Chapter 20: Applying Recommender Systems for Learning Analytics: A Tutorial

Soude Fazeli[1], Hendrik Drachsler[2], Peter Sloep[2]

[1] Welten Institute, Research Centre for Learning, Teaching and Technology, The Netherlands
[2] Open University of the Netherlands, The Netherlands

ABSTRACT

This chapter provides an example of how a recommender system experiment can be conducted in the domain of learning analytics (LA). The example study presented in this chapter followed a standard methodology for evaluating recommender systems in learning. The example is set in the context of the FP7 Open Discovery Space (ODS) project that aims to provide educational stakeholders in Europe with a social learning platform in a social network similar to Facebook, but unlike Facebook, exclusively for learning and knowledge sharing. In this chapter, we describe a full recommender system data study in a stepwise process. Furthermore, we outline shortcomings for data-driven studies in the domain of learning and emphasize the high need for an open learning analytics platform as suggested by the SoLAR society.

Keywords: Recommender system, offline data study, user study, collaborative filtering, information search and retrieval, information filtering, research, methodology, sparsity

With the emergence of massive amounts of data in various domains, recommender systems have become a practical approach to provide users with the most suitable information based on their past behaviour and current context. Duval (2011) introduced recommenders as a solution "to deal with the 'paradox of choice' and turn the abundance from a problem into an asset for learning" (p. 9), pointing out that several domains such as educational data mining, big data, and Web analytics all try to find patterns in large amounts of data. For instance, data mining approaches can make recommendations based on similarity patterns detected from the collected data of users. Furthermore, a survey conducted by Greller and Drachsler (2012), identified recommender systems and personalization as an important part of LA research.

Recommender systems can be differentiated according to their underlying technology and algorithms. Roughly, they are either content-based or use collaborative filtering. Content-based algorithms are one of the main methods used in recommender systems; they recommend an item to the user by comparing the representation of the item's content with the user's preference model (Pazzani & Billsus, 2007). Collaborative filtering is based on users' opinions and feedback on items. Collaborative filtering algorithms first find like-minded users and introduce them as so-called nearest neighbours to some target user; then they predict an item's rating for that user on the basis of the ratings given to this item by the target users' nearest neighbours (co-ratings) (Herlocker, Konstan, Terveen, & Riedl, 2004; Manouselis, Drachsler, Verbert, & Duval, 2012; Schafer, Frankowski, Herlocker, & Sen, 2007).

In the past, we have applied recommender systems in various educational projects with different objectives (Drachsler et al., 2010; Fazeli, Loni, Drachsler, & Sloep, 2014; Drachsler et al., 2009). In this chapter we want to share some best practices we have identified so far regarding the development and evaluation of recommender system algorithms in education; we especially want to provide an example of how to set up and run a recommender systems experiment.

As described by the RecSysTEL working group for Recommender Systems in Technology-Enhanced Learning (Drachsler, Verbert, Santos, & Manouselis, 2015) it is important to apply a standard evaluation method. The working group identified a research methodology consisting of four critical steps for eval-

uating a recommender system in education:

1. A *selection of dataset(s)* that suit the recommendation task. For instance, the recommendation task can be finding new items or finding relevant items for a user.

2. An *offline data study* of different algorithms on the selected datasets including well-known datasets (if possible, education-oriented datasets such as MovieLens makes movie recommendations) to provide insights into the performance of the recommender systems.

3. A comprehensive *user study* to test psycho-educational effects on learners as well as on the technical aspects of the designed recommender system.

4. A *deployment* of the recommender system in a real life application, where it can be tested under realistic, normal operational conditions with actual users.

The above four steps should be accompanied by a complete description of the recommender system according to the classification framework presented (Drachsler et al., 2015). The dataset used should be reported in the special section on educational datasets of the Journal of Learning Analytics[1] and made available for other researchers under certain conditions (Dietze, Siemens, Taibi, & Drachsler, 2016). This would allow other researchers to repeat and adjust any part of the research to gain comparable results and new insights and thus build up a body of knowledge around recommender systems in learning analytics.

In this chapter, we present an example of an experimental study that followed the research methodology described above for recommender systems in education. The rest of the chapter is organized as follows: In the next section, we present an example of a recommender system study that followed the methodology described above step by step. Next, we explain the practical implications of the experiment; then, we conclude.

