelona, Spain.

[1] `http://www.dai-labor.de/camra2010/`



(a) Moviepilot. (b) Filmtipset weekly. (c) Filmtipset social.

Figure 1: Number of ratings vs. the rate occurrence of the training and testing datasets.

Section 4), while the use of external information sources, like IMDb, Wikipedia, or NetFlix, was prohibited.

The datasets released for the challenge were partitioned into a training set and a test set (presented in Section 3). The latter contained the ratings provided by users, for which the participants were requested to generate recommendations according to the conditions of their track. The participants were requested to evaluate their recommendations with the ratings in the test sets and report the obtained Precision@5, Precision@10, Mean Average Precision and Area Under the Curve scores. In addition, final evaluation datasets were withheld by the organizers and released only for the final identification of the challenge winners.

3. TRACKS

There were three tracks in the challenge, each focusing on a different aspect of context-awareness. The first *weekly* track was divided into two sub-tracks and could exploit either the Moviepilot or the Filmtipset dataset. The other two, *mood* and *social* tracks, were dataset specific and focused on explicit features, available in, respectively, Moviepilot or the Filmtipset datasets.

3.1 Weekly Recommendation

The weekly recommendation track addressed the contextual temporal dimensions. The task was to recommend movies for two specific weeks: the week leading up to Christmas 2009 (Xmas), and the week leading up to the Oscars ceremony 2010 (Oscar). The rationale for this track implies that events taking place on a certain week can influence user preferences and the movies watched. For example, during the Christmas week there is a higher number of Christmas related movies, whereas during the week leading up to the Oscars ceremony there is a higher number of past Oscar winning movies. Hence, this is expected to influence a users movie selection.

The participants of this track were requested to recommend a set of movies for the two above mentioned weeks for users found in the test sets. The track was divided into two sub-tracks: for the Moviepilot dataset and for the Filmtipset dataset. As the datasets contained different features, different algorithms and feature sets could be used in the recommendation process. The evaluation users have been stripped of some of the movies they had rated during these two weeks.

3.2 Mood - Moviepilot

The Moviepilot track addressed the contextual dimension related to a user's mood. The Moviepilot dataset contains hierarchically organized features such as the movie mood, movie location, plot, and the intended audience. The rationale for this track implies that the mood of a movie could implicitly reflect the mood of a user at the time of watching the movie. This information could then be exploited to generate mood-based recommendations for certain other point of time, when a similar mood is conjectured.

The participants were of this track were requested to recommend a set of movies for a selection of users based on a given mood. The mood selected for this track was "Eigenwillig", i.e. "weird". However, the participants were not provided with the textual description of the mood, but rather with an anonymized numeric representation of it. The evaluation users have been stripped of some of the movies labeled with the "Eigenwillig" mood.

3.3 Social - Filmtipset

The Filmtipset track addressed the social links between users in the Filmtipset dataset. Many Filmtipset users contribute to the social network of the service. In particular, users can befriend each other asymmetrically, similarly to the follower/following relation in Twitter. The dataset contains additional features, such as movie comments, comments on actors/directors/writers, movie reviews, review ratings, lists of favorite movies, and links between similar movies. The rationale for this track implies that online friendship links between users reflect their implicit similarity and can influence user preferences and the movies watched, as shown in [12].

The participants of this track were requested to recommend a set of movies for users in the test set. The evaluation users were stripped of some of the movies they had seen at later points in time than at least three of their friends.

4. DATASETS

The datasets used in the challenge were released exclusively for the challenge by movie recommendation websites Moviepilot[2] and Filmtipset[3]. Due to privacy considerations, the datasets were anonymized prior to the release. Four versions of the datasets were created: two Moviepilot sets and two Filmtipset sets, i.e. one version for each track and subtrack. Each dataset was bundled with a test set, which the participants were expected to use in the evaluations of their

[2] `http://www.moviepilot.com`

[3] `http://www.filmtipset.se`

	Users	Movies	Ratings
Train	105,137	25,058	4,544,409
Xmas Test	160	3,377	16,174
Xmas Eval	80	2,153	6,701
Oscar Test	160	2,144	8,277
Oscar Eval	80	1,520	4,169
Mood Test	160	251	2,656
Mood Eval	80	220	1,421

Table 1: The number of movies, users and ratings in the Moviepilot datasets.

