The challenge and workshop on Context-Aware Movie Recommendation (CAMRa2010) were conducted jointly in 2010 with the Recommender Systems conference. The challenge focused on three context-aware recommendation scenarios: 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 accuracy of recommendations. The datasets contained contextual features, such as tags, annotation, social relationships, and comments, normally not available in public recommendation datasets. More than 40 teams from 21 countries participated in the challenge. Their participation was summarized by 10 papers published by the workshop, which have been extended and revised for this special section. In this preface we overview the challenge datasets, tasks, evaluation metrics, and the obtained outcomes.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Retrieval models, Information filtering, Selection process

General Terms: Algorithms, Design, Experimentation, Human Factors

Additional Key Words and Phrases: Recommender systems, context-awareness, context-aware recommendations, social network analysis, user modeling, datasets.

ACM Reference Format:
DOI = 10.1145/2414425.2414438 http://doi.acm.org/10.1145/2414425.2414438

1. INTRODUCTION

Recommender systems have been studied extensively for more than 20 years [Ricci et al. 2011]. The first Content-Based (CB) systems used the textual content of documents to compute inter-document similarity and the similarity scores were used to recommend other relevant documents. CB systems were followed by more elaborate ways of finding relevant items using Collaborative Filtering (CF). CF computes the usage-based similarity of users, that is, the degree to which the similarity between users correlates to the number of common items with which they have interacted.

With the constantly growing volume of available information, CB and CF techniques have reached levels where improvement in the accuracy and quality of
recommendations provided to users becomes increasingly difficult. An example of this could be observed at the Netflix Prize competition,\(^1\) which aimed at decreasing the predictive error of the deployed recommendation engine by 10%. Despite the hype, monetary prize of $1,000,000, and number of active participants (more than 2,000 teams), it took the winners almost three years to reach this goal.

The recently emerged area of Context-Aware Recommender Systems (CARS) has provided novel research challenges and recommendation opportunities. According to the widely accepted definition, “context is any information that can be used to characterize the situation of entities” [Dey 2001]. Applying this to a recommendation scenario, context can be considered as any feature that affects a user’s or an item’s situation, such as the time of day, day of the week, location, weather, mood, presence of other users, and many others.

The importance of context and contextualized user data for accurate recommendations has been widely recognized [Adomavicius et al. 2005; Berkovsky et al. 2006; Baltrunas et al. 2011]. However, the majority of existing techniques focus on recommending the most relevant items and take the context into consideration only to a limited extent. This results in somewhat static recommendations, which can change only after a substantial amount of user interaction with the system. Hence, the recommendations a user is presented with are often not reflective of real-time contextual conditions. Although these recommendations can be reasonably accurate, there is room for improvement by incorporating contextual factors, the problem that was addressed by the 2010 challenge on Context-Aware Movie Recommendation\(^2\) (CAMRa2010).

CAMRa2010 was not the first challenge dealing with recommender systems. A number of related challenges were held over the last decade, including recurring events, such as the ECML/PKDD Discovery Challenge\(^3\) (focused on recommendations in folksonomies in 2008 and 2009) or the KDD Cup\(^4\) (focused on movie recommendations in 2007 and on music recommendations in 2011), and one-time events, such as the already mentioned Netflix Prize competition. Traditionally, the challenges introduced a new dataset and defined specific tasks to be addressed by the participants. The newly released datasets boosted research in recommender systems and has been subsequently used in countless publications beyond the scope of the challenges.

Following the preceding challenges, the CAMRa2010 [Said et al. 2010] challenge was organized. The goal of the challenge was to increase research of context-awareness in recommender systems and bring together industrial and academic researchers from various backgrounds to tackle practical challenges faced by online movie recommender systems. Two datasets gathered by the Moviepilot\(^5\) and Filmtipset\(^6\) movie recommendation websites were released exclusively for CAMRa2010. These were used for three recommendation tasks (described in detail in 2.1):

— recommending movies for a week associated with special events – the weekly track;
— recommending movies based on implied user mood – the mood track; and
— recommending movies based on social relations between users – the social track.

The participants were requested to evaluate the performance of their solutions using the Mean Average Precision (MAP), Precision@K, and Area Under Curve (AUC) metrics [Herlocker et al. 2004]. In this homogeneous environment, the participants were able to

\(^1\)http://www.netflixprize.com/.
\(^2\)http://www.dai-labor.de/camra2010/.
\(^3\)http://www.ecmlpkdd.org/.
\(^4\)http://kddcup.yahoo.com/.
\(^5\)http://www.moviepilot.de.
\(^6\)http://www.filmtipset.se.
identify important aspects in context-aware recommendations, share and merge their solutions, and compare the obtained results.

