International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution

Evaluating Recommender Systems

Zied Zaier
UQAM University
zaier.zied@courrier.uqam.ca

Robert Godin
UQAM University
godin.robert@uqam.ca

Luc Faucher
UQAM University
faucher.luc@uqam.ca

Abstract

Recommender systems are considered as an answer to the information overload in a web environment. Such systems recommend items (movies, music, books, news, web pages, etc.) that the user should be interested in. Collaborative filtering recommender systems have a huge success in commercial applications. The sales in these applications follow a power law distribution. However, with the increase of the number of recommendation techniques and algorithms in the literature, there is no indication that the datasets used for the evaluation follow a real world distribution. This paper introduces the long tail theory and its impact on recommender systems. It also provides a comprehensive review of the different datasets used to evaluate collaborative filtering recommender systems techniques and algorithms (EachMovie, MovieLens, Jester, BookCrossing, and Netflix). Finally, it investigates which of these datasets present a distribution that follows this power law distribution and which distribution would be the most relevant.

1. Introduction

The exponential increase in sources of information available in the World Wide Web has created a challenge to find ways to identify relevant information and knowledge that it contains. Indeed, according to a new research, 93% of information produced worldwide is in digital format and the amount of data exceeds 161 exabytes, while more than 1 billion people around the world are connected to the internet making around 213 million queries each day looking for relevant information resources [1, 2].

In recent years, recommender systems have become part of the solution to the information overload problem faced by consumers on the net. Collaborative filtering recommender systems provide users with suggestions regarding which information is most relevant to them. These systems have proved to be some of the most successful techniques to help people find contents that are most valuable to them [3].

Recommender system plays an important role in e-commerce applications. *Amazon* recommends all kinds of products (movies, music, books, etc.) to consumers; *Netflix*, *moviecritics*, and *Cinemax* propose movies to users; *TiVo* recommend TV shows and movies; *iTunes*, *Pandora, Last.fm* and *Rhapsody* advise music; *Trabble* suggests restaurants; *Expedia* and *Travelocity* recommend travels; *CNN* advises News; *Jester* proposes jokes; *Google* suggests web pages [4]. These applications show that the sales follow what is called a *power law distribution*. Indeed, a small number of items sell a lot and a large number of items sell a little [5]. However, the economic impact of these items that sell little is important [6].

For the providers, it is important to ensure a good coverage of the information sources and products. This coverage can be achieved using collaborative filtering recommender systems [6]. With the popularity of these systems, a considerable number of recommendation algorithms have been developed [3]. A number of datasets exist which help researchers to evaluate these algorithms (*EachMovie*, *MovieLens, Jester, BookCrossing, and Netflix*, etc.) [7]. In this paper, we study the distribution of these datasets and its impact on recommendation quality.

The paper is organized as follows. In Section 2 we briefly describe the theory of the *Long Tail* and its importance for recommender system applications. We then give an overview of available datasets used to evaluate recommender systems techniques and algorithms in Section 3. We compare the performance of different recommendation techniques based on different dataset distributions in Section 4. Finally, in Section 5, we make a brief concluding remark and mention some future directions for research.

2. Long Tail concept

This section introduces the reader to the *Zipf's law* and its extension. It also describes in detail the *Long Tail* concept and the different domain of study where this *power law distribution* is observed. Finally, it highlights the importance of the *Long Tail* markets to collaborative filtering recommender systems.

978-0-7695-3406-0/08 $25.00 © 2008 IEEE
DOI 10.1109/AXMEDIS.2008.21

2.1. Zipf-Mandelbrot law

The *Long Tail* usually refers to the fact that a small number of events happen with very high frequency and a large number of events happen with very low frequency. George Kingsley Zipf, a Harvard linguistics professor, observed that the frequency of use of the word in a random corpus of writing is generally inversely proportional to its rank. Thus, if the most frequent word appears 120 times in a text, the second most frequent word will appear 60 times, the third most frequent word will appear 40 times and so on [5].

Zipf proposed the following empirical *power law distribution*:

$$f(r) = k / r^{\beta}$$

Where $f(r)$ is the frequency, r is the *rank* and k is a constant for the corpus and $\beta = 1$. This is what is now known as *Zipf's law*. The following figure represents the anatomy of the *Long Tail* (see figure 1). It should be noted that a curve that follows a *Zipf's* distribution has:

- ✓ A few elements that occur with high amplitude (the left tail in the curve in figure 1);
- ✓ An intermediate number of elements occur with middle amplitude (the middle part of the curve in figure 1);
- ✓ A huge number of elements occur with low amplitude (the right tail in the curve in figure 1).