	Assignments
Mood	6,712
Plot	92,124
Time	3,687
Place	8,586
Audience	2,436

Table 2: Tag assignments in the Moviepilot dataset.

solutions. In addition, four evaluation sets were withheld till after the submission deadline and were subsequently released to the participants. The results obtained using these datasets were used to identify the winners of the challenge.

The test datasets were generated through randomly seeded algorithms. In order to verify the soundness of the datasets, some statistical features were observed prior to release. Figure 1 shows the number of ratings vs. rate occurrence in the datasets. The rating distribution in all the datasets follows power law distribution, demonstrating that there were no rating related anomalies in the datasets.

4.1 Moviepilot

Moviepilot is the leading online movie and TV recommendation community in Germany. It has over 100,000 registered users and a database of over 40,000 movies with roughly 7,5 million ratings. The Moviepilot datasets used for CAMRa were extracted from a snapshot of the dataset taken in April 2010.

The Moviepilot datasets were based on ratings provided between January 31st 2008 and March 10th 2010. The ratings were split into seven parts: the training set, the test set for the Christmas week, the evaluation set for the Christmas week, the test set for the Oscars week, the evaluation set for the Oscars week, the test set for the mood track, and the evaluation set for the mood track. The details of these datasets are shown in Table 1.

In addition to ratings, the datasets included several other features, such as movie-mood tags, intended audience tags, favorited/hated movies, place and time of the movie, and plot tags. An abstract entity-relationship diagram of the datasets and their features is shown in Figure 2. All the data referring to user-movie pairs found in the test and evaluation sets and in the corresponding weeks (Christmas 2009 and Oscars week 2010) were removed in order to not reveal information not available at the time of the recommendations. The tag assignment statistics in the training dataset are shown in Table 2.

The weekly recommendations dataset consisted of two track-specific rating data files, and one training file which

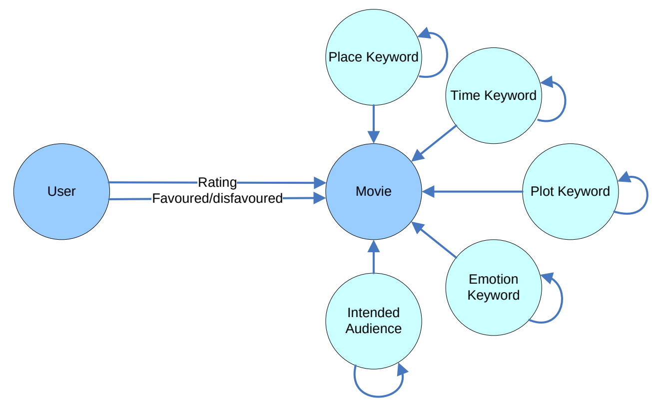


Figure 2: An abstract entity-relationship diagram of the Moviepilot dataset.

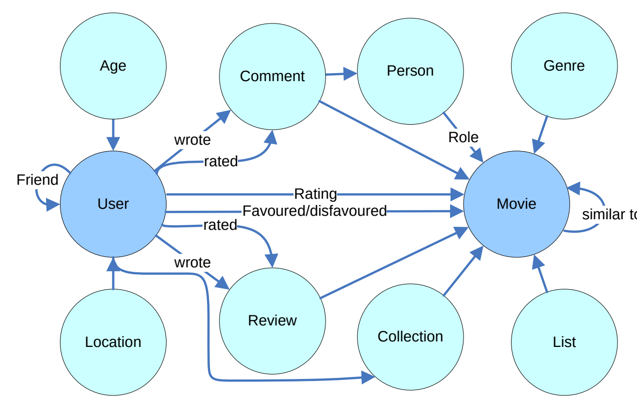


Figure 3: An abstract entity-relationship diagram of the Filmtipset dataset.

was used for both Moviepilot tracks. Participants were initially provided with the training and test sets, the evaluation dataset was withheld until the final evaluation process. The users in each of the test and evaluation sets were not overlapping between the datasets. Similarly to the weekly datasets, the mood recommendations track dataset also consisted of two track-specific rating data files and one training file.

