ONLINE EVALUATION OF RECOMMENDER SYSTEM WITH MOVIELENS DATASET

Asmir Handžić

Faculty of Information Technologies Pan-European University "APEIRON" (PhD Student)
E-mail: asmir.handzic@handzic-it.com

Case study

DOI: **10.7251/JIT16020H** UDC: **004.652.8**

Abstract: The purpose of this paper is to explore the advantages of recommender systems based on the matrix factorization in respect to classical first neighbor recommender systems to real users through A/B test, as these studies are more significant. The results presented in this paper confirms the hypothesis that the recommender systems based on the models of matrix factorization are superior in relation to classical nearest-neighbor recommender systems.

Keywords: Recommender systems, online evaluation, MovieLens, A/B test

INTRODUCTION

The recommender systems are created with a purpose of assisting users in dealing with problems arising from information overload by building a prediction model which will evaluate the preference rate that a user will give to each recommended item [17].

The recommender systems have become extremely popular in short amount of time, as in the research as well as in the business sector. By 1996, several companies had marketing recommender systems in place ("Agents Inc" originating in "Ringo" project, and "Net Perceptions" originating in Group Lens project), and the first exploratory workshop in this field was held in Berkley in March 1996.

Since its beginning, this area has had a significant progress, both in science and in business application so that today the recommender systems are built-in in many commercial and other applications, many articles and books [9], [18], have been published, many universities are offering courses in this field, and there is an annual conference dedicated to this subject (the ACM Recommender Systems Conference). To attest today's popularity to use these systems in the e-commerce is the fact that out of 100 % series/sitcoms and movies which Netlifx users choose more than 75 % has been selected based on the recommendation from the system.

Research topic

When introducing the recommender systems there are certain problems: the problem of cold start and the selection of appropriate algorithm. The problem of cold start appears in cases where there is not enough data on a new user or a new item, therefore the system cannot create the prediction of preference. The other problem is related to the selection of appropriate algorithm which will, within the given system or business case, offer the good-quality recommendations to users.

Having in mind that the recommender systems are modern technology and are constantly evolving, there are datasets and open source algorithms which are developed by the academic community and are offered for the purpose of further scientific research.

Such dataset and the recommender system is used by a portal/site pogledajfilm.info, which is newly in-

troduced non-commercial portal for the purpose of scientific research and to provide recommendations as to which movie to use to users. Using the MovieLens 10M dataset the problem of cold start at the aforementioned portal/site has been dealt with (10 000 000 votes for 10 000 movies by 72 users). Still there is a problem of selecting the appropriate algorithm from a variety of modern and complex algorithms from MyMedia lite package.

The process of evaluation is highly important when validating/selecting the recommender systems. The evaluation can be offline and online. Offline evaluation is generally conducted when the system has not yet been implemented and we have more algorithm candidates, while online evaluation with implemented varieties of the system which are valid with real users of the system, where results are recorded and compared. The real value of these systems is online evaluation, where the system is used by the real users performing the real actions. In some cases these experiments are risky. For example if the test system offers irrelevant recommendations this can discourage test users and turn them from ever using this system. Therefore the experiment can have negative effects on the system, which may be undesirable in commercial applications. For this reason it is best to do the online evaluation after offline evaluation which will confirm that the candidates were reasonable.

In addition, most of the research results are given on bases of offline evaluation, less on the online, especially from the commercial systems which do not provide results due to competition. This further supports the need for the research conducted in this paper, as it will be conducted on the real users on the real (online) system, which is the best indicator of the system's quality. The results of this research will be used to support the decision on which of the algorithms will be used as the main algorithm in making the personalized recommendations to the users of a portal pogledajfilm.info, and which algorithm is better for the given dataset, that is the given case, which can help someone who uses the same dataset as a bases for research but also for commercial purposes (commercial systems). On basis of results from offline evaluation research, the online evaluation will be conducted with the real users of algorithms which have better results of offline tests.