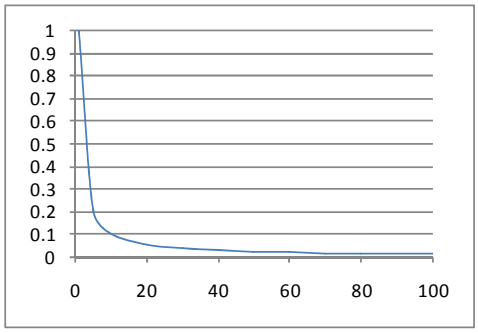

Figure 1. Anatomy of the Long Tail.

A generalization from the *Zipf's law* is the *Zipf-Mandelbrot law*, proposed by the mathematician Benoît Mandelbrot. The frequency is computed using the following equation:

$$f(r) = k / (a + br)^{c}$$

Where $f(r)$ is the frequency, r is the *rank* and *k, a, b and c* are constants for the corpus.

2.2. Recommendation and Long Tail

The observations of the curve in figure 1 are found in several other fields of study. Indeed, the *Long Tail* practically is everywhere, from words to web pages, from politics to public relations, from lands to cities, from earthquakes to DNA sequences, and from music sheet to college sports [5, 6, 8].

In e-commerce applications, Chris Anderson, editor-in-chief of *Wired Magazine*, was the first to point out that there is a strong demand for unknown and obscure products [5]. For instance, for Rhapsody, 40% of the sales do not come from the 39.000 most sold tunes. 21% of the sales of Netflix are done by movies which do not make part of the 3000 most rented one. 20% of Amazon sales do not come from books of the 130.000 most popular one [6].

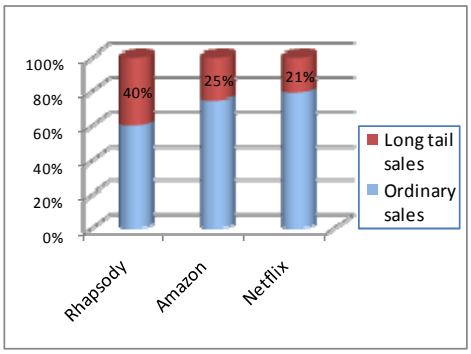

Figure 2. Total sales [6].

It is evident that there is a very important market in the small disparate elements which seem unimportant if they are taken separately, but when they are added together, give a considerable impact (see figure 2). As a result, there is now an important market for the *Long Tail*. Indeed, *Wal-Mart* has found ways to exploit the *Long Tail* for the supply chain; *Google* is employing the *Long Tail* for advertising; *Microsoft* is using the *Long Tail* for the *Xbox* video games online distribution [5, 6, 8].

One of the rules proposed by Anderson to address the *Long Tail* market is to use recommendations to drive demand down the Long Tail [5]. In other words, collaborative filtering recommender systems must be able to cover most obscure items.

3. Recommender system datasets

Recommender system researchers use a number of different datasets for evaluating the performance of the collaborative filtering recommendation algorithms [7]. In this paper, our attention is focused on the choices of the collections. One of

our concerns was to choose datasets that are easily accessible, that have been used in the past, and that are likely to be used in the future. We hence opted for the *MovieLens, Jester, BookCrossing,* and the *Netflix* datasets, which exhibit these three properties.

In this empirical evaluation on collaborative filtering recommender systems datasets, we ran experiments where, successively, we began by computing the number of rating for each item of a given dataset, then we ranked the collected results, and we finally draw the curve for the rating distribution. This evaluation uses respectively the 8 *MovieLens* collections, the 3 *Jester* datasets, the *Netflix* dataset, and the *BookCrossing* collection.

3.1. MovieLens datasets

The popular *MovieLens* datasets (www.grouplens.org) have already been the subject of numerous empirical investigations [2, 4, 9]. These publicly available datasets contain explicit ratings about movies, demographic information about users (age, gender, occupation, zip code), a brief description of the movies (Title, release year, genres.), and have very high densities. In these datasets, there are movie ratings raging from 1 to 5 made by 943 users (only the users that have rated at least 20 movies are considered.) on 1682 movies.

