Characterization of Public Datasets for Recommender Systems

Characterization of Public Datasets for Recommender Systems

Erion Çano
Politecnico di Torino
Email: erion.cano@polito.it

Maurizio Morisio
Politecnico di Torino
Email: maurizio.morisio@polito.it

***Abstract*— As Recommender Systems are becoming very common and widespread, there is an increasing need to evaluate their characteristics such as accuracy, diversity, scalability etc. One of the most fruitful ways to do this is by using public datasets with explicit user feedback about the items. In this paper we present and describe more than 20 available datasets covering different domains such as movies, books, music etc. Each dataset is described over a number of attributes such as size, domain, format of the data, type of access. Unfortunately we did not find any information about the quality of the data contained, that remains an open issue. We also refer to examples from the literature about using the datasets to evaluate recommendation algorithms or solutions. Overall aim of the paper is to offer a convenient resource for finding and selecting datasets as a support for the empirical evaluation of recommendation algorithms and techniques.**

***Index Terms*—Public Datasets, Recommender Systems, Recommendation Evaluations.**

I. INTRODUCTION

Recommender Systems (RS) are now being used to recommend any kind of product or service such as movies, books, music, food, software etc. Evaluating a recommender system with respect to different quality criteria such as accuracy, diversity or novelty is a very important step. One way to do that is by running public user evaluation campaigns based on questionnaires which provide rich feedback but are often heavy to process. It is also difficult to find a significant number of feedback providers. Another very common technique is based in using public datasets with real user feedback information. In this paper we describe some of the most common public datasets available for research in RS. These datasets can be used to compare the newly developed algorithms with the existing ones in given settings. In such datasets a representation of implicit or explicit feedback from the users regarding the candidate items is stored in order to allow the RS to produce a recommendation. The feedback is in different forms. In many cases it can be ratings or votes upon the items. The user-rating matrix used in collaborative filtering is a well-known example. In this case the evaluation consists in comparing the predicted ratings with the real ones. In case of content-based or other RS types it can be item reviews or simple tags (keywords) that users provide for items.

The public datasets are usually made available by university research groups or similar institutions. In many cases they also publish their own results obtained using the datasets. The aim of this paper is to describe the main characteristics of the datasets and provide examples of using them, which can be helpful to the many researchers who are currently working in the field of RS. We present 26 datasets, 20 of which are public and active whereas 6 are retired or restricted. They come from different domains and have various characteristics. Eight are used for movies, 5 for books, articles or other learning materials, 4 for music, 3 for food or healthcare and 6 from other domains. We also provide examples of using the datasets in different contexts and for different purposes extracted from the literature. More specific and technical details about how to use the datasets for recommendation evaluations can be found at [1]. The rest of this paper is structured as follows: In section II we describe the active datasets with respect to the domains they pertain. In section III we describe the restricted or retired datasets. In section IV we conclude by making a quantitative discussion.

II. AVAILABLE DATASETS

The datasets were found as part of a systematic literature review on RS we are conducting. This paper is actually byproduct of the systematic review in which we address different research questions