Various studies and literatures [13] give advantages to recommender systems that are based on matrix factorization that is the dimensionality reduction, especially for the sake of performance and an increase in prediction's accuracy. This has been demonstrated with Netflix Prize competition [13], where they have been far superior, and it is a movie domain as well as a domain which the portal pogledafilm.info is using. With the increase of matrix, the classical collaborative filtering systems suffers from synonymy, performance declines (more computer power is needed with the increase of matrix) and sparsity, hence the prediction is logically worse. Systems based on dimensionality reduction, according to the latent factors research are easier at revealing connections between users and items and, with the more compact matrix, have better performances.

On basis of the problems and research objectives we can identify hypothesis:

H1. Recommender systems based on the matrix factorization model are superior compared to classical nearest-neighbor recommender systems, and expecially with an increase of matrix.

The main objective of this study was to choose the good-quality algorithm of the recommender systems out of many which are available online, based on the rating by the users, since such studies are less common but more important bearing in mind validity of the system, and that the research results serve as basis for the selection of the main algorithm, which will be used at pogledajfilm.info. Furthermore, the aim of this study is to test the hypothesis that the models based on the matrix factorization (matrix reduction) are far superior to the classical nearest-neighbor algorithms (user-user, item-item algorithm). Aforementioned research results [13] which give advantages to the matrix factorization models are mostly based on using offline evaluation. The offline metrics only assess the ability to "recommend" items that have already been consumed or rated. Real recommenders should usually be suggesting new items not already known to the user. Hence, something with low of-

fline metrics might actually be better at finding new items of interest. Therefore the main research method is the online evaluation.

METHODOLOGY

Offline test

Since the algorithms will be tested on the accuracy of predicting of how much a user will like the movie (by appointing 1-5 stars) the following tests will be used:

- RMSE (root mean square error)
- MAE (mean absolute error)

In essence, both tests indicated the deviation of the mean value that is how much of an error/deviation the system has made. The system is better if MAE and RMSE are smaller.

Online test

Online test is performed on the actual system users, and during the work A/B test will be used. A/B test operates in a way to implement two recommendation variants; the users are given recommendations for both of them, than the results are compared through (through feedback information, site visitation, sales, etc.)

The testing will be conducted on portal/site pogledajfilm.info, where registered users, who voted for at least 20 movies, will be given two options of the personalized recommendation (in forms of 10 movies with the highest prediction) and will be asked to rate the option which appeal to them more, not knowing which system algorithm (candidate) is the in question. The minimum of two candidate algorithms will be taken for the online testing, one of the most contemporary representative of algorithms based on the matrix factorization (Biased Matrix Factorization [13]), and the other classical closest-neighbor algorithms, Slope One[14]. Slope One is a simple implementation of item-item recommender system, which is efficient and accurate as many more complicated algorithms, and will therefore be interesting to research as an alternative to A/B test.

Recommender systems algorithms, which will be used, are part of My MyMedia lite tool. It is an open source tool which is available for use and development for non-commercial purposes and contains the most contemporary algorithms of collaborative filtering. MyMedia lite system comes as an open source based on C# programming language or as a batch file, which generates text file with the prediction/recommendation. During the work, the batch file "RatingPrediction.exe" will be used, which offers prediction (users_id, movie_id, rating) based on the standard of explicit rating from 1-5.

RESULTS

Results of offline test

Results of offline test have confirmed the thesis that the matrix factorization models are superior to classical nearest-neighbor technics. This is visible in two different offline tests, that is, with the increase of difference reciprocated with the increase of dataset.