This special issue features the extended versions of three selected papers from CAMRa2010. The authors had more time to improve their solutions, to evaluate additional aspects of context-aware recommendations, and to discuss the outcomes of their solutions in more detail.

2. THE CAMRA CHALLENGE

In this section, we present the CAMRa2010 challenge. We first overview the recommendation tasks and tracks of the challenge. Then, we discuss the datasets released for the challenge and the evaluation metric. Finally, we summarize the results obtained by the participating teams.

2.1. Tracks

The challenge was divided into three tracks: the weekly track, the mood track, and the social track. Each track focused on a different dimension of context and was provided with a track-specific dataset. The weekly track was divided into two subtracks, each based on a subset of the Moviepilot or Filmtipset dataset. The mood and social tracks were dataset-specific and focused on explicit features available in, respectively, the Moviepilot or the Filmtipset datasets (see Section 2.2).

2.1.1. Weekly Recommendation Track. The weekly recommendation track addressed the temporal dimension of context. 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 2010 Oscar ceremony (Oscar). The rationale for this track implies that events taking place on a given week can potentially influence user preferences and movies they watch. For example, during Christmas week, there is a higher number of children- and family-oriented movies, while during the week leading up to the Oscars ceremony, higher attention is paid to the nominated movies and past Oscar winners.

The participants of this track were requested to recommend a set of movies for each user for these weeks. The users in the datasets were stripped off a randomly selected subset of movies they rated over the course of these weeks. Since the track was subdivided into the Moviepilot and Filmtipset subtracks that contained different information, different features could be used in the recommendation process and the performance of the proposed solutions could vary across the subtracks.

2.1.2. Moviepilot Mood Track. The mood track addressed the contextual dimension related to a user’s mood. The Moviepilot dataset contains hierarchically organized features related to the movie mood. The rationale for this track implies that the mood of a movie could implicitly reflect the mood of a user at the time she watches it. This information could be exploited to generate mood-based recommendations for other conditions, when a similar mood is conjectured.

The participants of this track were requested to recommend a set of movies with a given mood for a selection of users. The selected mood for this track was “Eigenwillig”, that is, “weird” or “unconventional”. However, the participants were not provided with the actual mood, rather with its anonymized code. The users in the dataset were stripped off some of their watched movies labeled with the “Eigenwillig” mood.

2.1.3. Filmtipset Social Track. The social track addressed the social dimension of context, that is, social relationships formed between users of the Filmtipset website. Many Filmtipset users contribute to the social network offered by the service and befriend each other asymmetrically, similar to the follow relation in Twitter. The rationale for
Table I. Statistics of Ratings in the Datasets

<table>
<thead>
<tr>
<th></th>
<th>Users$_{mp}$</th>
<th>Movies$_{mp}$</th>
<th>Ratings$_{mp}$</th>
<th>Users$_{ft}$</th>
<th>Movies$_{ft}$</th>
<th>Ratings$_{ft}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>105,137</td>
<td>25,958</td>
<td>4,544,409</td>
<td>34,857</td>
<td>53,600</td>
<td>5,862,464</td>
</tr>
<tr>
<td>Xmas Test</td>
<td>160</td>
<td>3,377</td>
<td>7,114</td>
<td>2,500</td>
<td>5,110</td>
<td>23,393</td>
</tr>
<tr>
<td>Xmas Eval</td>
<td>80</td>
<td>2,153</td>
<td>2,953</td>
<td>1,000</td>
<td>3,450</td>
<td>9,250</td>
</tr>
<tr>
<td>Oscar Test</td>
<td>160</td>
<td>2,144</td>
<td>7,583</td>
<td>2,500</td>
<td>5,670</td>
<td>33,548</td>
</tr>
<tr>
<td>Oscar Eval</td>
<td>80</td>
<td>1,520</td>
<td>4,098</td>
<td>848</td>
<td>2,325</td>
<td>11,486</td>
</tr>
<tr>
<td>Train</td>
<td>105,137</td>
<td>25,958</td>
<td>4,544,409</td>
<td>16,473</td>
<td>24,222</td>
<td>3,075,346</td>
</tr>
<tr>
<td>Test</td>
<td>160</td>
<td>251</td>
<td>2,656</td>
<td>439</td>
<td>1,915</td>
<td>15,729</td>
</tr>
<tr>
<td>Eval</td>
<td>80</td>
<td>220</td>
<td>1,421</td>
<td>153</td>
<td>1,449</td>
<td>6,224</td>
</tr>
</tbody>
</table>

this track implies that online friendship reflects users’ implicit similarity and can influence user preferences for movies.