Table 1: Complete list of Datasets and their attributes

Name	Domain	Size	Collected	Status	Format	URL
MovieLens	Movies	DS1: 1K users, 1700 movies, 100K ratings DS2: 6K users, 4K movies, 1M ratings DS3: 72K users, 10K movies, 10M ratings DS4: 138K users, 27K movies, 20M ratings DS5: 230K users, 27K movies, 21M ratings	Released 4/1998 Released 2/2003 Released 1/2009 Released 4/2015 Released 4/2015	Active/ Non-commercial	Csv	[2]
Yahoo Movie Ratings	Movies		November 2003	Request/ Academic		[3]
MovieTwittings	Movies	12425 users, 8458 movies, 65115 ratings	Released 3/1/2013	Active/ Non-commercial	Txt	[4]
IMDB	Movies		Since 1998	Active/ Restricted		[37]
EachMovie	Movies	72,916 users, 1628 movies, 2,811,983 ratings		Retired in 2004		
Netflix	Movies	480k users, 17770 movies, 100M ratings	1998 -2005	Inactive/ Non-commercial	Txt	[38]
Cornell University	Movies	DS1: 1k positive, 1k negative reviews DS2: 1770, 902, 1307, 1027 movie reviews DS3: 5k subjective, 5k objective sentences	Released June 2004 Released July 2005 Released June 2004	Active/ Non-commercial	Txt	[8]
Rotten Tomatoes	Movies	222352 reviews about 11855 movies		Active/ Non-commercial	Txt	[9]
Last.fm	Music	DS1: 1915086 lines, 992 users, 69420 artists DS2:17559530 lines, 359347 users, 107373 artists	May 2007	Active/ Non-commercial	Txt	[10]
Yahoo Music Ratings	Music	15,400 users, 1K songs, 300K ratings	2002 - 2006	Request/ Academic		[3]
Audioscrobbler	Music		May 2005	Merged with Last.fm	Txt	[40]
Million Songs	Music	1M songs, 44745 artists, 7643 unique terms, 2321 unique tags		Active/ Non-commercial	HDF5	[13]
Citation Papers	Research	DS1: 28 researchers x 597 papers DS2: 50 researchers x 100531 papers	June 2010	Active/ Non-commercial	Txt	[21]
Mendely	Research	1857912 articles, 200K users 254681 tags to 27652 articles by 4099 users	Since 2009	Request/ Non-commercial		[16]
MACE	Learning	1148 users, 150K resources, 47K tags	2006-2009	Active/ Private		
Apostle-DS	Learning	1500 user activities of 6 users for 3 months	3 months	Request/ Non-commercial		[17]
BookCrossing	Books	278,858 users, 271379 books, 1149780 ratings	Aug - Sep 2004	Active/ Non-commercial	Sql/Csv	[22]
Organic.Edunet	Food	345 tags, 250 ratings, 325 reviews	Jan 2010 - Sep 2010	Active/ Non-commercial	Rdf/Txt	[24]
Chicago Entree	Food		Sep 1996 - Apr 1999	Active/ Non-commercial	Txt	[25]
Mediacare	Health	Ratings of 15K nursing homes, 4K Hospitals		Active/ Non-commercial	Mdb/Csv	[28]
Dating website	Dating	135,359 users, 17,359,346 ratings	April 2006	Active/ Non-commercial	Txt	[29
WikiLens	Various		Feb. 2008	Retired in 2009	Txt	[41]
Epinions	Various	131228 users, 317755 items, 1127673 reviews	June 2011	Active/ Non-commercial	Sql	[30]
Yahoo Front Page	News	28041015 user visits	2 - 16 Oct. 2011	Request/ Academic		[32]
Tags2Con	URLs	1397 users, 1569 tags, 1681 ratings, 739 URLs, 603 domains	Dec 2007 - Apr 2008	Active/ Non-commercial	Rdf	[33]
Jester	Jokes	DS1: 4.1M ratings -10.00 - +10.00 of 100 jokes from 73421 users, DS2: 1.7 M ratings -10.00 - +10.00 of 150 jokes from 59132 users, DS3: Update of DS2 with 500 K new jokes by 79681 users	Apr 1999 - May 2003 Nov 2006 - May 2009 Nov 2006 - Nov 2012	Active/ Non-commercial	Xls	[35]

that have to do with the construction and evaluation of RS. We searched in SpringerLink, ScienceDirect, IEEExplore and ACM using keywords like "Evaluating Recommender Systems", "Public Datasets for Recommender Systems" etc. In Table 1 we summarize some basic characteristics of the 26

datasets we present. In 'Size' column we give the number of users, number of items managed and number of ratings (in some cases reviews or tags). The 'Collected' column contains the period when the dataset was started or the release date in some cases. In the 'Status' column we show the current status of the dataset and the usage conditions/constrains. Most of the datasets are "Active" (in the sense that they are freely available and usually updated). Some of the datasets can be obtained upon request to the owner/maintainer (denoted as "Request"). Most of the datasets can be used for any non-commercial purpose. Some datasets (denoted as "Academic") can be used for research purposes only. In the 'Format' column the data format is given.