BiasedMatrixFactorization:
RMSE:0,95
MAE:0,74
MatrixFactorization:
RMSE:0.97
MAE:0.75
Slope One:
RMSE:0.95
MAE:0.74
ItemKNN
RMSE:0.94
MAE:0.74

Figure 1. Results for offline validation on the 100K dataset

Both test methods have low RMSE and MAE and are close to each other at MovieLens 100K (100 000 votes) dataset, although at a significantly smaller dataset from the base one, Biased Matrix Factorization method provides significantly better performance. Test indicates that the classical Item KNN algorithm give good predictions at a smaller dataset (alongside others) but is the slowest in the performance. Classical Matrix Factorization algorithm is slightly worse than the advanced Biased Matrix Factorization algorithm which was expected.

Metod	RMSE	MAE
GlobalAverage	1.117	0.934
UserAverage	1.036	0.827
ItemAverage	0.983	0.783
SlopeOne	0.902	0.712
UserItemBaseline	0.908	0.719
ItemKNNPearson	0.871	0.683
FactorWiseMatrixFactorization	0.860	0.673
MatrixFactorization	0.857	0.675
BiasedMatrixFactorization	0.854	0.674
SVDPlusPlus	0.851	0.668

Figure 2. Results for offline validation on the 1M dataset

On a larger dataset (1 million of ratings) again we see confirmation of the thesis that is the increase of matrix increases advantages of the Biased Matrix Factorization method. In this example we have for 0,003 slightly less error of SVD PlusPlus method in comparison to Biased Matrix Factorization, however SVD PlusPlus will not be tested in the online evaluation because Biased Matrix Factorization through Netflix prize is more accurate, which is proved (Netflix prize award-given methods), and is more contemporary method which also incorporates biases (circumstances and prejudice). If we take the indicator of precision of Biased Matrix Factorization model from this example (Figure 1) and compare it to Netflix Prize data, Netflix's big award demanded RMSE to be 0.8563, while here RMSE is 0.854, hence we speak of very precise methods at the given dataset, at least as far as offline evaluation is concerned. Considering results of 100k and 1M datasets there is no need to make an evaluation with 10M dataset, but the online evaluation will be conducted, which is at the same time the main method of the research in this paper.

Results of online test – A/B test

The evaluation was carried out on 19 registered users in the system, from user ID 71568 to user ID 71586. The users had to have at least 20 explicit ratings. The users were then given, based on the outcome of their voting, two generated options of recommendations on 10 movie titles which had had the highest assumption that the user would personally like (with the highest prediction). The option 1 was personalized recommendation generated by Biased Matrix Factorization model, and option 2 by Slope One model. Than the user should have voted for the option which was more appealing to him, and to grade it from 1 to 5. Out of 19 participants in the analyses, 15 participated till the end of the analyses, which leads to conclusion that the users find it hard to explicitly participate in the evaluation of the system. Out of 15 participants, 10 voted for option number 1, and 5 for option number 2, thus individually graded option number 1 with the average grade of 4.11. In addition it was noted that the Matrix Factorization model (option 1) has a far better performance than Slope One at the given dataset. Option One training time is about 5 minutes, while option 2 takes 39 minutes.

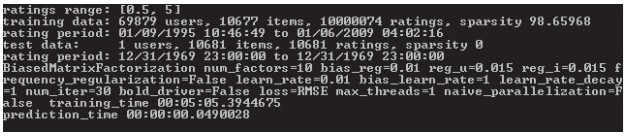

Figure 3. Biased Matrix Factorization training time

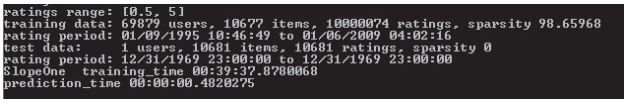

Figure 4. Slope One training time

CONCLUSION

Based on the research results of this work, confirmation of the thesis that the recommender systems based on the Matrix Factorization models provide better preferences to users than the classical nearby-neighbor recommender system models can be concluded. Furthermore, it is noted that the Matrix Factorization model (option 1) has far better performance than Slope One at a given dataset. Training time of the option 1 is about 5 minutes, while the option 2 takes 39 minutes, and here we can conclude how significantly advanced the Matrix Factorization Model is and that in the following phase of development of portal pogledajfilm.info the option 1 will be used to generate personalized predictions to site users. In addition, from the research conducted on the users, it can be also concluded that the users will harder agree to participate in the evaluation of the system explicitly, and this may be a reason why these studies are few and far in between but of a very high importance because they provide the overall picture on the system evaluation.