A RECOMMENDER SYSTEM EXPERIMENT IN THE EDUCATIONAL DOMAIN

In this section, we describe how one should evaluate a recommender system in learning, making use of an experimental study presented in our 2014 EC-TEL paper (Fazeli et al., 2014). This study follows the standard methodology described above. To this methodology, however, we added an additional step: that of developing a conceptual/theoretical model (Fazeli et al., 2013)., which is presented in a RecSysTEL special issue

(Manouselis et al., 2012).

In our study, our target environment is social learning platforms in general. Social learning platforms work similarly to social networks such as Facebook but, unlike Facebook, they are developed exclusively for the purpose of learning and knowledge sharing. They often serve, therefore, as a common place exclusively for educational stakeholders such as teachers, students, learners, policy makers, and so on. Our target social learning platform is Open Discovery Space (ODS).[2] As indicated on the ODS homepage,

> ODS addresses various challenges that face the eLearning environment in the European context. The interface has been designed with students, teachers, parents and policy makers in mind. ODS will fulfill three principal objectives. Firstly, it will empower stakeholders through a single, integrated access point for eLearning resources from dispersed educational repositories. Secondly, it engages stakeholders in the production of meaningful educational activities by using a social-network style multilingual portal, offering eLearning resources as well as services for the production of educational activities. Thirdly, it will assess the impact of the new educational activities, which could serve as a prototype to be adopted by stakeholders in school education.

The main goal of our study is to find out which recommender system can best suit the data and information needs of a social learning platform, the main recommendation task being to finding relevant items for users. In the following sub-sections, we describe the study step by step.

Dataset Selection

Most data studies target a specific environment or specific group of users and thus require a specific type of data. In our case, the target social learning platforms is ODS. Consequently, we tried to find data collected from learning platforms similar to ODS. We chose the MACE and OpenScout datasets for the following reasons:

1. The datasets provide social data of users (ratings, tags, reviews, et cetera) on learning resources. So, the structure, content, and target users of the datasets are similar to those of ODS.

2. Running recommender algorithms on these datasets helps us to evaluate their performance before going online with the actual users of the ODS.

3. Both the MACE and OpenScout datasets comply with the CAM (Context Automated Metadata) format (Schmitz et al., 2009), which offers a standard metadata specification for collecting and storing

Table 20.1. Details of the Selected Datasets

Dataset	# of users	# of items	Transactions	Sparsity (%)	Source
MACE	631	12,571	23,032	99.70	MACE portal
OpenScout	331	1,568	2,560	99.50	OpenScout portal
MovieLens	941	1,512	96,719	93.69	GroupLens research

social data. CAM has also been applied in the ODS for storing social data.

Besides these two datasets, we also tested the MovieLens dataset as a reference since, up until now, the educational domain has been lacking reference datasets for study, unlike the ACM RecSys conference series, which deals with recommender systems in general. Table 20.1 provides an overview of all three datasets (Fazeli et al., 2014). Note that the educational datasets MACE and OpenScout clearly suffer from extreme sparsity. All the data are described more fully our EC-TEL 2014 article (Fazeli et al., 2014).

Offline Data Study

Algorithms. In this second step, we tried to select algorithms that would work well with our data. First, it is important to check the input data to be fed into the recommender algorithms. In this case, the ODS data, thus the data of the selected datasets, includes interaction data of users with learning resources (items). Therefore, we chose to use the Collaborative Filtering (CF) family of recommender systems. CF algorithms rely on the interaction data of users, such as ratings, bookmarks, views, likes, et cetera, rather than on the content data used by content-based recommenders. CF recommenders can be either memory-based or model-based, according to the "type"; they can be either item-based or user-based, referring to the "technique." For a detailed description of these distinctions, please see Section 4 of Fazeli et al. (2014). In our study, we made use of all types and techniques: both memory-based and model-based, as well as both user-based and item-based. Figure 20.1 shows our experimental method, consisting of three main steps:

1. We compared performance of memory-based CFs, including both user-based and item-based, by employing different similarity functions.