Several datasets are available. The first one consists of the full dataset of 100,000 ratings. This set is called *U*. The next five are divided into 10,000 ratings. These sets are called *U1* to *U5*, respectively. The two remaining sets are divided into 10,000 ratings. These sets are called *UA*, and *UB*, respectively.

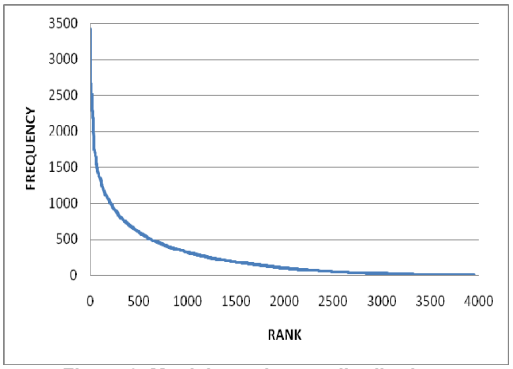

Figure 3. MovieLens dataset distribution.

All 8 *MovieLens* collections demonstrate a significant equivalence in the results. The result shown in figure 3 is the one collected using the entire *MovieLens* dataset. In the curve, we observe that, a few movies that are highly rated (the left tail in the curve in figure 3), an intermediate number of movies with middle number of rates (the middle part of the curve in figure 3), and a huge number of movies that are lowly rated (the right tail in the curve in figure 3). Thus, the *MovieLens* datasets follow a *power law distribution*.

3.2. Jester datasets

The *Jester* dataset released by *Ken Goldberg* from *Jester* Joke Recommender System website (www.ieor.berkeley.edu/~goldberg/jester-data/) contains anonymous ratings of 100 jokes from 73,496 users**.** Ratings are real values scaling from -10.00 to +10.00 [10]. Some users end up reading and rating all the jokes, so Jester is much denser than the other datasets we considered. Almost all users have rated those jokes.

This collection is divided on three different datasets:

- ✓ *Jester-data-1*: Contains the rating from 24,983 users who have rated 36 or more jokes.
- ✓ *Jester-data-2*: Contains the evaluation from 23,500 users who have rated 36 or more jokes.
- ✓ *Jester-data-3*: Contains the data from 24,938 users who have rated between 15 and 35 jokes.

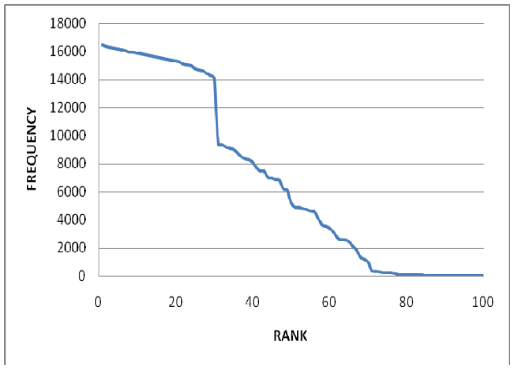

Figure 4. Jester dataset distribution.

Figure 4 presents only the *Jester-data-1* dataset distribution, since the result shown for the *Jester-data-1, Jester-data-2, and Jester-data-3* are similar. We can notice a modest number of jokes that are lowly rated (the right tail in the curve in figure 4). In addition, it shows that a huge number of jokes that are highly rated (the left side of the curve in figure 4). Thus, the *Jester* datasets do not follow a *power law distribution*.

3.3. Netflix dataset

The *Netflix* dataset is currently the subject of a considerable number of empirical studies. In 2006, the online movie renter, announced *The Netflix Prize* (www.netflixprize.com). An award of one million dollars to the first person who can attain a

defined accuracy goal in recommending movies. *Netflix* released a huge dataset including more than 100 million movie ratings scaling from 1 to 5, which were performed by about 480,189 real anonymous users on 17,770 movies. Also, rating dates and movie names are provided [11].