The users in the dataset were stripped off some of the movies they had watched shortly after at least three of their friends had watched the same movie. The participants of the track were requested to recommend a set of movies for a selection of users for whom this social influence was identified.

2.2. Datasets

Similar to many of the mentioned challenges, CAMRa2010 was centered around new real-life datasets released by Moviepilot and Filmtipset. Due to potential privacy breaches, the datasets were anonymized prior to the challenge, and the participants were prohibited from using external information sources. This allowed us to compare the accuracy of the proposed solutions in a controlled environment and to identify algorithmic aspects appropriate to specific recommendation tasks.

Different datasets were created for each track and subtrack: two Moviepilot datasets (for the weekly track and mood track) and two Filmtipset datasets (for the weekly track and the social track). Each of these datasets was bundled with a separate test set created using randomly seeded models, and the participants used the test sets to evaluate the performance of their solutions. In addition, four evaluation sets were withheld and released to the participants at the end of the challenge. The results obtained using the evaluation sets allowed us to identify the winners of the challenge.

2.2.1. Moviepilot. Moviepilot is the leading online movie recommendation community in Germany. At the time of CAMRa2010, it contained over 100,000 users and roughly 7.5 million ratings. The Moviepilot datasets for CAMRa2010 were extracted from a snapshot taken in April 2010 and were based on ratings provided between January 2008 and March 2010. Ratings referring to the target weeks of the weekly track (Christmas 2009 and Oscars 2010) were removed from the test and evaluation sets in order not to reveal information unavailable at the time of the recommendations.

All the extracted ratings were split into seven datasets: the same training set used for both tracks, test set for Christmas week, evaluation set for Christmas week, test set for the Oscar week, evaluation set for the Oscar week, test set for the mood track, and evaluation set for the mood track. The number of movies, users, and ratings in the Moviepilot datasets is shown in the left part of Table I. The top-left part of Table I refers to the weekly track and the bottom-left to the mood track.

In addition to ratings, the Moviepilot datasets included several other features, such as plot tags, mood tags, audience tags, lists of favorite/hated movies, and place/time of watching. An abstract entity-relationship diagram of the Moviepilot dataset is shown in the left part of Figure 1. The tag assignment statistics in the dataset are shown in the left part of Table II.

2.2.2. Filmtipset. Filmtipset is Sweden’s largest movie recommendation community. At the time of CAMRa2010, it contained over 90,000 users and more than 20 million
ratings. The Filmtipset datasets for CAMRa2010 were extracted from a snapshot taken in March 2010 and were based on ratings provided between February 2008 and February 2010. Again, ratings referring to the target weeks of the weekly track were removed from the test and evaluation sets.

All the extracted ratings were split into eight datasets: the training set for the weekly track, test set for Christmas week, evaluation set for Christmas week, test set for Oscar week, evaluation set for Oscars week, training set for the social track, test set for the social track, and evaluation set for the social track. The number of movies, users, and ratings in the Filmtipset datasets is shown in the right part of Table I. Again, the top-right part of Table I refers to the weekly track and the bottom-right to the social track.

The ratings were bundled with other features, such as movie comments, friend relations in the underlying social network, actor/writer/director details, lists of preferred movies, and others. An abstract entity-relationship diagram of the Filmtipset dataset is shown in the right part of Figure 1. The statistics of other features in the dataset are shown in the right part of Table II. All users from the Filmtipset weekly dataset were removed from the social recommendations dataset.

### 2.3. Evaluation

In order to identify the winners of each track, the evaluation included several facets. First, the participants were requested to address a predefined set of evaluation metrics and specify the performance of their solutions with respect to these metrics. Second, the workshop submissions were reviewed by the members of the program committee. Third, the quality of the proposed algorithms was subjectively assessed by the members of the expert panel.
2.3.1. Evaluation Metrics. The participants were requested to use the following set of evaluation metrics in order to provide comparable results.

— **Precision@5** - One of the widely-used measures in information retrieval and recommender systems. It communicates the classification accuracy of a short list of five recommended items.

— **Precision@10** - Similar to Precision@5, is used very often. It communicates the classification accuracy of a longer list of ten items, roughly referring to the “first page” of recommendations.