4.2 Filmtipset

Filmtipset is Sweden's largest movie recommendation community. It has over 90,000 registered users and a database of more than 20 million ratings. The Filmtipset datasets used for CAMRa were extracted from a snapshot of the dataset taken in March 2010. The datasets were divided into several files, each track having a training dataset, a test set, and an evaluation set. This rating data was bundled with other features, such as comments, friend relations, actor/writer/director details, and details. An abstract entity-relationship diagram of the datasets and their features is shown in Figure 3. Table 3 shows some of the non rating related numbers in the dataset.

The training set for the weekly recommendations track was based on all ratings provided between February 1st 2008 and February 25th 2010. The test and evaluation sets for the Christmas week sub-track were based on ratings pro-

	Weekly	Social
Collection	307,131	102
Favorites	44,765	15,283
Friends	83,966	12,171
Genres	143,316	67,997
Lists	519,515	438,643
Movie comments	289,586	146,510
People in movies	452,074	224,410
Person comments	322,555	2,822
Review ratings	37,491	2,423
Reviews	1,341	1,044
Movie similarities	35,925	28,372

Table 3: The number of collections, favorites, etc. in the Filmtipset datasets.

	Users	Movies	Ratings
Train	34,857	53,600	5,862,464
Xmas Test	2,500	5,110	23,393
Xmas Eval	1,000	3,450	9,250
Oscar Test	2,500	5,670	33,548
Oscar Eval	848	3,235	11,486

Table 4: The number of users, movies and ratings in the Filmtipset weekly dataset.

vided between December 21st 2009 and December 27th 2009. The participants were prohibit from using "future ratings", i.e. training set ratings provided after December 21st. The Oscars week sub-track training set was the same one as in the Christmas track, such that the test and evaluation sets were based on ratings provided between February 27th and March 7th 2010. The details of these datasets are shown in Table 4.The dataset for the social recommendations track was extracted from the Filmtipset dataset. In order to not reveal information sought for in the weekly track, data in this dataset contained ratings provided between April 18th 2000 (the date Filmtipset was launched) and January 31st 2008 (the day before the first rating in the weekly training set). Furthermore, all users found in the Filmtipset weekly dataset were removed from the social recommendations dataset, practically creating two user-wise disjoint datasets. The details of these datasets are shown in Table 5.

	Users	Movies	Ratings
Train	16,473	24,222	3,075,346
Test	439	1,915	15,729
Eval	153	1,449	6,224

Table 5: The number of users, movies and ratings in the Filmtipset social dataset.

5. OUTCOMES

The challenge attracted over 40 participating teams from 20 countries. The primary outcomes of the challenge were presented at the event which took place on September 30th 2010 in Barcelona. Ten papers were accepted for the proceedings, five long and five short. Out of the 40+ teams that were participating, twelve submitted papers to the challenge, whereas another 20 sent notifications they were still working on their approaches and would submit papers to upcoming conferences and journals. The ten accepted papers covered all tracks and sub tracks. Both weekly tracks and datasets were covered by [10] where the authors implemented a time-aware collaborative filtering model using matrix factorization. The weekly track using the Moviepilot data was covered by [7] where the authors used an approach from tag recommendation, Pairwise Interaction Tensor Factorization (PITF) where weeks were used to form user-movie-weeks tensors. Two attempts at the weekly track using the Filmtipset data were presented by the authors of [4, 5], the papers presented a time-based kNN recommender [5] and a regression models-based approach [4].
The mood track was covered by [13–15] where the approaches used were a mood and user-based hybrid kNN weighted mean [14], a extended matrix factorization model that included mood information [13] and a nearest-neighbor collaborative filtering algorithm utilizing expert users [15].
Finally the social track was covered by the authors of [6, 9–11]. The social approach in [10] was, similarly to the weekly approach covered by this paper, based on matrix factorization. The approach in [6] was a random-walk model utilizing the implicit information in friendships, [9] presented and extension of traditional collaborative filtering where social data was taken into consideration, and [11] presented two approaches: a kNN approach based on linear combinations of similarity measures between users, and one approach based on inductive logic programming. The results of each approach are presented in Table 6.