2. We ran the model-based CFs, including state-of-the-art Matrix Factorization methods, on our sample data.

3. We performed a final comparison of best-performing algorithms from steps 1 and 2. In addition to the baselines, we evaluated a graph-based approach proposed to enhance the mechanism of finding neighbours using the conventional k-nearest neighbours (kNN) method (Fazeli et al., 2014).

Performance Evaluation. After choosing suitable datasets and recommender algorithms, we arrive at the task of evaluating the performance of candidate algorithms. For this, we need to define an evaluation protocol (Herlocker et al., 2004). A good description of an evaluation protocol should address the following questions:

Q1. What is going to be measured?

Typically, in most offline recommender system studies, we measure the *prediction accuracy* of the recommendations generated. By this, we want to measure how much the rating predictions differ from the actual ones by comparing a training set and a test set. The training and test sets result from splitting our user ratings data (the same as user interaction data). In our
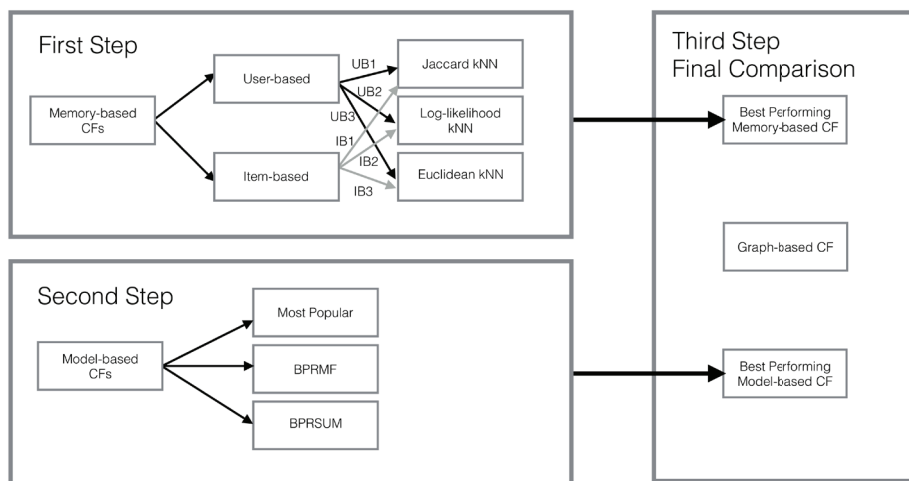

Figure 20.1. Experimental method used in Fazeli et al., 2014.
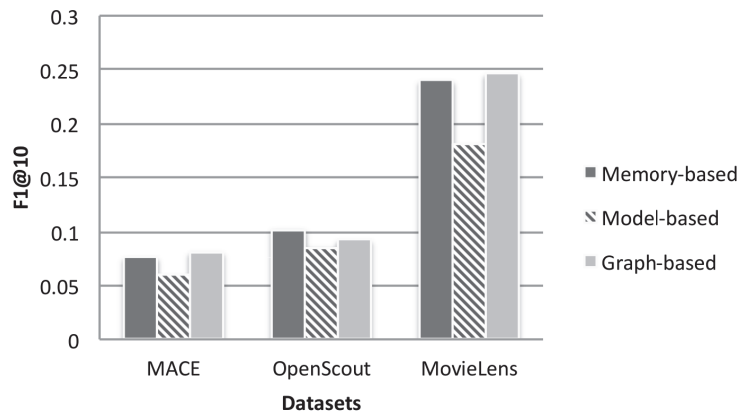
Figure 20.2. F1 of the graph-based CF and the best performing baseline memory-based and model-based CFs for all the datasets used (Fazeli et al., 2014).

EC-TEL 2014 study, we split user ratings into 80% and 20% for the training set and the test set, respectively. This kind of split is commonly used in recommender systems evaluations (Fazeli et al., 2014).

Q2. Which metrics are suitable for a recommender system study?

If our input data contains explicit user preferences, such as 5-star ratings, we can use MAE (mean average error) or RMSE (root mean square error). MAE and RMSE both follow the same range as the user ratings; for example, if the data contains 5-star ratings, these metrics range from 1 to 5.

If the input data contains implicit user preferences, such as views, bookmarks, downloads, et cetera, we can use Precision, Recall, and F1 scores. We made use of the F1 score since it combines precision and recall, which are both important metrics in evaluating the accuracy and coverage of the recommendations generated (Herlocker et al., 2004). F1 ranges from 0 to 1.

