MovieTweetings: a Movie Rating Dataset Collected From Twitter

DATASET · JANUARY 2013

CITATIONS
8

READS
273

3 AUTHORS:

Simon Dooms
Ghent University
13 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Toon De Pessemier
Ghent University
46 PUBLICATIONS **141** CITATIONS
SEE PROFILE

Luc Martens
Ghent University
378 PUBLICATIONS **2,620** CITATIONS
SEE PROFILE

MovieTweetings: a Movie Rating Dataset Collected From Twitter

Simon Dooms
iMinds-Ghent University
G. Crommenlaan 8, box 201
Ghent, Belgium
Simon.Dooms@UGent.be

Toon De Pessemier
iMinds-Ghent University
G. Crommenlaan 8, box 201
Ghent, Belgium
Toon.DePessemier@UGent.be

Luc Martens
iMinds-Ghent University
G. Crommenlaan 8, box 201
Ghent, Belgium
Luc1.Martens@UGent.be

ABSTRACT

Public rating datasets, like MovieLens or Netflix, have long been popular and widely used in the recommmender systems domain for experimentation and comparison. More and more however they are becoming outdated and fail to incorporate new and relevant items. In our work, we tap into the vast availability of social media and construct a new movie rating dataset 'MovieTweetings' based on public and well-structured tweets. With currently over 60,000 ratings and the addition of around 500 new ratings per day we believe this dataset can show to be very useful as an always up-to-date and natural rating dataset for movie recommenders.

Categories and Subject Descriptors

E.0 [**Data**]: General; H.1.2 [**Information Systems**]: Models and Principles—*User/Machine Systems, Human Information Processing*

General Terms

Human Factors

Keywords

Dataset, Twitter, MovieTweetings, MovieLens

1. INTRODUCTION

Ratings are used by Recommender systems to learn user preferences and so they are an indispensable component of the recommendation process. Their availability is a requirement for high quality recommendations and so new systems sometimes suffer from cold start issues when they lack sufficient rating data. To jump start these systems, often existing datasets are imported which contain user ratings of other systems in the same item domain [1]. For movie recommenders for example, datasets like MovieLens [2] and Netflix [3] are available and widely used. Especially for research purposes, where multiple recommendation algorithms or methods are compared, public datasets offer a very useful means for evaluation.

The relevance of these datasets however fades over the years as they become outdated (e.g., most recent movie in MovieLens 100K dataset is from 1998). While still useful for offline evaluation, online experiments with actual users may fail because of the lack of recent and relevant movies in the dataset. Moreover, datasets themselves are also often filtered so to only contain users with e.g. a minimum number of ratings (e.g., 20 ratings for MovieLens). Because of this filtering, a systematic bias is introduced which may prevent experimental results to be generalizable to real-life scenarios [4].

Focusing on the movie domain, in this work we propose and publish a new unfiltered movie rating dataset '*MovieTweetings*' which we constructed (and extend on a daily basis) from ratings contained in structured tweets posted on Twitter. Because the ratings originate from social media, they are likely to involve recent and relevant movies and may therefore serve as useful input data for modern online or user-centered evaluation experiments in the recommender systems domain.

2. RATINGS FROM TWITTER

Through Twitter, *tweets* (i.e., short text messages) can be posted and shared among its users. Such tweets are restricted in length to 140 characters and can be annotated by means of user-generated hashtags. The use of these hashtags often facilitates data extraction processes as is the case for this very work.

Since user-generated data is very noisy and hard to objectively interpret, we searched for well-structured tweets that may contain movie ratings ready for extraction. A usable stream of tweets we found through the popular Internet Movie Database (IMDb). This website (owned by Amazon) provides extensive information (e.g., director, cast, genre, plot, etc.) on a very large number of movies. For every movie a separate page details the relevant movie attributes together with some additional options. One of these options is to rate the movie on a 1 to 10 scale and more interestingly, this user rating can be posted to Twitter by means of a *share* button. When the button is clicked, a tweet is auto-generated by the system and proposed to the user in a pre-defined format. An example of such a tweet is the following.

```
'I rated The Matrix 9/10
http://www.imdb.com/title/tt0133093/ #IMDb'
```

As can be noted, this tweet is well-structured and apt for information extraction. The tweet contains the movie title, user rating, and more importantly a link to the relevant IMDb page which unambiguously identifies the rated movie. The same kind of structure applies to auto-generated tweets sourcing from smartphone and tablet apps associated with the IMDb platform.