The study was carried out in the same domain as was the domain used as a basis to form the hypothe-

sis, which is to recommend the movies, but not using the same dataset. The recommendation is for further research study is to be conducted, to repeat the same study but in some other domain, regardless whether it is commercial or non-commercial.

Biography:
Asmir Handžić Senior Teaching Assistant at University of Bihac. PhD student in Informatics from the Faculty of Information technology of the University ''APEIRON'' Banja Luka and MSc in Informatics from the Faculty of Information technology of the University ''Dzemal Bijedic'' Mostar. Research activity focused on recommender systems and human-computer interaction.
Faculty of Information Technologies Pan-European University "APEIRON" (PhD Student).

REFERENCES

[1] Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction., 331-370

[2] Bennet, J. and Lanning, S. (2007). "The Netflix Prize," KDD Cup and Workshop; www.netflixprize.com.

[3] Claypool, M., et al. (2001). Implicit interest indicators. In Proceedings of the 6th international conference on Intelligent user interfaces (IUI '01). ACM, New York, NY, USA, 33-40

[4] Eckhardt, A., Vojtáš, P. (2009). Combining Various Methods of Automated User Decision and Preferences Modelling, MDAI '09 Springer-Verlag Berlin, Heidelberg, 172-181.

[5] Eckhardt, A., Vojtáš, P. (2009). How to learn fuzzy user preferences with variable objectives. In proc. Of IFSA/EUSFLAT Conf. 2009: 938-943

[6] Freyne, J., et al. (2011). Recipe recommendation: accuracy and reasoning. In Proceedings of the 19th international conference on User modeling, adaption, and personalization (UMAP'11). Springer-Verlag, Berlin, Heidelberg, 99-110

[7] Gunawardana, A., Shani, G.A. (2009). Survey of Accuracy Evaluation Metrics of Recommendation Tasks, Jorunal of Machine Learning Research 10, 2935-2962

[8] Hu, Y., et al. (2008). Collaborative Filtering for Implicit Feedback Datasets. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08). IEEE Computer Society, Washington, DC, USA, 263-272

[9] Jannach, D., et al. (2011). Recommender systems: an introduction. Cambridge University Press, New York

[10] Jawaheer, G., et al. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. In Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '10). ACM, New York, NY, USA, 47-51

[11] Konstan, J. and Riedl, J. (2012). Recommender systems: from algorithms to user experience. User Modeling and User-Adapted Interaction,, 22, 101-123

[12] Konstan, J., et al. (1997). Applying collaborative filtering to usenet news: the GroupLens system. Commun. ACM 40(3), 77–87

[13] Koren, Y., et al. (2009). Matrix Factorization Techniques for Recommender Systems, IEEE Computer Society

[14] Lemire, D. and Maclachlan, A. (2005). Slope One Predictors for Online Rating-Based Collaborative Filtering. SIAM Data Mining

[15] Maxwell, F. H. and Konstan, A. J. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

[16] Mukherjee, R., et al. (2003). A Movie Recommendation System\—An Application of Voting Theory in User Modeling. User Modeling and User-Adapted Interaction 13, 1-2 (February 2003), 5-33. DOI=10.1023/A:1024022819690

[17] Peska, A., et al. (2011). UPComp - A PHP Component for Recommendation Based on User Behaviour. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT '11), Vol. 3. IEEE Computer Society, Washington, DC, USA, 306-309. DOI=10.1109/WI-IAT.2011.180

[18] Ricci, F., et al. (2011). Recommender Systems Handbook, Springer

Submitted: April 5, 2016.
Accepted: May 5, 2016.