In addition, we need to define the n in top-n recommendations on which a metric is measured, also known as a *cut-off*. In Fazeli et al. (2014), we computed the F1 for the top 10 recommendations of the result set for each user.

Finally, we present the results of running the candidate algorithms on the datasets following the defined evaluation protocol. Due to limited space, we only present the final results of our EC-TEL 2014 article here. Please see Sections 5.1 and 5.2 of the original article (Fazeli et al., 2014) for more results.

Figure 20.2 shows the F1 results of best performing memory-based CF (Jaccard kNN), model-based CF (a Bayesian method), compared to the graph-based CF. The x-axis indicates the datasets used and the y-axis shows the values of F1. As Figure 20.2 shows, the graph-based approach performs best for MACE (8%)

and MovieLens (24%) and the selected memory-based and model-based CFs come in second and third place right after the graph-based CF. For OpenScout, the memory-based approach performs better with a difference of almost 1%.

In conclusion, according to the results presented in Figure 20.2, the graph-based approach seems to perform well for social learning platforms. This is reflected by an improved F1, which is an effective combination of precision and recall of the recommendation made.

Deployment of the Recommender System and User Study

In the educational domain, the importance of user studies has become ever more apparent (Drachsler et al., 2015). Since the main aim of recommender systems in education goes beyond accurate predictions, it should extend to other quality indicators such as usefulness, novelty, and diversity of the recommendations. However, the majority of recommender system studies still rely on offline data studies alone. This is probably because user studies are time consuming and complicated.

After running the offline data study on the ODS data, we furthered the work reported in Fazeli et al. (2014) by conducting a user study with our target platform. For this, we integrated the algorithms that performed best with ODS. We asked actual users of ODS whether they were satisfied with the recommendations we made for them. For this we used a short questionnaire using five metrics: usefulness, accuracy, novelty, diversity, and serendipity. The full description and results of this data study and the follow-up user study have not been published yet. The user study does not confirm the results of the data study we had run on the actual ODS data, showing that it is quite necessary to run user studies that can go beyond the success indictors of data studies, such as prediction accuracy.

Accuracy is one of the important metrics in evaluating recommender systems but relying solely on this metric can lead data scientists and educational technologists down less effective pathways.

PRACTICAL IMPLICATIONS AND LIMITATIONS

Accessing most educational datasets is challenging since they are not publicly and openly available, for example through reference links. Moreover, it is often difficult to compare the findings of related studies, for instance those by Verbert et al. (2011) and Manouselis, Vuorikari, and Van Assche (2010). Although we applied the same datasets, and some of the algorithms used in those two studies, the results of our example study differ from their results. Therefore, we could not gain additional information from the comparisons regarding the personalization of learning resources. One possible reason is that the studies use different versions of the same dataset because the collected data belongs to different periods of time. For the MACE dataset, for instance, different versions are available. In fact, no unique version has been fixed for running experiments nor for comparison in the recommender system community.

This problem originates from the fact that, unfortunately, there is no gold-standard dataset in the educational domain comparable to the MovieLens dataset[3] in the e-commerce world. In fact, the LA community is in need of several representative datasets that can be used as a main set of references for different personalization approaches. The main aim is to achieve a standard data format to run LA research. This idea was initially suggested by the dataTEL project (Drachsler et al., 2011) and later followed up by the SoLAR Foundation for Learning Analytics (Gašević et al., 2011). In the domain of MOOCs, Drachsler and Kalz (2016) have discussed this lack of comparable results and the pressing need for a research cycle that uses data repositories to compare scientific results. Moreover, an EU-funded project called LinkedUp[4] follows a promising approach towards providing a set of gold-standard datasets by applying linked data concepts (Bizer, Heath, & Berners-Lee, 2009). The LinkedUp project aims to provide a linked data pool for learning analytics research and to run several data competitions through the central data pool.

Overall, the outcomes of different recommender systems or personalization approaches in the education domain are still hardly comparable due to the diversity of algorithms, learner models, datasets, and evaluation criteria (Drachsler et al., 2015; Manouselis et al., 2012).

