Data Sets and News Recommendation

Özlem Özgöbek, Nafiseh Shabib, Jon Atle Gulla

Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
{ozlemo,shabib,jag}@idi.ntnu.no

Abstract. Datasets are important for training and testing many information processing applications. In the field of news recommendation, there are still few available datasets, and many feel obliged to use non-news datasets to test their algorithms for news recommendation. This paper presents some of the most common datasets for recommender systems in general, and explains why these datasets do not fully satisfy the needs in news recommendation. We then discuss the ongoing process of building up an entirely new dataset for Norwegian news in the SmartMedia project. In particular, we go through some of the features of news datasets that separate them from many other datasets and are crucial for their use in news recommendation.

1 Introduction

A dataset is a collection of data that is used to train and test new systems under development. As real systems work on data, it is vital to validate and verify their behavior with extensive datasets prior to their deployment. Moreover, with the increasing popularity of data-driven learning applications, high-quality datasets have become critical for training these applications to perform at an acceptable level of precision.

Scientific methods rest on systematic use of measurements and their subsequent analysis. According to [4], datasets serve at least four different purposes in scientific research: Verification of publications (scientific publications can be verified by repeating the same study with the same data), longitudinal research (long term availability of the data for a long period of research), interdisciplinary use of data (usage of the same dataset for different purposes may lead to new insights and scientific development), and valorisation (Dataset ownership enables the acquisition of new research projects [4]). Also, Dekker claims that datasets are becoming more valuable as products by themselves and justify their own publications in the scientific community.

The nature of datasets depends on both the type of application and the choice of domain. In news recommendation, both machine learning techniques and traditional search technologies are applied and need to be verified against suitable datasets. Machine learning techniques require training and test datasets that are feature-rich and may involve aspects that are directly present in the news itself, while search applications are usually tested with a narrower focus and no regard of user differences.

More complicated, though, is the fact that it is difficult to replicate the news domain as a fixed controlled document set. We want the datasets to mirror the users preference of news in real news contexts, which means that we need the dynamic and unpredictable nature of news to be reflected in the way these datasets are built up. This is a challenging task and partly explains why there are only a few small news datasets and no large-scale datasets available.

2 Related Work and Comparison of Existing Datasets

Recommender system has been identified as the way to help individuals to find information or items that are most likely to be interesting to them or to be relevant to their needs [1] and it is still very interesting area in the research and real world setting. Thus, monitoring the operation of a recommender system is a challenging task and it is common to evaluate recommendation algorithms with available public dataset (e.g. MovieLens, Netflix, Million Song Dataset). Furthermore, the datasets are used as benchmarks to develop new recommendation algorithms and to compare them to other algorithms in given settings [10]. In the news domain, recommender systems are increasingly applied, but still we are facing lack of publicly available dataset that completely interoperate in news domain. In this section, we present an overview of different datasets, which are available in different domains and then in the next section we introduce our dataset.

2.1 MovieLens Dataset

MovieLens is a movie recommender system project at the University of Minnesota, led by the GroupLens Research Group [1]. There are three datasets of different sizes that have been collected in different time periods [2]. All data is collected through the MovieLens web site. The 100K and 1M datasets contain simple demographic information about the users (age, gender, occupation, zip) while the 10M data set only contains user id. For the 100k dataset the data was collected during the seven-month period from 19 September 1997 to 22 April 1998. For the 1M dataset the data was collected from 6040 users who joined MovieLens in 2000. The 10M dataset contains $10,000,054$ ratings (ranging from 1 to 5) and $95,580$ tags applied to $10,681$ movies by $71,567$ users [3].ngs and 95580 tags applied to 10681 movies by 71567 users [4].