A. Movies

MovieLens. This is one of the most used datasets for algorithms evaluation in the community of Recommender System Research. The dataset was collected and made available by GroupLens Research, a research lab in the department of computer science, University of Minnesota. There are four stable versions of the dataset which vary in size and release date (Table 2). The fifth version is the most recent. It changes over time and it is not appropriate for reporting research results [2]. The datasets are public and open for non-commercial use only. The many user ratings these datasets contain make them very suitable for evaluating different versions of Collaborative Filtering recommendation algorithms.

Table 2: Versions of MovieLens Dataset

Version	Users	Movies	Ratings	Released	Format
100k	1k	1.7k	100k	4/1998	Txt
1M	6k	4k	1M	2/2003	Csv
10M	72k	10k	10M	1/2009	Csv
20M	138k	27k	20M	4/2015	Csv
Latest	230k	27k	21M	4/2015	Csv

Yahoo Movie Ratings. This dataset contains a small sample of the Yahoo Movies community's preferences for various movies. The movies are rated on a scale from A+ to F. The dataset also contains a large amount of descriptive information about many movies released prior to November 2003, including cast, crew, synopsis, genre, average ratings, awards, etc. It may be used by researchers to validate recommender systems of different algorithms, including hybrid content-based and collaborative filtering. The dataset is available for download upon request from [3] and can be used for academic purposes only.

MovieTwittings. This is a very recent movie dataset which collects the ratings from user twits. It is composed of 2 files which contain the movies and the ratings respectively. Unlike MovieLens and other filtered datasets which contain only users with a minimum number (i.e. 20) of ratings, MovieTwittings has a number of ratings which varies from 1 to 305 per user. The authors started querying the Twitter search API on March 2013. Since then the dataset continuously grows based on the daily twits of different users. Currently the dataset contains 12425 unique users, 8458 unique items and 65115 ratings. The many and recent movie ratings it contains make it suitable for the evaluation of various versions of Collaborative Filtering algorithms. The dataset is open for public use and can be downloaded from [4]. More details about MovieTwittings can be found at the authors' publication [5].

Movie Review Dataset. This dataset was created by researchers at Cornell University and includes subjective movie reviews. There are actually 3 datasets: Sentiment polarity, Sentiment scale and Subjectivity dataset. The Sentiment polarity dataset is a collection of 1000 positive and 1000 negative movie reviews. The Sentiment scale dataset is a collection of four subjective review sets (with 1770, 902, 1307 and 1027 movie reviews each) which can be used to infer ratings. The Subjectivity dataset is a collection of 5000 subjective and 5000 objective movie review sentences. The datasets were released in June 2004 the first and the third and July 2005 the second. All the data are in Txt format. These datasets are very suitable to evaluate Machine Learning, Natural Language Processing or Text Processing algorithms/techniques used for ratings or recommendations. The authors use the Sentiment scale dataset in [6] to evaluate their algorithm that addresses the rating-inference problem. They also use the first and third datasets in [7] to train and evaluate

the Naive Bayes and SVM text categorizers they use to determine the sentiment polarity of the subjective users' reviews. The datasets are publicly available for non-commercial use. More info about the datasets and the download links can be found at [8].

Rotten Tomatoes. This is a highly viewed movie review website with a rich movie dataset. A subset of its dataset can be freely downloaded from [9]. It is comprised of tab separated textual movie review phrases from the original Rotten Tomato dataset. The dataset contains 222352 reviews about 11855 movies. There is a train/test split of the dataset to make it more suitable for benchmarking algorithms. The structure of the dataset makes it very suitable for sentiment analysis and machine learning research and benchmarking.

