ABSTRACT
We propose a method for estimating the expected economic value of multi-dimensional data sets in recommender systems and illustrate the proposed approach using a unique data set combining implicit and explicit ratings with rich content as well as spatio-temporal contextual dimensions and social network data.

Categories and Subject Descriptors
E.0 [Data]: General; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords: Business Value, Context, Dataset

1. INTRODUCTION
Although collaborative filtering (CF) recommender systems (RSes) have been very successful during the last decades, they have certain limitations; traditional RSes operate in the two-dimensional User × Item space and do not take into consideration additional contextual information, such as time and location, that may be crucial in many applications. At the same time, data related to social networks and other informative dimensions is widely available nowadays but it usually comes at significant monetary cost and/or engineering effort. Hence, data should be treated as an investment and the expected costs and benefits of acquiring and using it should be carefully considered and evaluated.

In this paper, we illustrate how we can estimate the expected economic value (gain or loss) of such multi-dimensional data sets or specific sets of features can lead to better and more profitable managerial decisions through more informed and data-driven decision making in the future. Besides, the proposed approach can be used to derive even more useful evaluation metrics in the field of RSes.

In the rest of the paper, we first use the matrix factorization framework to show how various dimensions can be incorporated into a single model for recommendations and then discuss how the added predictive power of the inducted model translates into monetary value for businesses. Then, we introduce a novel multi-dimensional data set and illustrate the aforementioned approach. Due to space limitations, we focus on the task of item prediction; this method can be extended to rating prediction as well.

2. MODEL
We build a (hybrid) model incorporating the extra information of temporal, social and location dynamics as well as the content of items, using a feature-based factorization model [2]. In particular, the prediction score \( \hat{y}_{u,i} \) is modeled as:

\[
\hat{y}_{u,i} = \mu + \left( \sum_{g \in G} \gamma_g b_g + \sum_{m \in M} \alpha_m b_m^u + \sum_{n \in N} \beta_n b_n^i \right) + \left( \sum_{m \in M} \alpha_m p_m \right)^T \left( \sum_{n \in N} \beta_n q_n \right)
\]

where \( \mu \) is the base score of the predictions, \( G, M, N \) the indexes of global features, user features, and item features, respectively, \( \gamma, \alpha, \beta \) the corresponding feature vectors, and \( \gamma_g, \alpha_m, \beta_n \) the feature values. In the specific example presented in the rest of this paper, the global features include the location and temporal information (context) of the rating events, the item features the content information of the items, and the user features the social network information of the users (see Section 5). In addition, a vector of latent factors is included as well. The model can be further extended in order to incorporate social relationships of the users or other relevant information.

To estimate the model, we use the logistic function as activation function and the negative log-likelihood as loss function:

\[
\text{Loss} = \sum_{u,i} \left( r_{u,i} \ln f(\hat{y}_{u,i}) - (1-r_{u,i}) \ln(1-f(\hat{y}_{u,i})) \right) + \text{regularization},
\]

where \( f(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}} \) and \( r_{u,i} \in \{0, 1\} \) the true rating.

3. DATA
Similar to [3], we construct a new data set, titled “ConcertTweets”, based on publicly available and well-structured tweets referring to music concerts [1]. This data set is collected and analyzed in real time using the Twitter streaming API. We decided to collect, use, and release this data set because it contains rich feature dimensions as well as novel and relevant activity from a domain of significant academic and business interest. As of June 2014, this data set contains information on 30,178 distinct Twitter users and 100,000 personal ratings, both implicit and explicit, referring to more than 50,000 concerts of 13,578 music artists and bands.

The unique characteristics of our data set allow reconciling it and linking it to popular databases leveraging rich semantic information, such as the musical genres of the artists. Besides, both the geolocation information of the concert and the user (as publicly disclosed based on the application set-
Table 1: Cost/Benefit matrix.

<table>
<thead>
<tr>
<th>Recommended (R)</th>
<th>Used (U)</th>
<th>Not Used (NU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Recommended (NR)</td>
<td>b(R,U)</td>
<td>c(NR,NU)</td>
</tr>
<tr>
<td></td>
<td>c(R,U)</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Evaluation results.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Accuracy $E[\Delta Value]$ per user-item pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>0.7548</td>
</tr>
<tr>
<td>MF + Item Content</td>
<td>0.7567</td>
</tr>
<tr>
<td>MF + User Social Network data</td>
<td>0.7767</td>
</tr>
<tr>
<td>MF + Location-based features</td>
<td>0.8819</td>
</tr>
<tr>
<td>MF + Temporal features</td>
<td>0.7667</td>
</tr>
<tr>
<td>MF + All features</td>
<td>0.8826</td>
</tr>
</tbody>
</table>

In this paper, we propose a method for estimating the expected economic value of multi-dimensional data sets in RSes and illustrate the proposed approach using a unique data set combining implicit and explicit ratings with rich content, spatio-temporal contextual dimensions, and social network profiles. This approach can lead to better and more profitable managerial decisions as well as more useful evaluation metrics. As part of the future work, we plan to extend the proposed approach to the task of rating prediction as well as estimate the value of different dimensions in various recommendation domains and settings.

6. CONCLUSIONS

In this paper, we propose a method for estimating the expected economic value of multi-dimensional data sets in RSes and illustrate the proposed approach using a unique data set combining implicit and explicit ratings with rich content, spatio-temporal contextual dimensions, and social network profiles. This approach can lead to better and more profitable managerial decisions as well as more useful evaluation metrics. As part of the future work, we plan to extend the proposed approach to the task of rating prediction as well as estimate the value of different dimensions in various recommendation domains and settings.

7. REFERENCES