Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 1 (2010) 2859–2864

Procedia Computer Science

www.elsevier.com/locate/procedia

ELSEVIER

1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010)

Exploring user-based recommender results in large learning object repositories: the case of MERLOT

Miguel-Ángel Sicilia[a]*, Elena García-Barriocanal[a], Salvador Sánchez-Alonso[a], Cristian Cechinel[b]

[a] *Computer Science Department, University of Alcalá, Ctra. Barcelona km. 33.6, Alcalá de Henares 28871, Spain*
[b]*Computer Engineering Course, Federal University of Pampa, Caixa Postal 07, Bagé 96400-970, Brazil*

Abstract

Collaborative filtering (CF) techniques have proved to be effective in their application to e-commerce and other application domains. However, their applicability to the recommendation of learning resources deserve separate attention as seeking learning resources can be hypothesized to be substantially different from selecting information resources or products for purchase. To date there are only a few scattered studies reporting on the application of well known user-based CF algorithms to learning object repositories. This paper reports an empirical study carried out by using MERLOT data and existing user-based CF algorithms. The aim of this preliminary study was that of finding evidence on accuracy measures of existing CF algorithms, and the relation of the items recommended with other elements of the repository. The results can be used as a starting point for future studies that account for the specific context of learning object repositories and the different aspects of preference in learning resource selection.

Keywords: Collaborative filtering; learning objects; learning object repositories; recommender systems; MERLOT

1. Introduction

Collaborative filters predict someone's personal preferences for information and/or products by keeping track of their likes and dislikes, and then connecting that information with a database of other peoples' preferences. Similarity of user profiles is thus the key element of collaborative filtering. Collaborative filtering has been successfully applied to build e-commerce recommenders [1] and it has the potential to be useful for any kind of item collection in which a community of users is enabled to express their preferences about the items (be them digital files, products or other kind of items). Learning object repositories (LOR) represent a special kind of digital collection, in which the preferences about items can be considered to be related to contingent learning needs. In consequence, seeking learning resources can be hypothesized to be substantially different from selecting products

* Corresponding author. Tel.: +34-91-885-4000; fax:+34-91-885-6646.
E-mail address: msicilia@uah.es.

doi:10.1016/j.procs.2010.08.011

for purchase or information resources. Particularly, it seems apparent that the criteria for quality of learning objects are complex. For example, the *Learning Object Review Instrument* (LORI) evaluates the quality using nine criteria: Content Quality, Learning Goal Alignment, Feedback and Adaptation, Motivation, Presentation Design, Interaction Usability, Accessibility, Reusability, and Standards Compliance [2]. This appears to call for multi-attribute CF algorithms that consider different user profiles, as the different criteria are reflecting different concerns (e.g. standards compliance might be an issue of interest only to some users). Also, in LOR both instructors and learners are users, so that it might be possible that mixing preferences of both kinds would produce poor results when used with existing CF algorithms that do not consider different types of users. Finding response to those questions requires first an assessment of existing CF algorithms applied to LOR in order to contrast if they are applicable "as is" without considering any specific aspects. It is known that recommender system evaluation can be approached from different perspectives [3], including accuracy, coverage, usefulness or novelty among others. Among them, accuracy evaluation can be performed offline by using datasets of ratings, which has the benefit of providing a first assessment before evaluations with users is approached. This paper describes a prediction accuracy evaluation of well-known recommender algorithms using a dataset gathered from the popular MERLOT repository. This initial study is intended as a first step prior to a more comprehensive evaluation combining several criteria.

The rest of this paper is structured as follows. Section 2 describes related research and the motivation for the present study. Section 3 reports an internal evaluation of two CF algorithms on a rating dataset extracted from MERLOT. Then, Section 4 reports an analysis of recommendations generated. Finally, conclusions and outlook are provided in Section 5.

2. Background

Learning object repositories provide a platform for the sharing of educational resources on the Web, and most of them provide some mechanisms for building community dynamics around their resources. The community dimension and its social dynamics have been found to be an important aspect for the success of these repositories. For example, Brosnan [4] provided a conceptualization for that importance based on social capital theory, and Monge, Ovelar and Azpeitia [5] analyzed the potential impact of Web 2.0 strategies to foster social dynamics and participation in repositories. In a similar direction, Han et al. [6] reported an empirical study on the LON-CAPA repository in which a non-explicit community model was identified on the basis of co-contribution of resources to the same courses by popular authors. Several repositories have started to provide services by which members can share their personal collections of favorite resources and comment or review other's resources. This information is in those sites openly available in some cases, and represents an objective account of usage and expression of preference.