[1] http://grouplens.org

[2] http://grouplens.org/datasets/movielens/

[3] http://files.grouplens.org/datasets/movielens/ml-10m-README.html

[4] http://files.grouplens.org/datasets/movielens/ml-10m-README.html

2.2 Netflix Dataset

On October 2, 2006, Netflix, the world's largest online DVD rental service, announced the 1-million Netflix Prize for improving their movie recommendation service [5]. To aid contestants, Netflix publicly released a dataset containing 100,480,507 movie ratings, created by 480,189 Netflix subscribers between December 1999 and December 2005.

2.3 MoviePilot Dataset

The MoviePilot dataset was released as part of the Context-Aware Movie Recommendation 2011 Challenge at ACM RecSys. There were two tracks in this challenge. In the Context-Aware Movie Recommendation (CAMRa) Challenge [9] they requested participants to identify which members of particular households were responsible for a number of event interactions with the system in the form of ratings. The contest provided a training dataset with information about ratings in a movie RS, including the household members who provided the ratings, and the associated time stamps. The goal was to identify the users who had been responsible for certain events (ratings), and whose household and time stamp were given in a randomly sampled test dataset. This task is assumed to be equivalent to the task of identifying active users requesting recommendations at a particular time. In another track, the main task of the challenge was recommending a given set of items to a household of users. The MoviePilot dataset contains 290 unique households with between two to four members, and a total of 602 users, of which the majority has been assigned to a particular household. The dataset contains information about which user rated which movie at which time. More details are shown in Table 1.

Datasets	Movies	Users	Ratings
Training	23,974	171,670	4,536,891
Household in training	7,710	602	145,069.
Test	811	594	4482

Table 1. The MoviePilot dataset characteristics

2.4 Million Song Dataset

The Million Song Dataset (MSD) [2] is a collection of music audio features and metadata that has created to support research into industrial-scale music information retrieval [6]. The Million Song Dataset (MSD), a freely-available collection

[5] http://www.netflixprize.com

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

of meta data for one million of contemporary songs (e.g,. song titles, artists, year of publication, audio features, and much more) [7].
The Million Song Dataset is a cluster of complementary datasets contributed by the community: SecondHandSongs dataset for cover songs, musiXmatch dataset for lyrics, Last.fm dataset for song-level tags and similarity, and Taste Profile subset for user data. Comprising several complementary datasets that are linked to the same set of songs, the MSD contains extensive meta-data, audio features, and song-level, lyrics, cover songs, similar artists, and similar songs. In Lastfm dataset, songs have different tags with different degrees. The tag's degree shows how much the song is linked to a particular tag. Some of the characteristics of Millions of song Million Song Dataset are shown in Table 2.

Songs	1,000,000
Data	273 GB
Unique artists	44,745
Unique terms	7,643
Unique musicbrainz tags	2,321
Artists with at least one term	43,943
Asymmetric similarity relationship	2,201,916
Dated tracks starting from 1922	515,576

Table 2. The Million Song Dataset characteristics [2]

2.5 Last.fm Dataset

Last.fm dataset is one of the largest music recommender system datasets [3]. It contains 359,347 unique users and 17,559,530 of total lines which includes -user, artist, plays- tuples collected from Last.fm API [7]. This data was collected by Oscar Celma @ MTG/UPF, during Fall 2008 [8], and the it is available for non-commercial use.

This dataset contains user profile information as gender, age, subscription date, country, name. It also contains information about which user listened to which artist and how many times as the user name, artist id, artist name, number of plays.

The Last.fm dataset contains only the artist information that a user listened to. By looking at the number of plays we can figure out the users' most popular artists and the similarities between users' preferences. It is not possible to assess the similarities between artists or songs. So this dataset is mostly suitable for training collaborative filtering methods for artist recommendation. Since there is

[7] http://last.fm/

[8] http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html

no information about the individual songs, it is not possible to recommend a song that the user has not listened to yet. The age, gender and country information can be used for group recommendations. For example, if there is an artist who is mostly listened to 22-25 year old people, it may be possible to recommend it to other users in the same age group and have not listened to it yet.