— **Mean Average Precision (MAP)** - Standard metric used in the TREC community. It provides a single-figure measure of accuracy across various sizes of the list of recommended items.

— **Area Under the Curve (AUC)** reflects the overall ability of a recommender system to differentiate between relevant and irrelevant items.

Formal definitions and elaborate discussion on these and many other evaluation metrics used in recommender systems can be found in Herlocker et al. [2004].

2.3.2. Subjective Evaluation Criteria. Members of the expert panel were asked to evaluate the algorithmic summaries of the proposed solutions according to their impression and subjective “gut feeling”. The following evaluation criteria were used.

— **Context.** Does the algorithm use the contextual (temporal/mood/social) data?

— **Contextualization of recommendations.** Does the algorithm provide context-dependent recommendations?

— **Extensibility.** Is the algorithm applicable to other contextual variables?

— **Serendipity.** Does the algorithm provide serendipitous recommendations?

— **Creativity.** Is the algorithm creative/innovative?

— **Scalability.** Can the algorithm scale up to millions of users/items?

— **Sparsity.** Does the algorithm perform well for sparse/incomplete data?

— **Domain dependence.** Is the algorithm configurable to other domains/constraints?

— **Adoptability.** Do you see the algorithm adopted by real recommender systems?

2.4. Results

The challenge attracted more than 40 participating teams from 21 countries. Out of these teams, twelve submitted papers to the workshop. Ten papers (five long and five short) covering all tracks and subtracks were accepted for publication and presented at the workshop, which was held jointly with the 2010 Recommender Systems conference.

2.4.1. Weekly Recommendation Track Results. Both weekly tracks were covered by Liu et al. [2010], where the authors implemented a time-aware CF model using matrix factorization. The Moviepilot weekly track was covered by Gantner et al. [2010], where the authors used Pairwise Interaction Tensor Factorization, an approach from tag recommendations that used the weeks to form user-movie-weeks tensors. Two attempts at the Filmtipset weekly track developed a time-based kNN recommender [Campos et al. 2010] and a regression model-based approach Brenner et al. [2010]. The winner of the track was Gantner et al. [2010], which showed that recommending movies that were popular in the ten days before the Oscar ceremony is preferable over using item-based CF on the full dataset.

2.4.2. Moviepilot Mood Track Results. The mood track was covered by three teams. The proposed solutions to this track involved a mood and user-based hybrid kNN weighted mean [Wang et al. 2010], extended matrix factorization model including mood information [Shi et al. 2010], and a kNN CF algorithm utilizing the expertise of users [Wu et al. 2010]. The winner of the track was Shi et al. [2010], which showed that using one
Table III. Summary of the Obtained Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>Team</th>
<th>P@5</th>
<th>P@10</th>
<th>MAP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPXmas</td>
<td></td>
<td>Liu et al. 2010</td>
<td>0.3637</td>
<td>0.3188</td>
<td>0.1854</td>
<td>0.9212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gantner et al. 2010</td>
<td>0.1418</td>
<td>0.1281</td>
<td>0.9680</td>
<td></td>
</tr>
<tr>
<td>MPOscar</td>
<td></td>
<td>Liu et al. 2010</td>
<td>0.2775</td>
<td>0.2237</td>
<td>0.1362</td>
<td>0.9556</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gantner et al. 2010</td>
<td>0.2039</td>
<td>0.1822</td>
<td>0.9623</td>
<td></td>
</tr>
<tr>
<td>FTXmas</td>
<td></td>
<td>Liu et al. 2010</td>
<td>0.0917</td>
<td>0.0526</td>
<td>0.0902</td>
<td>0.9233</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Campos et al. 2010</td>
<td>0.0070</td>
<td>0.0044</td>
<td>0.0405</td>
<td>0.4552</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brenner et al. 2010</td>
<td>0.0795</td>
<td>0.0821</td>
<td>0.0973</td>
<td>0.9231</td>
</tr>
<tr>
<td>FTOscar</td>
<td></td>
<td>Liu et al. 2010</td>
<td>0.1087</td>
<td>0.0708</td>
<td>0.0911</td>
<td>0.9467</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Campos et al. 2010</td>
<td>0.0034</td>
<td>0.0028</td>
<td>0.0359</td>
<td>0.4161</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brenner et al. 2010</td>
<td>0.0942</td>
<td>0.0655</td>
<td>0.0849</td>
<td>0.9295</td>
</tr>
<tr>
<td>Mood</td>
<td></td>
<td>Shi et al. 2010</td>
<td>0.3390</td>
<td>0.2970</td>
<td>0.2840</td>
<td>0.8690</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wu et al. 2010</td>
<td>0.0037</td>
<td>0.0037</td>
<td>0.0658</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td></td>
<td>Liu et al. 2010</td>
<td>0.5144</td>
<td>0.4185</td>
<td>0.3103</td>
<td>0.9782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diez et al. 2010</td>
<td>0.0802</td>
<td>0.0704</td>
<td>0.0596</td>
<td>0.4276</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Liu and Yuan 2010</td>
<td>0.4167</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rahmani et al. 2010</td>
<td>0.1480</td>
<td>0.1230</td>
<td>0.0970</td>
<td>0.9880</td>
</tr>
</tbody>
</table>