Recker, Walker & Wiley [7] reported on three experiments of a collaborative filtering system, with 15, 63 and 375 participants respectively, but provided only user experience measures, not an actual evaluation of the accuracy of the collaborative filtering algorithm. Manouselis, Vuorikari and Van Assche [8] reported on a dataset of 2000+ multi-attribute learning object evaluations from the CELEBRATE portal. The study contrasted variations of three options of similarity calculation (Pearson, similarity and vector/cosine) and two neighborhood selection criteria (correlation weight threshold and maximum number of neighbors). Their results found several algorithms performing reasonably well and with acceptable performance figures. This paper aims at complementing these existing studies with the evaluation of several recommender variants evaluated against a data set of rating data extracted from the MERLOT learning object repository[b]. Having MERLOT data allows also to contrast the recommended items with other quality or endorsement mechanisms in the repository, aiming to explore potential relations between them.

3. Evaluating CF Algorithms with MERLOT Data

A database from the MERLOT repository was gathered May 2009 by using a crawler that systematically traversed the Web pages of the repository. Information of a total of 69,248 users was extracted, of which 1,393 were

[b] http://www.merlot.org/

also recognized as resource authors. 434 of these authors had no declared organization, and the rest of the frequencies of occurrence of individuals from the same organization were below 10, which allows us to discard a possible bias coming from a dominant institution behind the community of users. The fact that there is a significantly higher number of contributors than authors is relevant as it points out that MERLOT is more a community of contributors than of authors. This can be a result of MERLOT being a repository only storing metadata and not the contents themselves, as occurs in other systems as Connexions[c].

Ratings from comments in MERLOT were extracted, totaling 6103 ratings. The distribution of ratings among users is depicted in Figure 1 (three users with 495, 334 and 194 contributed ratings respectively have been omitted from the Figure), and a similar distribution is found for ratings among learning objects. Both distributions appear to follow a typical power law distribution, which is consistent with existing studies on the unequal distribution of contributions in learning object repositories [9].

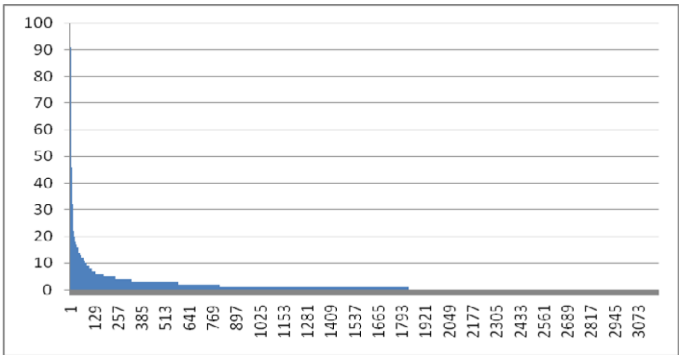

Fig.1. Distribution of ratings among users

The evaluations were done using Apache Mahout version 0.3[d]. Concretely, the algorithm used was a *Generic User Based Recommender* using two different similarity measures: *Pearson Correlation Similarity* and *Euclidean Distance Similarity*. The algorithm variants tested were dependant on the following parameters:

- Neighborhood size, i.e. the maximum amount of similar users taken for the computation of predicted ratings. Variations from 1 to 20 users were experimented.
- Minimum similarity, i.e. the minimum similarity required to include a neighbor in the computation. Variations from zero to one in increments of 0.1 were considered.

The evaluation was based on the commonly used measure of absolute average error, using Mahout's `AverageAbsoluteDifferenceRecommenderEvaluator`, applied to the different variants of both algorithms for different values of the two parameters. The process of computation used a split of the sample data taking 90% of the data for computing the predicted ratings, and 30% for evaluation[e]. The computation of the average error was done repeatedly (100 times) per each algorithm configuration and the average of the error of the 100 runs was used for the analysis. This was done to overcome the effect of some particular splits of the evaluation and testing data.

The first evaluation was the influence of parameters on the quality of prediction. Figure 2 depicts average error for the Pearson configuration and for different neighborhood sizes. A similar distribution is found when using the Euclidean metric. In both cases, overall average absolute error is around one.