The movie domain is obviously the most preferred for recommender systems research. This is partly because there are many public and rich datasets (apart from the 5 above, we describe 3 other restricted/retired movie datasets in the next section) with explicit user feedback (movie ratings). The regular structure and contents of these datasets makes them very suitable for evaluating different versions of Collaborative Filtering or hybrid recommendation algorithms. MovieLens is probably the most used dataset for research purposes.

B. Music

Last.fm. This is a subset of last.fm website music dataset, one of the most important in the domain of music. There are actually two versions of the dataset: Last.fm-360k which includes the top artists of 360k users and Last.fm-1k which includes the full listening history of 1k users. The datasets contain information about 69420 and 107373 artists respectively. They also contain information about track preferences of the users and some attributes of every user. This dataset is suitable for content-based or hybrid algorithms evaluation. All the data is stored in tab separated textual values. The dataset can be freely downloaded from [10] for non-commercial use only. It is also required to reference last.fm website when using it.

Yahoo Music Ratings. This dataset contains music ratings supplied by users while browsing Yahoo Music services and ratings for randomly selected songs collected during an online survey conducted by Yahoo Research [11]. The rating data includes 15.4k users, and 1k songs. The dataset includes approximately 300k user-supplied ratings, and exactly 54k ratings for randomly selected songs. The data were collected between 2002 and 2006. The rich user feedback (many explicit ratings) it contains makes it very suitable to evaluate different versions of collaborative filtering or hybrid recommendation algorithms. It can be downloaded upon request from [3] and used for academic purposes only.

The Million Song Dataset. The Million Song Dataset (MSD) is an attempt to help researchers in the field of Music Analysis and Recommendations by providing a large-scale dataset. The main purposes of the dataset are:

- encouraging research on algorithms that scale to commercial sizes
- providing a reference dataset for evaluating algorithms by using the audio features
- being a shortcut alternative for creating a large dataset with the Echo Nest's API
- helping new researchers get started in the MIR field, develop music recommendations and study music similarity.

A large dataset helps to reveal problems with algorithms scaling and discover rare phenomena or patterns that may not be discoverable in small datasets. The dataset contains 1M songs/files, 44745 unique artists, 7643 unique terms (Echo Nest tags) and 2321 unique tags (more detail at [12]). The data is stored in HDF5 format. Its size and the many music features (i.e. pitches, timbre, loudness etc.) it contains make MSD very suitable to evaluate content-based or hybrid recommendation algorithms. The dataset can be downloaded from [13] for non-commercial use (be aware that the download size is about 300 GB).

For a long time Music Information Retrieval (MIR) research has suffered the lack of publically available and large-scale open data for personalized music recommendations, mainly because of the privacy and intellectual property concerns. Things seem to have changed with the partial release of last.fm data (march 2010) and the publication of MSD (march 2011). These datasets are a very good source for evaluating different kinds of content-based

or hybrid algorithms. The publication of MSD and also the MSD challenge [14] are hoped to greatly facilitate the academic research in MIR.

C. Learning Materials

Research and learning materials is another domain where Recommender Systems have gained significant applicability. As in the other domains, having publically available sources of data appropriate for development and evaluation of novel recommendation solutions is indispensable. For this reason the dataTEL Theme Team of the STELLAR Network of Excellence lunched the the first dataTEL Challange [15] which is a call to research groups to submit datasets from Technology Enhanced Learning applications.

Mendely's Data Set. The first to submit a dataset to the dataTEL challenge was Mendely, a research platform that helps users to organize their research, collaborate with colleagues and discover new knowledge. Their dataset contains data regarding the articles that Mendely records and analyses and is composed of five files: Online catalog, Online article view log, Library readership, Library stars and Article tags. Online catalog contains article metadata like title, year, number of readers etc. for 1857912 articles. Online article view log contains a random sampling of 200k users that are extracted from usage logs. Library readership contains 41220 user libraries that contain more than 20 articles. Library stars provides data on the 186976 articles that have been starred by users. Article tag contains 254681 tags applied to 27652 articles by 4099 users. The dataset can be downloaded from [16] upon request and for non-commercial use only.