	paper	P@5	P@10	MAP	AUC
MP_{Xmas}	$[10]_{TWMF}$	0.3637	0.3168	0.1654	0.9212
	[7]	0.1418	0.1281		0.9680
MP_{Oscar}	$[10]_{TWN}$	0.2775	0.2237	0.1362	0.9556
	[7]	0.2039	0.1822		0.9623
FT_{Xmas}	$[10]_{SMF}$	0.0817	0.0596	0.0902	0.9283
	[5]	0.0070	0.0044	0.0405	0.4552
	[4]	0.0795	0.0821	0.0973	0.9231
FT_{Oscar}	$[10]_{SMF}$	0.1087	0.0708	0.0911	0.9467
	[5]	0.0034	0.0028	0.0359	0.4161
	[4]	0.0942	0.0655	0.0849	0.9295
Mood	[13]	0.3380	0.2970	0.2940	0.8690
	[15]	0	0	0.0037	0.6548
Social	$[10]_{NRMF}$	0.5144	0.4185	0.3103	0.9782
	[6]	0.0802	0.0704	0.0596	0.4276
	[9]			0.4167	
	[11]	0.1480	0.1230	0.0970	0.9880

Table 6: The results obtained in each track by the participants (missing values were not provided).

6. REFERENCES

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin, 'Incorporating contextual information in recommender systems using a multidimensional approach', *ACM Trans. Inf. Syst.*, **23**(1), 103–145, (2005).

[2] Sarabjot S. Anand and Bamshad Mobasher,

'Contextual recommendation', in *WebMine*, pp. 142–160, (2006).

[3] Shlomo Berkovsky, Tsvi Kuflik, Lora Aroyo, Dominik Heckmann, Alexander Kröner, Francesco Ricci, and Geert-Jan Houben, 'Predicting user experiences through cross-context reasoning', in *LWA*, pp. 27–31, (2006).

[4] Antoine Brenner, Bruno Pradel, Nicolas Usunier, and Patrick Gallinari, 'Predicting most rated items in weekly recommendation with temporal regression', in *Proc. of CAMRa'10*. ACM, (2010).

[5] Pedro G. Campos, Alejandro Bellogín, Fernando Díez, and J. Enrique Chavarriaga, 'Simple time-biased knn-based recommendations', in *Proc. of CAMRa'10*. ACM, (2010).

[6] Fernando Díez, Enrique Chavarriaga, Pedro G. Campos, and Alejandro Bellogín, 'Movie recommendations based in explicit and implicit features extracted from the filmtipset dataset', in *Proc. of CAMRa'10*. ACM, (2010).

[7] Zeno Gantner, Steffen Rendle, and Lars Schmidt-Thieme, 'Factorization models for context-/time-aware movie recommendations', in *Proc. of CAMRa'10*. ACM, (2010).

[8] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Trans. Inf. Syst.*, **22**(1), 5–53, (2004).

[9] Bin Liu and Zheng Yuan, 'Incorporating social networks and user opinions for collaborative recommendation: Modified cf and local trust network based methods', in *Proc. of CAMRa'10*. ACM, (2010).

[10] Nathan N. Liu, Bin Cao, Min Zhao, and Qiang Yang, 'Adapting neighborhood and matrix factorization models for context aware recommendation', in *Proc. of CAMRa'10*. ACM, (2010).

[11] Hossein Rahmani, Beau Piccart, Daan Fierens, and Hendrik Blockeel, 'Two complementary approaches to context aware movie recommendation', in *Proc. of CAMRa'10*. ACM, (2010).

[12] Alan Said, Ernesto W. De Luca, and Sahin Albayrak, 'How social relationships affect user similarities', in *SRS Workshop*, (2010).

[13] Yue Shi, Martha Larson, and Alan Hanjalic, 'Mining mood-specific movie similarity with matrix factorization for context-aware recommendation', in *Proc. of CAMRa'10*. ACM, (2010).

[14] Licai Wang, Xiangwu Meng, Yujie Zhang, and Yancui Shi, 'A new approach to mood-based hybrid collaborative filtering', in *Proc. of CAMRa'10*. ACM, (2010).

[15] Pei Wu, Weiping Liu, and Cihang Jin, 'A novel recommender system fusing the opinions from experts and ordinary people', in *Proc. of CAMRa'10*. ACM, (2010).