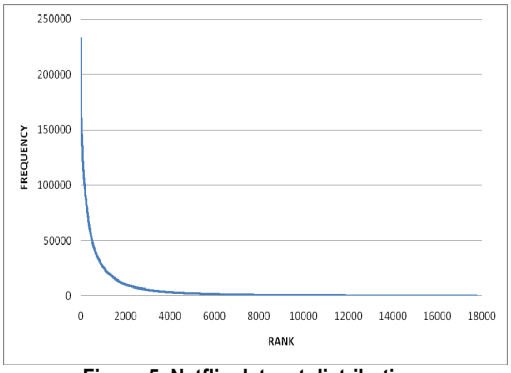

Figure 5. Netflix dataset distribution.

Figure 5 illustrates the distribution for *Netflix* dataset. We note that, as was for the *MovieLens* collections, the *Netflix* dataset shows a distribution that follows the anatomy of the *Long Tail*.

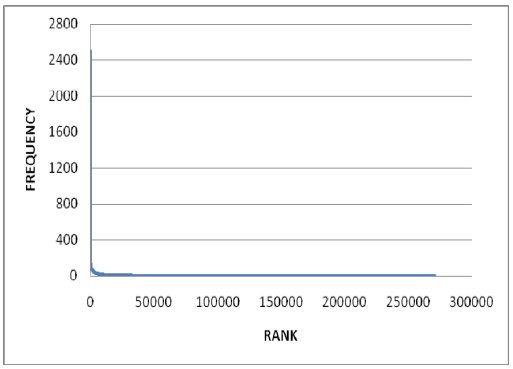

Figure 6. BookCrossing dataset distribution.

3.4. BookCrossing dataset

The *BookCrossing* (*BX*) (www.informatik.uni-freiburg.de/~cziegler/BX/) dataset was collected in 2004 by *Cai-Nicolas Ziegler* from the *Book-Crossing community* (www.bookcrossing.com). It contains 278,858 anonymous users providing 1,149,780 ratings about 271,379 books, expressed on a scale from 1 to10. This dataset, as was for the *MovieLens* collections, contains demographic information about users (age, location.) and a brief description of the books (book title, book author, year of publication, publisher.) [12].

Figure 6 shows the *BookCrossing* dataset distribution. We observe that, as was for the *MovieLens* collections and *Netflix* dataset, the BookCrossing dataset distribution follows the *power law distribution*.

4. Experimental evaluation

We used a recommendation algorithm based on neighborhood discrimination For our experiments to evaluate the impact of the dataset distribution on recommendation quality, [4]. In this section, we present a brief introduction of our experimental data collection, the evaluation metric, the experimental procedure followed by the experimental results and discussion.

4.1. Datasets

Several datasets were chosen for the evaluation. We hence first opted for the popular *MovieLens* datasets, which exhibits a *power law distribution* (see section 3.1). These datasets have been the subject of empirical investigation for the neighborhood discrimination recommendation approach [4]. Also, we chose the *Jester* datasets, which shows a distribution that does not follow a *power law distribution* (see section 3.2).

4.2. Evaluation metrics

As used in the empirical investigation for the neighborhood discrimination recommendation approach [4], we used for our experiments a widely popular statistical accuracy metric named *Mean Absolute Error (MAE)*, which measures the average magnitude of the errors. The *MAE* evaluates the quality of the prediction.

$$\left|\overline{E}\right| = \frac{\sum_{b_k \in B_i} \left| r_i(b_k) - w_i(b_k) \right|}{\left|B_i\right|}$$

where $r_i(b_k)$ is the real rating value on the item b_k, $w_i(b_k)$ is the rating value predicted by the system, and B_i is the number of items. Also, we use *Recall,* which measures the percentage of relevant items a recommendation system predicts. *Recall* is an indicator of the quality of the neighborhood.

$$R = \frac{N_{rc}}{N_r}$$

where N_{rc} is the number of relevant items predicted by the system, and N_r is the number of real relevant items. Finally, we used *Rating Coverage,* which measures the percentage of items for which a recommendation system can provide

predictions. This metric, as for the *Recall,* evaluates the quality of the neighborhood.

$$C = \frac{N_r - N_{rn}}{N_r}$$

where N_r is the number of real relevant items, and N_{rn} is the number of real relevant items that cannot be predicted by the system.