Table IV. Precision of Baseline Context-Ignorant Recommendations

<table>
<thead>
<tr>
<th>Method</th>
<th>Metric</th>
<th>Team</th>
<th>P@5</th>
<th>P@10</th>
<th>Mood</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPXmas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0925</td>
<td>0.0513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPOscar</td>
<td></td>
<td></td>
<td>0.0919</td>
<td>0.0494</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTXmas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0080</td>
<td>0.0073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTOscar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0073</td>
<td>0.0074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1119</td>
<td>0.1316</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

specific mood tag is preferable over using general mood tags. It was demonstrated that it is particularly important to learn latent movie features with respect to the specified mood.

2.4.3. Filmtipset Social Track Results. The social track was covered by four teams. The social approach in Liu et al. [2010] was, similar to the weekly approach covered by this article, based on matrix factorization. The approach in Diez et al. [2010] was a random-walk model utilizing the implicit information in friendships. In Liu and Yuan [2010], the authors presented an extension of traditional CF, where social data was taken into consideration. Rahmani et al. [2010] presented two solutions: a kNN approach based on linear combinations of user similarity and an approach based on inductive logic programming. The winner of the track was Liu et al. [2010], which showed that incorporating the social network of users into the matrix factorization model increases the accuracy of recommendations.

2.4.4. Comparison. The results obtained by each approach with respect to each of the predefined metrics are summarized in Table III. The missing values were not reported by the participants. As a comparison baseline, we provide in Table IV the precision scores obtained by a standard context-ignorant kNN CF approach using the same evaluation set as the participants of the challenge.

3. SUMMARY

CAMRa2010 was concluded by a one-day workshop, but the datasets and problems of the challenge continue to be investigated, both by the participants of the challenge and others. The released datasets were subsequently used in multiple works, some published beyond the recommender systems community. Furthermore, the second edition of the challenge and workshop, CAMRa2011, was organized jointly with the 2011 Recommender Systems conference and focused on group related contextual aspects [Said et al. 2011].

This special section features the extended versions of selected articles presented at CAMRa2010. The participants of CAMRa2010 were encouraged to continue their work on the released datasets and submit extended versions of their papers for the special
issue. Six articles were submitted and reviewed by distinguished researchers in the field (special 3 to 4 rounds of reviews). We are pleased to present in the current section of ACM Transactions on Intelligent Systems and Technology three high-quality articles that were eventually accepted for publication.

—“An Empirical Comparison of Social, Collaborative Filtering, and Hybrid Recommenders” by Alejandro Bellogin, Ivan Cantador, Fernando Diez, Pablo Castells, and Enrique Chavarriaga presents an evaluation of how social aspects affect recommendation in datasets, where not all users have social ties, and compares the performance with several standard and hybrid recommendation algorithms in social and regular settings.

—“Social Temporal Collaborative Ranking for Context Aware Movie Recommendation” by Nathan Liu, Luheng He, and Min Zhao presents a novel ranking-based matrix factorization model, a time-aware matrix factorization model, and a network regularization method to be used in a social recommendation scenario.

—“Mining Contextual Movie Similarity with Matrix Factorization for Context-Aware Recommendation” by Yue Shi, Martha Larson, and Alan Hanjalic presents a novel joint matrix factorization model for factorizing multiple matrices. The model is then applied as a hybrid recommender in the mood-based recommendation scenario.

Collectively, these three articles represent the state-of-the-art developments and emerging trends in context-aware movie recommender systems. We hope that the readers will find the approaches and ideas described in these articles interesting and informative and that the articles will contribute to the body of knowledge in the area of context-aware recommendations.

REFERENCES