2.6 Jester Dataset

Jester is an online joke recommender system which has three different versions of publicly available collaborative filtering dataset [5]. The first version of Jester dataset contains over 4 million continuous ratings collected from $73,421$ users. There are 100 jokes in the dataset and it is collected between April 1999 - May 2003. The second version contains over 1.7 million continuous ratings of 150 jokes from $59,132$ users and it is collected between November 2006 - May 2009. Also there is an updated version of the second dataset with over $500,000$ new ratings from $79,681$ total users. [9] The ratings of Jester dataset is in range between -10.00 and $+10.00$ as a floating number. The dataset contains two files where the first one includes the item ID and the jokes, and the other one includes user ID, item ID and ratings.

2.7 Book-Crossing Dataset

Book-Crossing dataset is collected by Cai-Nicolas Ziegler [12] in 4-weeks from August to September 2004. [10] The dataset contains $278,858$ users, about $271,379$ books and $1,149,780$ both explicit and implicit ratings. In the dataset the demographic information is also provided. For the user privacy the demographic data is anonymized. The Book-Crossing dataset includes 3 tables: BX-Users (user ID, location, age), BX-Books (ISBN, book title, author, publisher, year of publication) and BX-Book-Ratings (explicit ratings from 1 to 10, implicit ratings expressed by 0).

2.8 YOW Dataset

YOW dataset is collected at the Carnegie Mellon University for the Yow-now news filtering system. Yow-now was an information filtering system that delivered news articles to users from various RSS feeds. [11] Within this project the data is collected by a one month user study which includes approximately 25 people and 7000+ feedback entries from all users. In total 383 articles rated by each user. It is collected both implicit and explicit feedback from users. Explicit feedback is collected as rating from 1 to 5 and explicit feedback is collected by tracking the user actions (mouse, keyboard and scroll activities) during the usage of the system [11].

[9] http://eigentaste.berkeley.edu/dataset/

[10] http://www.informatik.uni-freiburg.de/ cziegler/BX/

[11] http://users.soe.ucsc.edu/ yiz/papers/data/YOWStudy/

The YOW dataset contains a lot of details about the user actions while reading news. Both explicit and implicit feedbacks are available in the dataset, making this dataset well suited for collaborative filtering. Since there is no information about news content, content-based filtering is not possible with this dataset. YOW dataset is the only publicly available dataset that we could find on the news domain.

	Domain	Size			Feedback	
		Items	Users	Ratings	Explicit	Implicit
MovieLens 100k	Movie	1682 movies	943	100,000	Ratings from 1 to 5	-
MovieLens 1M	Movie	3900 movies	6040	1,000,209	Ratings from 1 to 5	-
MovieLens 10M	Movie	10682 movies	71567	10,000,054	Ratings from 1 to 5	-
Netflix (Training)	Movie	17, 770 movies	480, 189	100, 480, 507	Ratings from 1 to 5	-
MoviePilot (Training)	Movie	23, 974 movies	171, 670	4, 536, 891	Ratings	-
Last.fm	Music	186, 642 artists	359, 347	17, 559, 530	Ratings	-
Million Song (cluster of complementary datasets)	Music	1, 000, 000 songs	-	-	-	-
Jester v1	Joke	100 jokes	73, 421	4, 000, 000	Ratings from -10.00 to +10.00	-
Jester v2	Joke	150 jokes	59, 132	1, 700, 000	Ratings from -10.00 to +10.00	-
Book-Crossing	Book	271, 379 books	278, 858	1, 149, 780	Ratings from 1 to 10	✓
YOW	News	383 articles	25	7000+	Ratings from 1 to 5	Mouse, keyboard and scroll activities

Table 3. Comparison of different datasets of recommender systems and their properties.

3 SmartMedia Dataset

The specific challenges of news domain requires the usage of a special dataset for testing the news recommender system. Within our SmartMedia project [6] we are building a dataset on Norwegian news domain.

