Harvesting Movie Ratings from Structured Data in Social Media

Simon Dooms
Ghent University
and
Luc Martens
Ghent University

Public movie rating datasets, like MovieLens or Netflix, have long been popular and widely used in the recommmender systems domain for comparison and experimentation. More and more however these datasets are becoming outdated and fail to incorporate new and relevant items.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. New data is added on a daily basis, which guarantees the MovieTweetings dataset to always incorporate ratings on the newest and most relevant movies.

DOI: 10.1145/2559858.2559862 http://doi.acm.org/10.1145/2559858.2559862

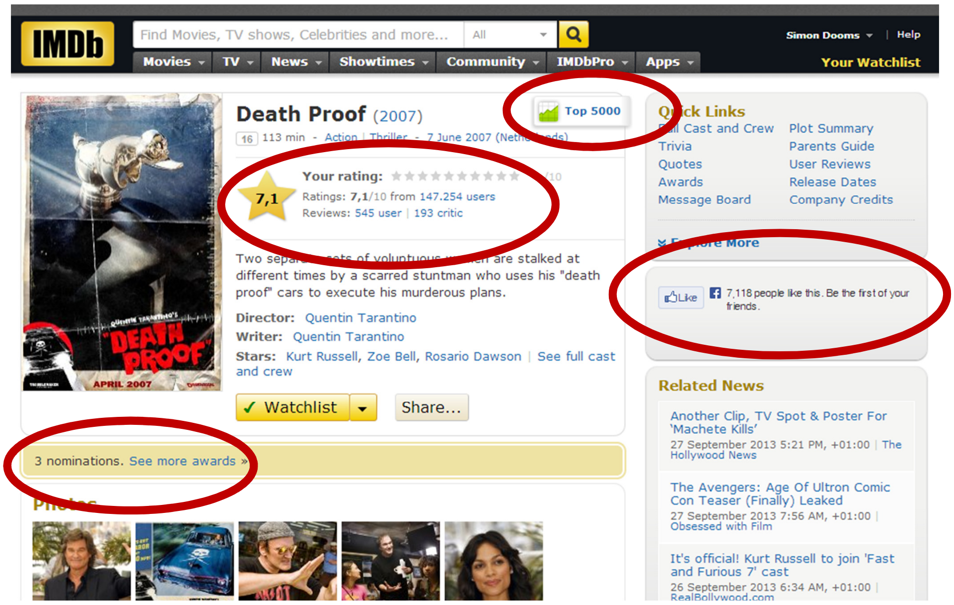
1. INTRODUCTION

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 new systems therefore 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 (e.g. [Dooms et al. 2013]). For movie recommenders for example, datasets like MovieLens [Herlocker et al. 1999] and Netflix [Bennett and Lanning 2007] 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 The relevance of these datasets however fades over the years as they become outdated (e.g., most recent movie in MovieLens 100K dataset is from 1998).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 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 [Shani and Gunawardana 2011].

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 [Dooms et al. 2013]. 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. RATING DATA FROM ONLINE SOURCES

Nowadays, data expressing user preferences can easily be found online. For the movie domain specifically, preferential data is abundantly present on the detailed information pages featured by the popular Internet Movie Database (IMDb). This website (owned by Amazon) provides extensive information (e.g., director, cast, genre, plot, etc.) on a large number of movies. For every movie a separate page details relevant movie attributes amongst other information (Fig. 1).



Death Proof (2007)

16 113 min - Action | Thriller - 7 June 2007 (Netherlands)

Ratings: 7,1/10 from 147.254 users
Reviews: 545 user | 193 critic

Two separate sets of voluptuous women are stalked at different times by a scarred stuntman who uses his "death proof" cars to execute his murderous plans.

Director: Quentin Tarantino
Writer: Quentin Tarantino
Stars: Kurt Russell, Zoe Bell, Rosario Dawson | See full cast and crew

3 nominations. See more awards »

Full Cast and Crew, Plot Summary, Trivia, Parents Guide, Quotes, User Reviews, Awards, Release Dates, Message Board, Company Credits

7,118 people like this. Be the first of your friends.

Another Clip, TV Spot & Poster For 'Machete Kills'
27 September 2013 5:21 PM, +01:00 | The Hollywood News

The Avengers: Age Of Ultron Comic Con Teaser (Finally) Leaked
27 September 2013 7:56 AM, +01:00 | Obsessed with Film

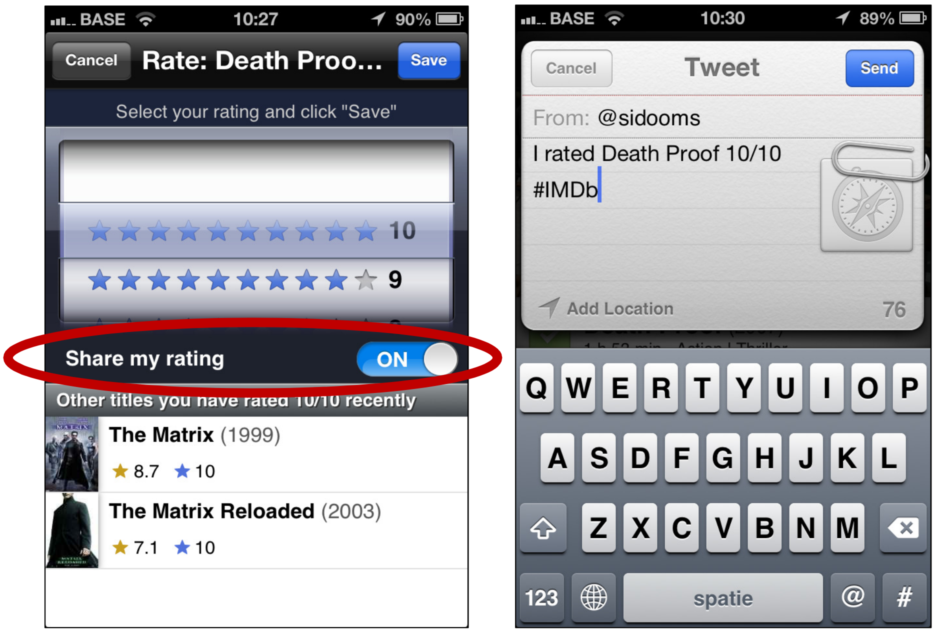
It's official! Kurt Russell to join 'Fast and Furious 7' cast
26 September 2013 6:34 AM, +01:00 | RealBollywood.com