Mendely dataset can be used for research on TEL recommender systems as it provides the possibility to extract users' interests in articles, identify users who share common interests etc. It was actually used at [18] to provide data about library readership, library start and article tags and experiment with user-based and item-based collaborative filtering algorithms for TEL. It can be suitable to evaluate different kind of recommendation algorithms like content-based, hybrid, etc.

Apostle-DS. This dataset originates from the APOSDLE EU project which is an adaptive work-integrated learning system aiming to improve knowledge worker productivity by supporting learning situations within everyday work tasks. It recommends resources (documents, videos, links) and knowledgeable persons (colleagues) based on the user's current context. The data have been collected from August 2009 till December 2009. The dataset captures 1500 user activities of 6 users during a period of 3 months. The activities captured are task performing, resource viewing, annotation editing, learning goal selection, adding resource to collection, contacting person, browsing data etc.

The dataset is downloadable upon request from [17] for non-commercial use only. The dataset was used at [18] to evaluate item-based and user-based collaborative filtering algorithms. This dataset provides data on the learning goal of the learner when s/he is performing a task. Such data is useful to improve similarity measures between users and to find users who share similar goals. Improving the recommendations of relevant learning resources is one of the best purposes this dataset can be used for.

Citation Papers Dataset. This is a recent and experimental dataset made available by two researchers of National University of Singapore, School of Computing. There are two versions of the dataset: small and large. The small version is a subset of Association for Computation Linguistics (ACL) Anthology Archive papers published between 2000 and 2006. It contains 597 papers to recommend and the interest of 28 researchers. The large dataset is a subset of ACL DL papers published between 2000 and 2010. It contains 100531 papers to recommend and the interest of 50 researchers. All the data are in Txt format. The datasets are suitable for evaluating Collaborative Filtering, Content-based Filtering and hybrid recommendation approaches. In [19] the authors provide a profile comparison method to recommend papers to researchers. They use the small dataset first to build the profiles of their users and then to recommend them the most similar candidate papers. In [20] they extend their previous solution by using the large dataset. Here they use Collaborative Filtering to find potential citation papers and alleviate

data sparsity. The datasets are publicly available at [21] for non-commercial use.

BookCrossing. This dataset was collected from the Book-Crossing community in August-September 2004. It is composed of 3 tables: BX-Users which contains user IDs for 278858 users and some demographic data (Location and Age) when available, BX-Books which contains the books identified by the ISBN and some content-based information like Book-Title, Book-Year, Year-Of-Publication etc, and BX-Book-Ratings which contains the ratings expressed on a scale from 1-10. The data is in Sql and Csv formats and can be downloaded from [22] for non commercial use only. The authors use this dataset in [23] to evaluate the topical diversity of the Item-based Collaborative Filtering recommendations. They conclude that the user's overall liking of recommendations goes beyond accuracy and involves other factors like diversity. The dataset is also suitable for the evaluation of content-based recommendation solutions as it contains different book attributes.

D. Food and Healthcare

Organic.Edunet. Organic.Edunet is a European initiative program that aims to facilitate access and exploitation of educational content related to Organic Agriculture. Part of the initiative is also the deployment of a learning repository (dataset) which provides access to more than 10500 learning resources from 11 institutional repositories. The dataset was collected from Jan. 2010 till Sep. 2010. It includes information about 345 tags, 250 ratings and 325 textual reviews that users have provided. The ratings are collected upon three different criteria: the usefulness of a learning resource, the relevance of the organic thematic and quality of the metadata. This makes it appropriate to evaluate multi-criteria recommender systems. The data are available in Rdf or Txt format and can be downloaded from [24] for non-commercial use only.