4.3. Experimental Protocol

The collaborative filtering approach was used in our experiments for the evaluation of the impact of dataset distribution. We did not include hybrid collaborative filtering and demographic hybrid collaborative filtering approaches, as was for the previous neighborhood discrimination recommendation approach study [4], because the Jester datasets does not provide content and demographic information. The collaborative filtering approach recommends items to a given user based on the ratings of other neighbors. Once a neighborhood of users is formed, the ratings dataset is represented as a *UserVSItem* matrix with users for rows and items for columns. The similarity between a user u_i and a user u_j can be computed using the *Pearson Correlation Coefficient* similarity measure based on the ratings made by the users. The result of this step is a *UserVSUser* matrix that defines the correlation between all users. In our work, the prediction of the user u_i on the item r_j is computed using the following equation [5]:

$$\overline{eval}(u_i) + \kappa \sum_{u \in U} (w(u_i, u)(eval(u, r_j) - \overline{eval}(u)))$$

Where $\overline{eval}(u_i)$ is the average evaluation of the user u, $eval(u, r_j)$ is the evaluation made by the user u in the item r_j, and the weight $w(u_i,u)$ is the similarity between the user u_i and the user u given by the *UserVSUser* matrix. We let, in our case, $\kappa = 1/\sum_{u \in U} w(u_i, u)$.

Also, we adopt the experimental procedure that was given for the previous neighborhood discrimination recommendation approach study [4]. In fact, in this empirical evaluation on collaborative filtering, we ran experiments where, successively, we began by a training cycle on a given training base for each one of the datasets, followed by a test cycle on the test base in which we choose a subset of neighbors with 20% of users with closest similarity and 20% of users with farthest one to generate a set of predictions, and to finish, we computed the metrics we presented in the above section and we collected the results. Subsequent to this operation and until we arrived to the total training base (943 users for *MovieLens* dataset and 100 users for *Jester* dataset), 50 new neighbors for *MovieLens* dataset and 5 new neighbors for *Jester* dataset are added to the training set and another training/test cycle is performed.

4.4. Results and discussion

All 8 *MovieLens* collections demonstrate a significant equivalence in the results. However, the results shown in figures 7, 9, and 11 are the ones collected using U1 MovieLens dataset for *Pearson Correlation Coefficient* similarity measure. Also, figures 8, 10, and 12 present only the *Jester-data-1* dataset distribution for *Pearson Correlation Coefficient* similarity measure, since the result shown for the *Jester-data-1, Jester-data-2, and Jester-data-3* are significantly equivalent. Each one of the metrics is applied to the collaborative filtering approach for the two datasets using a subset of neighbors in the extremities with the closest and the farthest similarities.

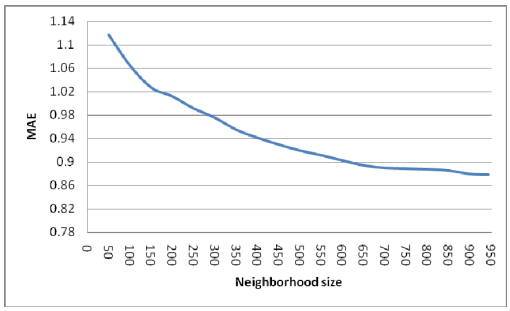

Figure 7. Collaborative filtering MAE as neighborhood size grows for MovieLens dataset [4].

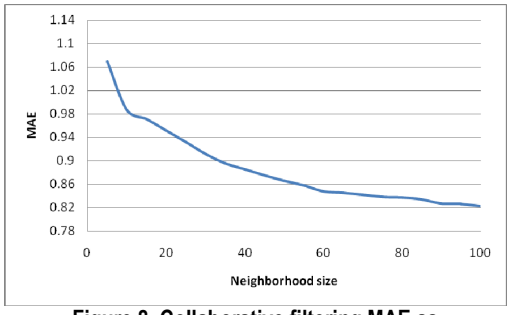

Figure 8. Collaborative filtering MAE as neighborhood size grows for Jester dataset.

Figures 7 and 8 present the prediction accuracy for the two datasets. We can note a modest difference between the MovieLens and Jester datasets. However, the *Jester* dataset demonstrates a considerable advantage.

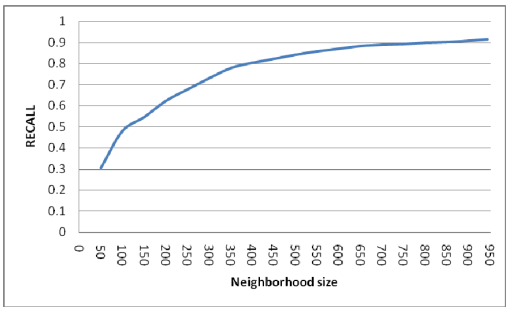

Figure 9. Collaborative filtering Recall as neighborhood size grows for MovieLens dataset [4].

