Estimating the Value of Multi-Dimensional Data Sets in Context-based Recommender Systems

Panagiotis Adamopoulos and Alexander Tuzhilin
Department of Information, Operations and Management Sciences
Leonard N. Stern School of Business, New York University
{padamopo,atuzhili}@stern.nyu.edu

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 $User \times Item$ 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 and translate the added predictive power into monetary units (such as U.S. dollars). This approach has important implications since determining the expected monetary value of 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.

RecSys 2014 Poster Proceedings, October 6–10, 2014, Foster City, Silicon Valley, USA.

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:

$$\hat{y}_{u,i} = \mu + \left(\sum_{g \in G} \gamma_g b_g + \sum_{m \in M} \alpha_m b^u_m + \sum_{n \in N} \beta_n b^i_n \right) + \left(\sum_{m \in M} \alpha_m \mathbf{p}_m \right)^T \left(\sum_{n \in N} \beta_n \mathbf{q}_n \right)$$

where μ is the base score of the predictions, G, M, N the index sets of global features, user features, and item features, respectively, γ, α, β 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 (i.e., the feature weights b_g,b^u_m,b^i_n and factors $\mathbf{p}_m$, $\mathbf{q}_n$), we use the logistic function as activation function and the negative log-likelihood as loss function:

$$Loss = \sum_{u,i} (r_{u,i} \ln f(\hat{y}_{u,i}) - (1 - r_{u,i}) \ln(1 - f(\hat{y}_{u,i}))) + regularization,$$

where $f(\hat{y}) = \dfrac{1}{1 + e^{-\hat{y}}}$ 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.

	Used (U)	**Not Used** (NU)
Recommended (R)	b(R,U)	c(R,NU)
Not Recommended (NR)	c(NR,U)	0

tings, self-reported by the user, or inferred based on the detailed meta-data about the time zone of the location of the user) are included. Other characteristics of this data set that allow for more thorough and extensive (both offline and online) experimentation are the combination of implicit (i.e., $r_{u,i} \in$ {'Yes', 'Maybe', 'No'}) and explicit (i.e., $r_{u,i} \in \{0.5, 1.0, \ldots, 5.0\}$) ratings, the presence of popular and recent events, and the availability of the timestamp information for both the item (i.e., concert) and the corresponding rating event. In addition, this data set includes information about the social presence of the users (e.g., number of followers, timeline, etc.) and can be easily extended to include their social network. Finally, using the unique Twitter user identifiers, this data set can be further enriched with cross-domain (e.g., books, movies) user activity [4].

4. BUSINESS VALUE

Working within the CF framework of RSes, we assume that data related to implicit and explicit ratings is already available and part of the baseline recommender. Hence, we illustrate how we can estimate the added economic value of data sets related to additional contextual dimensions. We also assume that either the complete data set or an initial representative sample from the additional dimensions is available in order to conduct the initial analysis before the decision to acquire the full data set and / or incorporate it into the production RS. Then, using the cost-benefit information of the business for the specific recommendation task (as in Table 1), we can estimate the expected value of predictions with and without using the additional dimensions. In particular, the added value per instance (i.e., rating tuple) for an additional dimension is estimated as:

$$\begin{aligned}\Delta\text{Value} = {} & p(\text{U}) \times \Delta\text{Recall} \times \text{b(R,U)} \\ & + p(\text{U}) \times (-\Delta\text{Recall}) \times \text{c(NR,U)} \\ & + p(\text{NU}) \times (-\Delta\text{Specificity}) \times \text{c(R,NU)},\end{aligned}$$

where $\Delta\text{Recall} = \text{Recall}_{\text{RS}'} - \text{Recall}_{\text{RS}}$, $-\Delta\text{Specificity} = \text{Specificity}_{\text{RS}} - \text{Specificity}_{\text{RS}'}$, RS the baseline recommender (or "random" predictions) and RS$'$ the recommender with the extended set of contextual dimensions.

Equivalently, for the task of top-N recommendations:

$$\begin{aligned}\Delta\text{Value} = {} & p(\text{U}) \times \Delta\text{Recall@}N \times \text{b(R,U)} \\ & - p(\text{U}) \times \Delta\text{Recall@}N \times \text{c(NR,U)} \\ & - p(\text{NU}) \times \Delta\text{Specificity@}N \times \text{c(R,NU)}.\end{aligned}$$

Similarly, the above approach is extended to the ranking task, using the area under the ROC curve, as well as applications with non-zero benefit for true negatives (i.e., not recommended and not used items) and variable costs.

Given the expected value of the additional dimensions introduced to the RS, we can then estimate whether adding such factors justifies the engineering cost and effort as well as the potential monetary cost of acquiring the data.

5. RESULTS

In the conducted experiments, we consider as positive instances ($r_{u,i} := 1$) all the items with an explicit rating equal to or greater than 4.0 or an implicit rating indicating that the user attended (i.e., labeled as 'Yes') or might attend (i.e., 'Maybe') the event; items with ratings less than 4.0 or events that a user will not attend (i.e., 'No') are consider negatives ($r_{u,i} := 0$). In addition, for each user we randomly select an equal number of non-rated items as negative examples in order to increase the accuracy of our predictions. Moreover, we employ a holdout evaluation scheme with 80/20 random splits into training and test sets without filtering any ratings and we evaluate each model in term of classification tasks based on accuracy. Also, we set the $L2$ regularization parameters at 0.004 and the constant bias for prediction at 0.5. The learning rate for stochastic gradient descent is 0.015.

Table 2: Evaluation results.

Specification	**Accuracy**	**E[ΔValue] per user-item pair**
MF	0.7548	-
MF + Item Content	0.7567	0.3688
MF + User Social Network data	0.7767	4.3084
MF + Location-based features	0.8819	25.4352
MF + Temporal features	0.7667	2.2869
MF + All features	0.8826	25.5928

For the various specifications of the factorization model of Section 2, apart from i) the basic model *(MF)* which includes 128 latent factors, we used ii) the *content* information of the 50 most frequent music genres of the artists as item features, iii) the *social* presence of the users (i.e., number of followers, friends, statuses posted, and tweets favorited) as user features, iv) *spatial* information of the 50 most popular locations and whether the user is located in the same geographical region with the event (locality) as global features, v) the *temporal* information (i.e., 'Friday', 'Saturday', 'Other') of the event again as global features, and vi) an *integrated* model combining all the aforementioned features.

Table 2 shows the experimental results using a cost of 100 units for wrong predictions and zero cost for correct predictions. We see that the various dimensions of this data set have very different monetary values and that the contextual information of location is the most informative dimension in this application offering significant return on investment. Even though the highest accuracy was achieved using the integrated model, the business value should be further considered and compared against the engineering effort and the monetary cost related to additional data.

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

[1] P. Adamopoulos. ConcertTweets: A Multi-Dimensional Data Set for Recommender Systems Research. http://people.stern.nyu.edu/padamopo/data/concertTweets.html.

[2] T. Chen, W. Zhang, Q. Lu, et al. Svdfeature: a toolkit for feature-based collaborative filtering. *JMLR*, 2012.

[3] S. Dooms et al. Movietweetings: a movie rating dataset collected from twitter. In *CrowdRec at RecSys*, 2013.

[4] S. Dooms et al. Mining cross-domain rating datasets from structured data on twitter. In *MSM at WWW*, 2014.