As a specific challenge to news domain, there can be hundreds of new articles every hour and it is not always possible to get enough ratings to overcome the problem of data sparsity. In SmartMedia dataset we are trying to build a dataset which is less sparse than other datasets in news domain by limiting the number of news articles gathered from different sources. As it is stated in [8] implicit feedback is one of the challenges that both it is needed to be collected and considered the user privacy issues.

SmartMedia dataset will contain both explicit and implicit feedback from users. As the explicit feedback, we get ratings from 1-5 for each news article. Implicit feedback contains the time spent on each article, current location and timestamp. The dataset will also contain some personal information like occupation, age and gender. We developed an application to collect the data from users. We recruited 20 users with different backgrounds, occupations and within different age groups to make the dataset more homogeneous and realistic. We asked the users to read and rate the news articles which are collected from different Norwegian news sources for a period of two weeks. As a result of our data collection process we expect nearly 8500 ratings of 3000 articles.

4 Discussion

The properties of different data sets of recommender systems is given in Table 3. We have chosen to compare these data sets because most of them are very well known, publicly available and regularly used data sets in the recommender system research. Since each domain have its own specific challenges, we also wanted to compare data sets from different domains as much as possible. Recommending news articles has different challenges than recommending movies or music [8]. For example, for recommending movies, learning the users' preferences/tastes about movies can be enough. But for the news domain one may find the article important even though she does not like the topic or she may not want to read the other articles in the same topic. For the news recommendation YOW data set was the only publicly available data set with ratings which is also suitable for collaborative filtering.

Nearly all the datasets that we compare have explicit feedback from the users. So asking the user how much she liked the recommendation (rating) is the most common way to get feedbacks. Usually the ratings are integers ranging from 1 to 5 or 1 to 10. Only the Jester dataset ratings are not integers.

Most of the datasets do not have implicit feedback. Only the YOW dataset has a lot of implicit feedbacks. Since recommending news articles is different than recommending items in other domains in many aspects like recency (the people usually want to read fresh news) and quick or instantaneous changes of user interest (age, cultural level, mood or on going circumstances in the world may affect the preferences of users), the need for implicit ratings is more demanding [8]. So to develop better recommender systems in specific domains it is important to choose the suitable dataset. By the SmartMedia dataset we are aiming to have a realistic and less sparse dataset (compared to the YOW dataset) in Norwegian news domain including explicit and implicit feedbacks.

5 Conclusion

Since each domain has its own specific requirements for the recommender systems, the need for choosing the suitable dataset for developing and improving systems is quite obvious. Especially the news domain is very different in many aspects like the item churn and recency compared to the other domains.

In this paper we provided a comparison of recommender system datasets from different domains and we presented our SmartMedia news recommender system dataset which will be the first publicly available dataset in the Norwegian news domain.

References

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.
2. T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
3. O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.
4. R. Dekker. The importance of having data-sets. 2006.
5. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.
6. J. A. Gulla, J. E. Ingvaldsen, A. D. Fidjestl, J. E. Nilsen, K. R. Haugen, and X. Su. Learning user profiles in mobile news recommendation. pages 183–194, 2013.
7. B. McFee, T. Bertin-Mahieux, D. P. Ellis, and G. R. Lanckriet. The million song dataset challenge. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 909–916. ACM, 2012.
8. O. Ozgobek, J. A. Gulla, and R. C. Erdur. A survey on challenges and methods in news recommendation. In *In Proceedings of the 10th International Conference on Web Information System and Technologies (WEBIST 2014)*, 2014.
9. A. Said, S. Berkovsky, and E. W. De Luca. Group recommendation in context. In *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, CAMRa '11, pages 2–4. ACM, 2011.
10. K. Verbert, H. Drachsler, N. Manouselis, M. Wolpers, R. Vuorikari, and E. Duval. Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 44–53, New York, NY, USA, 2011. ACM.
11. S. R. Wolfe and Y. Zhang. Interaction and personalization of criteria in recommender systems. In *User Modeling, Adaptation, and Personalization*, pages 183–194. Springer, 2010.
12. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32. ACM, 2005.