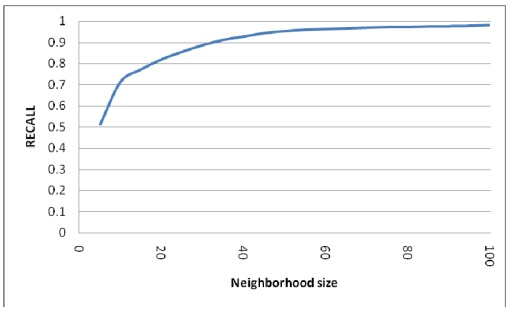

Figure 10. Collaborative filtering Recall as neighborhood size grows for Jester dataset.

Figures 9 and 10 illustrate the recommendation quality for the *MovieLens* and *Jester* datasets. We can observe that the two datasets are equivalent. However, the *Jester* dataset is the first to get to the optimal value. In addition, we can notice that, as was for the *Mean Absolute Error*, the *Jester* dataset shows a substantial advantage.

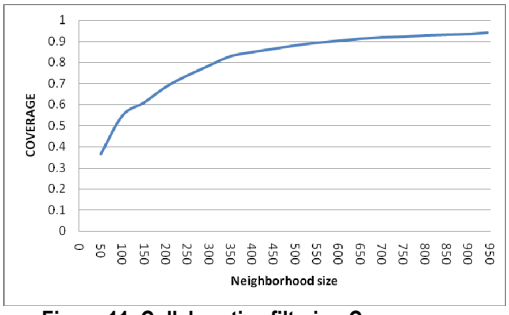

Figure 11. Collaborative filtering Coverage as neighborhood size grows for MovieLens dataset [4].

The curves showed in Figures 11 and 12 present the quality of the neighborhood. These figures, as for the *Recall*, reveal that *Coverage* is strongly equivalent between the *MovieLens* and *Jester* datasets. Also, we note that the *MovieLens* dataset reach their optimal value after the *Jester* dataset. In addition, we observe that, as was for the *Mean Absolute Error* and *Recall*, the *Jester* dataset shows a substantial advantage. Finally, we note that the curves in figures 9 to 10 and figures 11 to 12 are quite the same. This can be explained by the fact that these two metrics evaluate the quality of the neighborhood.

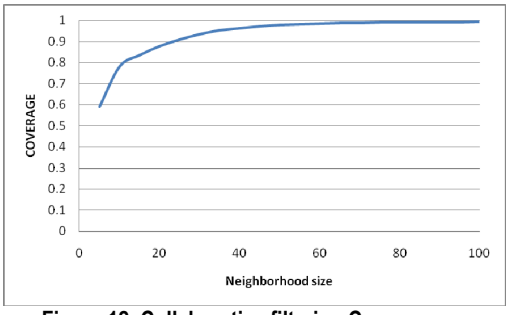

Figure 12. Collaborative filtering Coverage as neighborhood size grows for Jester dataset.

5. Conclusion and future directions

Recommender systems are, without contest, becoming increasingly popular. This success is probably due to the assistance provided by these systems to the users in their information search, by making it more personalized and at the same time more optimal. Moreover, these systems are not limited to e-commerce domain only, but are currently used in several other application domains. For example, *Abyss eurocom* uses recommender systems as guides to visit the virtual museum; *iKarma* and *eBay* exploit them to create user's reputation; *RAZOR2* and *CASSANDRA* use them as a spam filter; *NewsGator* and *Illumio* exploit them to recommend blog articles; *ColFi* uses them for a Dating Service; *Lifestyle* exploits them in the interactions of the avatar and the user [13-17]. However, it is evident that the *Long Tail* market is very important to collaborative filtering recommendation systems. Indeed, in most of the applications making recommendation, small number of items sale very high and a large number of items sale low [5, 6, 8]. To date, however, there has been no study that tries to investigate the coverage of these obscure items.