To harvest the information contained in these tweets we query the Twitter Search API on a daily basis for tweets con-

Metric	Value
Ratings	65,115
Unique Users	12,425
Unique Items	8,458
Sparsity	0.9993
Minimum ratings per user	1
Average ratings per user	5
Maximum ratings per user	308
Minimum ratings per item	1
Average ratings per item	8
Maximum ratings per item	1604
Minimum movie year	1898
Maximum movie year	2013
Minimum ratings per day	46
Average ratings per day	521
Maximum ratings per day	905

Table 1: Some dataset metrics at the time of writing. Since the dataset is updated daily, these may be subject to change.

taining the string '*I rated*' and hashtag '#IMDb'. Through a series of regular expressions, we extract relevant information such as user, movie and rating, and cross-reference this with the according IMDb page to provide also genre metadata.

3. THE DATASET

Starting March 7, 2013 we queried the Twitter search API on a daily basis and at the time of writing more than 60,000 tweets have been collected. We aim for this dataset to be a modern version of the popular MovieLens dataset and so similar file formats and metadata structures were adopted. Our dataset comprises two files: *ratings.dat* and *movies.dat* which respectively store the ratings and movie metadata (i.e., genre information). We adopted an IMDb identifier as item id to facilitate additional metadata enrichment.

Table 1 overviews some of the main characteristics of the MovieTweetings dataset. It contains over 60,000 ratings provided by more than 12,000 users on 8,000 unique items. Since the dataset is gathered from social media, the divergence of the rated items (i.e., movies) is very high, leading to a sparsity value (i.e., ratio of known and all possible ratings in the user-item matrix) of at least 0.9993.

This dataset is unfiltered and therefore unlike MovieLens, where every user has rated at least 20 movies, here the number of ratings per user varies from 1 to 305. Fig 1 shows the distribution of the number of ratings per user. As is the case for many online applications, we note a long tail distribution with most of the users having rated only 1 item. The MovieTweetings dataset is available for download on the GitHub platform[1]. We provide the dataset in two forms: the full dataset which is updated on a daily basis, and snapshots of fixed portions (e.g., 10K ratings, 20K, 50K, etc.) of the dataset to facilitate experimentation and reproducibility of research.

4. CONCLUSIONS

Public rating datasets like MovieLens and Netflix are slowly but surely becoming outdated and they are losing their relevance for online and user-centered experiments. In this work, we present the MovieTweetings dataset which we collect automatically from structured social media posts (i.e., Twitter). Although sparsity values for other datasets (e.g., MovieLens) might be lower, this dataset consists of natural and realistic data, and furthermore incorporates all the most recent and popular movies which can be crucial for the adoption of present-day user-centric evaluation experiments.

[1]http://github.com/sidooms/movietweetings

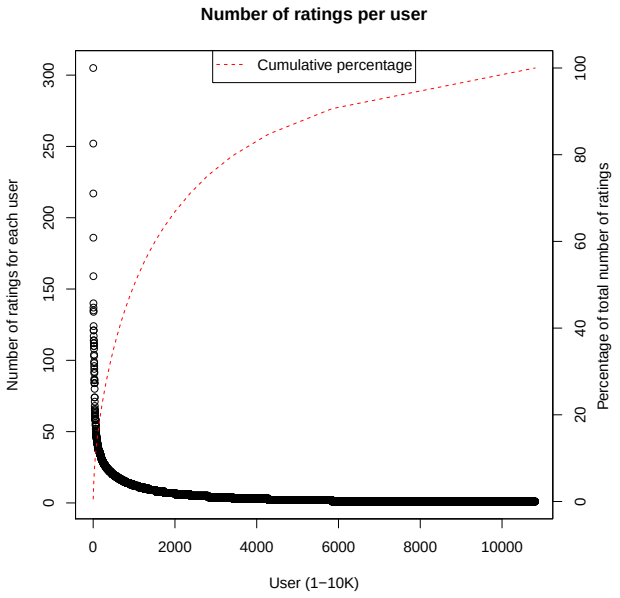

Figure 1: The distribution of the number of ratings per user and the cumulative percentage of the total amount of ratings.

5. ACKNOWLEDGMENTS

The described research activities were funded by a PhD grant to Simon Dooms of the Agency for Innovation by Science and Technology (IWT Vlaanderen). Our thanks goes out to Kris Vanhecke for providing the original idea.

6. REFERENCES

[1] Simon Dooms, Toon De Pessemier, Dieter Verslype, Jelle Nelis, Jonas De Meulenaere, Wendy Van den Broeck, Luc Martens, and Chris Develder. Omus: an optimized multimedia service for the home environment. *Multimedia Tools and Applications*, 2013.

[2] Jonathan L Herlocker, Joseph A Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237. ACM, 1999.

[3] James Bennett and Stan Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007.

[4] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.