ConcertTweets: A Multi-Dimensional Data Set for Recommender Systems Research

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

ABSTRACT

We present a multi-dimensional data set suitable for recommender systems research. This unique data set combines implicit and explicit user ratings with rich content as well as spatio-temporal contextual dimensions and social network data. The data set can be easily further enriched with additional dimensions and ratings.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors, Measurement

Keywords

Recommender Systems, Context, Dataset, Human Factors

1. INTRODUCTION

Research on recommender systems has strengthened significantly during the last years resulting in pioneering algorithms and scientific achievements that extend beyond the field of recommender systems. As in many other technological domains, one of the main enablers and facilitators of such advancements is the availability of open data sets. A stellar example is the Netflix prize data set [4] and the subsequent enhancements of collaborative filtering algorithms. However, even within the very same communities, other paradigms and perspectives remain relatively underexplored and thus have not yielded similar improvements yet. This is due, partially, to the fact that similar data sets suitable for these paradigms and perspectives are not available to researchers.

In an effort to facilitate (non-commercial) scientific research purposes and further enable collaborations among researchers in the data mining, machine learning, and computer science communities, we publicly release the 'ConcertTweets' data set. The specific data set is suitable for empirical research in the field of recommender systems and has the potential to contribute to significant advancements in less explored areas of research, such as contextual or location-based recommendations.

This data set possesses important characteristics that add significant value and differentiate it from existing data sets. Although during the last decade several user rating data sets have been publicly released and widely used, notably the 1 million rating dataset provided by MovieLens.org [1] and the Netflix Prize data set [4], existing data sets, including various data sets incorporating contextual information and data sets collected from the same source, typically do not contain such rich information. For instance, the data set used in [3] contains contextual information only about the time, location, and companion of the user and the 'MovieTweetings' data set [5] does not contain among others a geolocation dimension or a combination of both implicit and explicit ratings. Some of these unique characteristics are discussed in Section 2.

The 'ConcertTweets' data set is made available at the following web address: `http://stern.nyu.edu/~padamopo/data/`. We plan to continue maintaining this data set and release all future extensions in frequent increments.

2. DATA SET

Similar to [5], we construct a new data set, titled 'ConcertTweets', based on publicly available and well-structured tweets referring to music concerts. 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 user 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. These data correspond to publicly available tweets posted after February, 2014. Table 1 shows the characteristics of the data set. The next version of the 'ConcertTweets' data set, scheduled for release during October 2014, will include $150,000$ user ratings and will be labeled 'ConcertTweets v1.5'.

Aiming at facilitating experimentation, we adopt a file format similar to existing widely used data sets. In particular, the data set consists of the following files:

- `users.dat` Contains the information of the users. The user IDs correspond to the Twitter API user ID. The following information is included: user ID. Additional information can be retrieved using the Twitter API and the provided IDs. In order to obtain additional information, such as the corresponding social network and the description of the user profile and the number of followers, friends, and tweets of each account, corre-

Table 1: Statistical properties of the data set.

Metric	Value
Number of ratings	100,000
Number of unique users	30,178
Number of unique items	13,578
Minimum number of ratings per user	1
Average number of ratings per user	3.313
Standard deviation of number of ratings per user	6.345
Maximum number of ratings per user	477
Minimum number of ratings per item	1
Average number of ratings per item	7.364
Standard deviation of number of ratings per item	21.286
Maximum number of ratings per item	632

sponding to the time of each collected rating, you are encouraged to contact the authors.

- `events.dat` Contains the information of the musical concert. The following information is included: event ID, event date, city, state (or country), latitude, longitude, venue, and event URL.
- `ratings.dat` Contains the concert rating information. Ratings are either explicit, expressed on a scale of 5 (higher values denoting higher appreciation), or implicit, indicating whether a user will attend an event. The following information is included: user ID, band, rating, event ID, venue, event URL, and timestamp.

Complying with both Twitter's and users' rights, we do not redistribute Twitter content, such as datasets of Tweet text and follow relationships, but only Twitter IDs and derivative data. Such content and relationships can be easily retrieved using the Twitter API and the provided IDs (e.g., tweet IDs and user IDs) in the aforementioned data set. If you have difficulties retrieving such information or reconciling it with the released data set, you are encouraged to contact the authors of this document.

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 settings, self-reported by the user, or inferred based on the detailed meta-data about the time zone of the location of the user) can be used as in [2]. 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, the hierarchies of the included entities, 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, etc.) and can be easily extended as previously described to include their social network or other information. Moreover, using the Twitter streaming API we target at collecting all the corresponding ratings broadcasted on Twitter. Also, in an effort to avoid any systematic bias, we do not filter any ratings of the users or items (e.g., the ratings of users with fewer ratings than a predefined threshold). Finally, using the unique Twitter user identifiers, this data set can be further enriched with cross-domain (e.g., books, movies) user activity [6].

Examples of research using the specific data set include [2]. In this example, the authors have used the Twitter API in order to collect the additional Twitter content and freebase.com to retrieve semantic information about the artists included in the data set.

3. USAGE

Neither New York University nor any of the researchers involved can guarantee the correctness of the data, its suitability for any particular purpose, or the validity of results based on the use of the data set. The data set may be used for any non-commercial research purposes under the following conditions:

- The user may not state or imply any endorsement from the New York University or the researchers involved.
- The user should acknowledge the use of the data set in publications resulting from the use of the data set. Please use the following reference:
  - Adamopoulos, Panagiotis, and Alexander Tuzhilin. "Estimating the value of multi-dimensional data sets in context-based recommender systems." *Proceedings of the 8th ACM conference on Recommender systems*. ACM, 2014.
- As a courtesy, we would appreciate it if you send us an electronic or paper copy of those publications.
- The user may not redistribute the data without separate permission.
- The user may not use this information for any commercial or revenue-bearing purposes.

If you have any further questions or comments, please contact `padamopo@stern.nyu.edu`.

4. CONCLUSIONS

We present a multi-dimensional data set combining implicit and explicit ratings with rich content, spatio-temporal contextual dimensions, and social network profiles. This unique data set can significantly facilitate scientific research in the domain of recommender systems.

5. REFERENCES

[1] MovieLens Datasets. http://grouplens.org/datasets/movielens/.

[2] P. Adamopoulos and A. Tuzhilin. Estimating the value of multi-dimensional data sets in context-based recommender systems. In *RecSys*, 2014.

[3] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1):103–145, 2005.

[4] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup and Workshop*, 2007.

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

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