Average error was found to be non correlated with neighbourhood sizes for both Pearson and Euclidean, and having a very small positive and non statistically significant correlation with minimum similarity for Pearson. Figure 3 shows for the Euclidean case the relationship between average error and minimum similarity, with a large variation for minimum similarity one (which is the unlikely case of having users with identical preferences). In the case of Euclidean, a statistically significant negative correlation of 0.63 between minimum similarity and error was found, which is evident in Figure 3. This suggests that the best minimum similarity for Euclidean is relatively high, around 0.8.

[c] http://cnx.org/

[d] http://mahout.apache.org/

[e] This is not an introducing overfitting or bias as described in the documentation of Mahout.

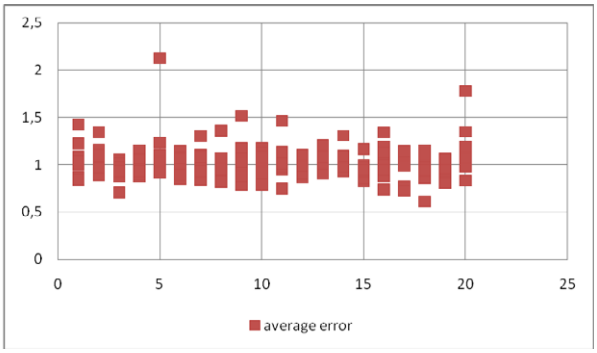

Fig. 2. Average error for the Pearson configuration for different neighborhood sizes

The average error results can be considered high especially in the context of the kind of digital repositories analyzed [8]. It is known that the distribution of ratings in MERLOT is not uniform, as high ratings are much more frequent than lower ones [10]. For practical purposes, this entails that the recommended items (which are usually highly rated ones) are very sensitive to prediction errors, as one point of average error in a one to five scale represents a significant variation.

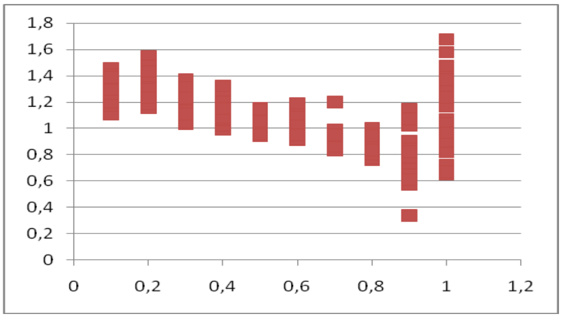

Fig. 3. Average error for different levels of minimum similarity (Euclidean distance)

4. Analysis of Recommended Items

The Pearson algorithm variant with best performance according to the above described analysis was used to generate systematically recommendations for all the users. Concretely, they were asked to generate at most 10 recommendations for each user. The Pearson algorithm generated 548 recommendations to 158 learning objects and 109 users. The distribution of number of recommendations per user is depicted in Figure 4.

As can be appreciated in Figure 4 the distribution shows most users distributed in two segments: those with 2 or less recommendations and those with 10 recommendations. However, as 10 was the maximum number generated, this distribution would need further exploration. In any case there is no apparent known distribution that adjusts to the sample.

One important aspect of the use of CF algorithms is to what extent recommendation algorithms complement existing endorsement or prestige mechanisms in repositories. In the case of MERLOT, peer reviews and ratings are used – but not as personalized rankings, together with some labeling as “Editor’s choice” or the like. An interesting contrast is analyzing if the algorithms generate recommendations to items that are already favored by these existing endorsement mechanisms. In that direction, it is interesting to note that only a 1.9% of the recommended learning objects had the “Editor’s Choice” label in MERLOT and only 9.5% of them had the “Merlot Classics” label. However, in MERLOT only 0.07% of resources are labeled as “Editor’s Choice”, and only 0.42% are labeled as “Merlot Classics”. This suggests a relation between these endorsement mechanisms and the recommendations outcomes that might be attributed to some intrinsic quality characteristics.

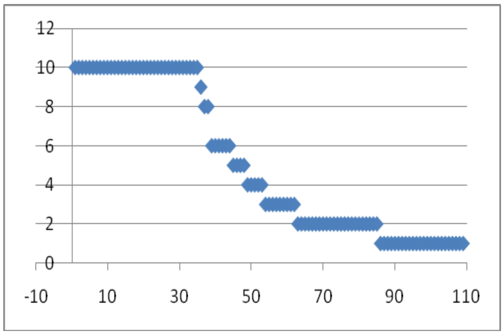

Fig. 4. Distribution of number of recommendations per user

In other direction, it is interesting to note that all of the recommended items generated have been peer reviewed (obviously all of them have also at least one comment, but this was expected as the ratings are associated to the comments). It should be noted that only a 12.65% of learning objects in MERLOT were peer reviewed at the time of taking the data used in this study, and only 3.38% are at the same time peer reviewed and commented.