Chicago Entree. This dataset contains recorded interactions with Entree Chicago restaurant recommendation system from September 1996 to April 1999. The system recommends restaurants to the user based on factors such as cuisine, price, style, atmosphere, etc. The user can then provide feedback such as find a nicer or less expensive restaurant. The data is in Txt format and organized into files in which each line represents a session of user interaction with the system. The (tab-separated) fields are as follows:

Date, IP, EntryPoint, R1, R2, ..., RN, EndPoint

R1, R2, ... RN are the rates of the rated restaurants. The dataset can be download from [25] for non-commercial use. It is suitable for research in Knowledge based and Case-based Recommender Systems. The dataset and the Chicago Entree Knowledge-based Recommender System are described by the author at [26]. He has also used the dataset in [27] to built a domain-independent Case-based Recommender System for on-line information access.

Mediacare. These are the official datasets of Mediacare.gov websites such as Nursing Home Compare, Hospital Compare, Physician Compare, Home Health Compare, Dialysis Facility Compare and Supplier Directory. The collected data are anonymous and reflects the preferences of different patients in the USA for about 15k Nursing Homes, 4k Hospitals, 77k Medical Suppliers from which 6357 Dialysis Facility Suppliers etc. They are in Csv format and regularly updated (last version of May 2015). These datasets can be used to evaluate Healthcare Recommender Systems built with different recommendation algorithms. All the datasets are public and can be downloaded without any permission from [28].

E. Other Domains

There are also some available datasets which are domain independent (contain different categories of objects) or used in less common application domains of Recommender Systems. Their usage is not limited only to Recommender Systems but also to other classes of Data Mining or information filtering applications.

Libimseti.cz. This is a dating dataset collected by Charles University using libimseti.cz dating forum. There are two Txt files which contain gender information of 135359 users and their 17359346 anonymous ratings of each-other profiles. The data were collected in April 2006. Dating websites collect data about users only (the object to recommend is a person) and thus are a very good source for the construction and evaluation of various Collaborative Filtering Recommendation algorithms (more than

17M ratings in this dataset). It is publicly available for download at [29] with few conditions.

Epinions. This dataset was extracted from Epinions website in June 2011 and contains reviews from users on items, trust values between users, items category, categories hierarchy and users expertise on categories. The dataset contains 131228 users, 317755 items and 1127673 reviews. 113629 users have at least one rating and 47522 users have at least one trust relation. It is available in Sql format and can be downloaded from [30] for non-commercial use. This is one of the few public datasets that includes trust relationships. The domain independence and the rich information structure makes this dataset suitable to evaluate various Recommender Systems such as Collaborative Filtering, Content-Based, Trust-Based or Hybrid. A more detailed description of the dataset can be found at [31].

Yahoo Front Page. This dataset contains a fraction of user click log for news articles displayed in the Featured Tab of the Today Module on Yahoo's front page. It contains 15 days of data (from October 2 to 16, 2011) with raw features (so that researchers can try out different feature generation methods). There are 28041015 user visits to the Today Module on Yahoo's FrontPage. For each visit, the user is associated with a binary feature vector of dimension 136 that contains information about the user like age, gender, behavior targeting features, etc. Online content recommendation represents an important example of interactive machine learning problems that require an efficient tradeoff between exploration and exploitation. It can be downloaded upon request from [32] and used for academic purposes only.

Tags2Con. This dataset was created from a subset of Delicious dump by linking the tags to their real meaning. It contains 1681 user-bookmark pairs from 1397 unique user with 1569 tags from 603 website domains. The dataset includes annotations from users which have less than 1k tags and have used at least 10 different tags in 5 different websites. It can be used to evaluate recommendation solutions of different websites. The dataset is available in Rdf format at [33] for non-commercial use. A more detailed description of the dataset can be found at [34] where the authors also explain the annotation process.