Fig. 1. The detailed information page for a movie on IMDb. Aside from specific movie information, user preference indicators are available.

Aside from the movie information, multiple user preference indicators are available (Fig. 1) such as the IMDb rating, Facebook likes, oscar nominations, etc. Most of these indicators are however unquantifiable or aggregated values, while for recommender systems rating data is often required on a numerical rating scale and expressed on a per user basis.

When we search for preferential movie indicators on social networks as Facebook and Twitter, similar observations can be made. Preferential data is abundantly available in online sources and social networks, but it is extremely unstructured and difficult to directly use as input for recommender systems.

Since user-generated data in online sources is so noisy and hard to objectively interpret, we searched for well-structured data that may contain movie ratings ready for extraction. A usable stream of data we found through the social 'share' feature of the IMDb platform. When a user provides a movie rating while using an IMDb app, an option is provided to share the rating to the user's social network e.g. Twitter (Fig. 2). Interestingly, the app makes a suggestion as to what the user may post.



Select your rating and click "Save"

★★★★★★★★★★ 10

★★★★★★★★★☆ 9

Share my rating ON

Other titles you have rated 10/10 recently

The Matrix (1999) ★ 8.7 ★ 10

The Matrix Reloaded (2003) ★ 7.1 ★ 10

From: @sidooms

I rated Death Proof 10/10 #IMDb

Fig. 2. Screenshots from the IMDb iOS app, allowing users to share their movie rating in a well-structured tweet.

As can be noted, this suggestion (or tweet in this case) 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.

To harvest the information contained in these tweets we query the Twitter Search API on a daily basis for tweets containing 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 include also genre metadata.

3. THE DATASET

Starting March 7, 2013 we queried the Twitter search API daily and at the time of writing more than 130,000 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 three files: *ratings.dat*, *movies.dat* and *users.dat*

Metric	MovieTweetings	MovieLens 100K
Ratings	135,099	100,000
Unique Users	20,595	943
Unique Items	12,318	1682
Sparsity	.9995	.9370
Minimum ratings per user	1	20
Average ratings per user	7	106
Maximum ratings per user	429	737
Minimum ratings per item	1	1
Average ratings per item	11	59
Maximum ratings per item	2013	583
Minimum movie year	1898	1922
Maximum movie year	2013	1998
Minimum ratings per day	46	1
Average ratings per day	565	469
Maximum ratings per day	1308	3550

Table I. Dataset metrics comparing the MovieTweetings dataset and the MovieLens 100K dataset. Since the MovieTweetings dataset is updated daily, these may be subject to change.

which respectively store the ratings, movie metadata (i.e., genre information) and user information. We adopted an IMDb identifier as item id and provide a link to the original twitter user to facilitate additional metadata enrichment for both items and users.

Table I overviews some of the main characteristics of the MovieTweetings dataset in comparison with the MovieLens 100K dataset. Currently, it contains over 130,000 ratings provided by more than 20,000 users on 12,000 unique items. Because the dataset is gathered from social media, the divergence of 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 0.9995. Compared to the MovieLens sparsity this is much higher, which results in a more challenging problem for recommender systems. On the other hand, the MovieTweetings 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 429. Because our dataset is more natural, experimental results will be more generalizable to real-life scenarios.

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) may 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

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

of present-day user-centric evaluation experiments.

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). Thanks to Kris Vanhecke for providing the original idea.

REFERENCES

BENNETT, J. AND LANNING, S. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*. Vol. 2007. 35.

DOOMS, S., DE PESSEMIER, T., AND MARTENS, L. 2013. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013*.

DOOMS, S., DE PESSEMIER, T., VERSLYPE, D., NELIS, J., DE MEULENAERE, J., VAN DEN BROECK, W., MARTENS, L., AND DEVELDER, C. 2013. Omus: an optimized multimedia service for the home environment. *Multimedia Tools and Applications*.

HERLOCKER, J. L., KONSTAN, J. A., BORCHERS, A., AND RIEDL, J. 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 230–237.

SHANI, G. AND GUNAWARDANA, A. 2011. Evaluating recommendation systems. In *Recommender systems handbook*. Springer, 257–297.

Simon Dooms is a PhD student at the Wireless and Cable (WiCa) research group of Professor Luc Martens at the Department of Information Technology (INTEC) of Ghent University. His research interests include personalization and recommendation systems, more specific hybrid recommender systems.

Luc Martens is a Professor in electrical applications of electromagnetism at Ghent University and head of the Wireless and Cable (WiCa) research group. He is responsible for the research on experimental characterization of the physical layer of telecommunication systems at the Department of Information Technology (INTEC), Ghent University.