In this paper, we provided a comprehensive review of the different datasets used to evaluate collaborative filtering recommender systems techniques and algorithms. We also studied the distribution of these datasets and its impact on recommendation quality. Our experiment shows that *MovieLens, BookCrossing,* and *Netflix* datasets follow a *power law distribution*. Moreover, datasets that not follow a *power law distribution* (*Jester*) seems to converge much faster than the ones that follow it (*MovieLens*). Of interest for us is *MovieLens* datasets that have been subject of several empirical researches. These datasets offer additional data that are useful for recommender systems by taking advantage of the content about

the movies (Title, release year, genres.) and the demographic information about the users (age, gender, occupation, zip code). Some additional content can also be added by loading movies description (Pilot, actors, countries.) from *Internet Movie DataBase* (*IMBD,* www.imdb.com) and associating it to the already existing data [2, 4].

While our study investigates a single-criterion rating datasets like most of the current collaborative filtering recommender systems, many commercial applications have begun using multicriteria rating. *Yahoo! Movies* asks users to provide four specific criteria about the movie: story, acting, direction, and visuals (http://movies.yahoo.com); *Zagat Survey* provide three criteria to evaluate a restaurant: food, decor, and service (www.zagat.com); *Circuitcity* offer four criteria for item ratings display, performance, battery life, and cost (www.circuitcity.com). Multicriteria rating is considered as an added-value to current recommender system [18]. In future work, we will investigate new recommendation techniques that take advantage of multicriteria rating to improve recommendation quality. Finally, due to the nature of the information involved in recommender systems, we will try to develop anonymity and encryption techniques to protect individual data privacy in distributed recommender systems [10, 19].

6. References

[1] J. K. Kim, H. K. Kim, and Y. H. Cho, "A user-oriented contents recommendation system in peer-to-peer architecture," *Expert Syst. Appl.,* vol. 34, pp. 300-312, 2008.

[2] Z. Zaier, R. Godin, and L. Faucher, "Recommendation Quality Evolution Based on Neighborhood Size," in *Third International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution. AXMEDIS '07* Barcelona, Spain, 2007, pp. 33-36.

[3] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *Transactions on Knowledge and Data Engineering,* vol. 17, pp. 734-749, 2005.

[4] Z. Zaier, R. Godin, and L. Faucher, "Recommendation Quality Evolution Based on Neighbors Discrimination," in *MCETECH Conference on e-Technologies* Montreal, 2008, pp. 148-153.

[5] C. Anderson, "The Long Tail," *Wired Magazine,* vol. 12, 2004.

[6] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*: Hyperion, 2006.

[7] L. J. Herlocker, A. J. Konstan, G. L. Terveen, and T. J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.,* vol. 22, pp. 5-53, 2004.

[8] A. Elberse, "Should You Invest in the Long Tail?," in *Harvard Business Review*, 2008.

[9] B. N. Miller, J. A. Konstan, and J. T. Riedl, "PocketLens: Toward a personal recommender system," *ACM Trans. Inf. Syst.,* vol. 22, pp. 437-476, 2004.

[10] J. Canny, "Collaborative filtering with privacy via factor analysis," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* Tampere, Finland: ACM, 2002.

[11] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* San Jose, California, USA: ACM, 2007.

[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* Chiba, Japan: ACM, 2005.

[13] A. Gray and M. Haahr, "Personalised, Collaborative Spam Filtering," 2004.

[14] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury, and P. O. Boykin, "Collaborative Spam Filtering Using E-Mail Networks," *Computer,* vol. 39, pp. 67-73, 2006.

[15] V. V. Prakash and A. O'Donnell, "Fighting spam with reputation systems," *Queue,* vol. 3, pp. 36-41, 2005.

[16] L. Brozovsky, "ColFi - Recommender System for a Dating Service," in *ZNALOSTI 2007 conference* Ostrava, 2007.

[17] S. Ujjin and P. J. Bentley, "Building a Lifestyle Recommender System," in *The Tenth International World-Wide-Web Conference.*, 2001.

[18] G. Adomavicius and K. YoungOk, "New Recommendation Techniques for Multicriteria Rating Systems," *Intelligent Systems,* vol. 22, pp. 48-55, 2007.

[19] S. K. Lam, D. Frankowski, and T. J. Riedl, "Do You Trust Your Recommendations? An Exploration of Security and Privacy Issues in Recommender Systems," in *Emerging Trends in Information and Communication Security*, 2006, pp. 14-29.