Jester. This is the dataset of Jester Online Joke Recommender System. There are three versions of it: Dataset1 contains more than 4.1M continuous ratings (-10.00 to +10.00) of 100 jokes from 73421 users collected between April 1999 to May 2003. Dataset2 contains over 1.7M continuous ratings (-10.00 to +10.00) of 150 jokes from 59132 users collected between November 2006 to May 2009. Dataset2+ is an updated version of Dataset2 with over 500k ratings from 79681 total users collected between November 2006 to November 2012. All the datasets are in Xls format and can be downloaded from [35] for non-commercial use. They are mostly appropriate to evaluate different versions of Collaborative Filtering Recommendation algorithms. In [36] the authors use the dataset to evaluate their own Collaborative Filtering algorithm called Eigenstate.

III. RETIRED OR RESTRICTED DATASETS

In this section we describe some other datasets which have been retired or which are still active bur are private and cannot be used even for research purposes.

IMDB. It is probably the largest movie collection with more than 3M movies and 60M registered users. A subset of the plain text data files is available for download [37]. Nevertheless the public use is restricted.

EachMovie. This movie recommender was run by HP Research and had a rich dataset collection. When EachMovie was shut down, the dataset was available to the public for use in research. MovieLens was originally based on this dataset. The dataset has been used by many researchers and referenced in different publications. It was retired by HP in October 2004 and it is no longer available for download.

NetFlix. This dataset was used to support the participating teams of the Netflix Price. The dataset stopped being available since the finish of the competition in 2009. However it can be still downloaded from [38] and used under netflix's copyright notice.

Audioscrobbler. Audioscrobbler was a University project of Richard Jones who developed the first plugins and opened the API to the community [39]. It was limited to recording music its users played on a registered computer, which allowed for charting and collaborative filtering. In August 2005 Audioscrobbler merged with last.fm. The original

Audioscrobbler dataset is still available for download at [40].

MACE. This dataset originates from the MACE project and is another dataset submitted to dataTEL challenge. It provides access to 150k learning resources in architecture. More than 12k of these resources have been accessed by registered users. They hold together about 47k tags, 12k classification terms and many other actions performed by the users such as viewing and downloading. Unfortunately the dataset is restricted to public use.

WikiLens. This was a generalized collaborative recommender system that allowed its community to define item types (e.g. beer) and categories (e.g. microbrews, pale ales, stouts), and then rate and get recommendations for items. It is no longer active but a dump of the data of February 2008 can be still downloaded at [41]. It is in Txt format and can be used to evaluate Collaborative Filtering algorithms of different items.

IV. DISCUSSION

In this paper we presented 26 public datasets that can be used to build and/or evaluate Recommender Systems. Our goal is not to make an exhaustive review of all the available datasets. We focused only in the presentation and characterization of the most widely used dataset that Recommender System researchers can use for empirical analysis or evaluation of their algorithms and methods. Most of the datasets contain information about popular items such as movies, music, books etc. However there are also public datasets of less common items such as jokes, restaurants etc. The collection period of the datasets ranges from 1996 (Chicago Entree) to 2013 (MovieTwittings). In some cases this period is several years long whereas in other cases it is short and only the release date is provided. The data are mostly provided as tab separated textual values but there are also other data formats such as Sql, Csv or Mdb.

Most of the datasets contain explicit ratings of the users about the items. This makes them very appropriate for experimenting with Collaborative Filtering algorithms, and user similarity measures. There are also datasets (i.e. music and books) rich of item features and appropriate for content-based algorithms. A couple of datasets contain subjective user reviews of the items. They can be used for different purposes like user profile building, sentiment analysis and machine learning natural language processing or text processing and categorization algorithms or techniques.

Some of the datasets are continuously updated and can be freely downloaded for non-commercial use. Some others are active and available for the public use but can be downloaded only upon submitting a request to the owner/maintainer. Few datasets are restricted to academic use only. There are also some datasets that are not being updated any longer or that are restricted, retired or merged with others. It is important to note that it was not possible for us to find or perform any data quality assessment of the described datasets. Researchers willing to use any of the datasets should first check if the available data meet their research requirements.

REFERENCES

[1] P. Cremonesi, R. Turrin, E. Lentini and M. Matteucci, "An Evaluation Methodology for Collaborative Recommender Systems", White paper, March 2010.