CONCLUSION

The main goal of this chapter has been to illustrate how to identify the most appropriate recommender system for a learning environment. To do so, we followed an example data study using the standard methodology presented in Drachsler et al. (2015) for evaluating recommender systems in learning. The methodology consists of four main steps:

1. Select suitable datasets preferably from the educational domain and, in case the actual data is not available yet, similarly to the target data.

2. Run a set of candidate recommender algorithms that best fits the input data. The output of this step should reveal which recommender algorithms best works with the input data.

3. Conduct a user study to measure user satisfaction on the recommendations made for them.

4. Deploy the best candidate recommender to the target learning platform.

The fact that our user study results did not confirm the results of the offline data study illustrates the importance of running user studies even though they are quite time consuming and complicated.

ACKNOWLEDGMENTS

REFERENCES

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.

Dietze, S., Siemens, G., Taibi, D., & Drachsler, H. (2016). Editorial: Datasets for learning analytics. *Journal of Learning Analytics*, 3(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

[3] http://www.grouplens.org/node/73
[4] www.linkedup-project.eu

Drachsler, H., Hummel, H., van den Berg, B., Eshuis, J., Waterink, W., Nadolski, R., Berlanga, A., Boers, N., & Koper, R. (2009). Effects of the ISIS recommender system for navigation support in self-organised learning networks. *Journal of Educational Technology & Society*, 12(3), 115–126.

Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning*, 32(3), 281–290. http://doi.org/ 10.1111/jcal.12135

Drachsler, H., Rutledge, L., Van Rosmalen, P., Hummel, H., Pecceu, D., Arts, T., Hutten, E., & Koper, R. (2010). ReMashed: An usability study of a recommender system for mash-ups for learning. *International Journal of Emerging Technologies in Learning*, 5. http://online-journals.org/i-jet/article/view/1191

Drachsler, H., Verbert, K., Santos, O., & Manouselis, N. (2015). Panorama of recommender systems to support learning. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook*, 2nd ed. Springer.

Drachsler, H., Verbert, K., Sicilia, M.-A., Wolpers, M., Manouselis, N., Vuorikari, R., Lindstaedt, S., & Fischer, F. (2011). *dataTEL: Datasets for Technology Enhanced Learning – White Paper*. Stellar Open Archive. http://dspace.ou.nl/handle/1820/3846

Duval, E. (2011). Attention please! Learning analytics for visualization and recommendation. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 9–17). New York: ACM.

Fazeli, S., Loni, B., Drachsler, H., & Sloep, P. (2014). Which recommender system can best fit social learning platforms? *Proceedings of the 9th European Conference on Technology Enhanced Learning* (EC-TEL '14), 16–19 September 2014, Graz, Austria (pp. 84–97).

Gašević, G., Dawson, C., Ferguson, S. B., Duval, E., Verbert, K., & Baker, R. S. J. d. (2011, July 28). *Open learning analytics: An integrated and modularized platform*. http://www.elearnspace.org/blog/wp-content/uploads/2016/02/ProposalLearningAnalyticsModel_SoLAR.pdf

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 42–57). New York: ACM.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), 5–53.

Manouselis, N., Drachsler, H., Verbert, K., & Duval, E. (2012). *Recommender systems for learning*. Springer Berlin Heidelberg.

Manouselis, N., Vuorikari, R., & Van Assche, F. (2010). Collaborative recommendation of e-learning resources: An experimental investigation. *Journal of Computer Assisted Learning*, 26(4), 227–242.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 325–341). Lecture Notes in Computer Science vol. 4321. Springer.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 291–324). Lecture Notes in Computer Science vol. 4321. Springer.

Schmitz, H., Scheffel, M., Friedrich, M., Jahn, M., Niemann, K., Wolpers, M., & Augustin, S. (2009). CAMera for PLE. In U. Cress, V. Dimitrova, & M. Specht (Eds.), *Learning in the Synergy of Multiple Disciplines: 4th European Conference on Technology Enhanced Learning* (EC-TEL 2009) *Nice, France, September 29–October 2, 2009 Proceedings* (pp. 507–520). Lecture Notes in Computer Science vol. 5794. Springer.

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11), 27 February–1 March 2011, Banff, AB, Canada (pp. 44–53). New York: ACM.