[2] http://www.grouplens.org/node/73

[3] http://webscope.sandbox.yahoo.com/catalog.php?datatype=r

[4] https://github.com/sidooms/MovieTweetings

[5] S. Dooms, T. Pessiemier, L. Martens , "MovieTweetings: a Movie Rating Dataset Collected From Twitter", Workshop on Crowd sourcing and Human Computation for Recommender Systems, 2013.

[6] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Based on Minimum Cuts", ACL '04, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Article No. 271

[7] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", ACL '05, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115-124

[8] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[9] https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data

[10] http://www.dtic.upf.edu/~ocelma/

MusicRecommendationDataset/index.html

[11] http://webscope.sandbox.yahoo.com/catalog.php?datatype=r

[12] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Songs Dataset", 12th International Society for Music Information Retrieval Conference, Ismir 2011.

[13] http://labrosa.ee.columbia.edu/millionsong/

[14] B. McFee, T. Bertin-Mahieux, D. P. W. Ellis and G. R. G. Lanckriet, "The Million Song Dataset Challange"

[15] http://adenu.ia.uned.es/workshops/recsystel2010/datatel.htm

[16] http://www.teleurope.eu/pg/pages/view/50632/

[17] http://www.teleurope.eu/pg/pages/view/50647/

[18] K. Verbert, H. Drachsler, N. Manouselis, M. Wolpers, R. Vourikari and E. Duval, "Dataset-driven Research for Improving Recommender Systems for Learning"

[19] K. Sugiyama and M. Kan, "Scholarly Paper Recommendation via User's Recent Research Interests", Proceedings of the 10th annual joint conference on Digital libraries, ACM 2010, pp. 29-38

[20] K. Sugiyama and M. Kan, "Exploiting Potential Citation Papers in Scholarly Paper Recommendation", Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, ACM 2013, pp. 153-162

[21] http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html

[22] http://www2.informatik.uni-freiburg.de/~cziegler/BX/

[23] C. Ziegler, S. McNee, J. Konstan and G. Lausen, "Improving Recommendation Lists Through Topic Diversification", WWW 2005 Proceedings of the 14th international conference on World Wide Web, ACM 2005, pp. 22-32

[24] http://project.organic-edunet.eu

[25] http://archive.ics.uci.edu/ml/datasets/Entree+Chicago+Recommendation+Data

[26] R. Bruke, "Knowledge-based recommender systems", In A. Kent (ed.), *Encyclopedia of Library and Information Systems*. Vol. 69, Supplement 32, 2000

[27] R. Bruke, "The Wasabi Personal Shopper: A case-Based Recommender System", AAAI '99/IAAI '99 Proceedings of the sixteenth national conference on Artificial intelligence, pp. 844-849.

[28] https://data.medicare.gov

[29] http://www.occamslab.com/petricek/data/

[30] http://liris.cnrs.fr/red/

[31] S. Meyffret, E. Gulliot, L. Medini and F. Laforest, "RED: Rich Epinions Dataset for Recommender Systems"

[32] http://webscope.sandbox.yahoo.com/catalog.php?datatype=r

[33] http://disi.unitn.it/~knowdive/dataset/delicious/

[34] P. Andrews, J. Pane and I. Zaihrayeu, "Semantic Disambiguation in Folksonomy: A Case Study", Proceedings of the 2009 international conference on Advanced language technologies for digital libraries, pp. 114-134.

[35] http://www.ieor.berkeley.edu/~goldberg/jester-data/

[36] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, "Eigentaste: A Constant Time Collaborative Filtering Algorithm", Journal of Information Retrieval, Vol. 4, Issue 2, July 2001, pp. 133-151

[37] http://www.imdb.com/interfaces

[38] http://www.lifecrunch.biz/wp-content/uploads/2011/04/nf_prize_dataset.tar.gz

[39] http://en.wikipedia.org/wiki/Last.fm

[40] http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html

[41] http://grouplens.org/datasets/wikilens/