The median of the ratings of recommended items (sample=1) is 4.5 while the median of the ratings in the whole MERLOT data set is 4 (sample=2). There is a statistical significant difference among the medians. However, this could be explained by the fact that recommender algorithms tend to recommend highly rated items. The density plots in Figure 5 show the difference in the distribution of ratings.

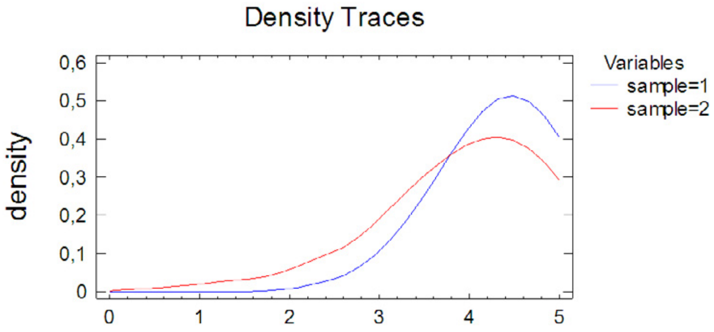

Fig.5. Distribution of ratings

Another interesting contrast is that of personal collections (bookmark collections of each user) and the recommended items. In this case, the differences are high and statistically significant, as recommended items have an average of occurrences in personal collections of 15, while the overall average is 1.5. As personal collections are known to be a good predictor of high ratings [10], this is reflecting that items included in many personal collections are more likely to have high ratings and to be recommended. This opens the question of to what extent personal collections can be used as an alternative source of preference expressions with similar or even better properties than ratings associated to comments.

5. Conclusions and Outlook

Standard CF algorithms have been evaluated with learning object ratings extracted from MERLOT in May 2009. Concretely, user based recommendations using Pearson and Euclidean similarity metrics have been explored, generating algorithm variants with neighborhood size and minimum similarity as parameters. Results show relatively high mean absolute errors around 1.0 in a scale of 1 to 5 (especially if we contrast with the results provided in [8]). Given the distributions of ratings in this kind of repositories (which is unequally distributed to high values, i.e. people tend to provide high ratings), these figures appear too high to make a proper discrimination of preferences with predicted ratings.

It has been found also that recommendations generated are somewhat related to other endorsement mechanisms in MERLOT as “Editor’s choices” or “MERLOT Classics”, in consequence, CF algorithms appear to select resources that are given some quality assessment otherwise. Further, all the recommended items have been peer reviewed and tend to be objects included in many personal collections, which deserves further exploration. The distribution of ratings for recommended items is similar to that of the overall dataset with some shift to higher values in the former, which for the moment discards some bias in the amount of ratings of recommended ones.

This exploratory study has provided evidence about a relatively poor performance of some existing CF algorithms, but their potential to select high quality resources, complementing other endorsement mechanisms. Recommendations tend to be to objects that were peer reviewed and that are included in many personal collections, which might be attributed to the fact that all these mechanisms are reflecting quality, but this opens the question of what kind of relationship exists among them.

Further contrasts with other elements of MERLOT are required in future work. For example, it would be interesting to contrast if recommendations for a give user fall in the disciplinary area of that user or are crossing disciplines. Also, it would be interesting to contrast if using personal collections as zero/one ratings would produce similar results to those reported by the CF algorithms studied. Also, user studies and other evaluation aspects need to be addressed in future work.

Acknowledgements

The results presented in this project have been partially funded by the University of Alcalá and the CAM (Comunidad de Madrid), as part of project MARIA, code CCG08-UAH/TIC-4178, and by the Spanish Ministry of Education through project MAPSEL, code TIN2009-14164-C04-01.

References

1. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce (Minneapolis, Minnesota, United States, October 17 - 20, 2000). EC '00. ACM, New York, NY, 158-167
2. Vargo, J., Nesbit, J. C., Belfer, K., & Archambault, A. (2003). Learning object evaluation: Computer mediated collaboration and inter-rater reliability. International Journal of Computers and Applications, 25(3), 198-205.
3. Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 1 (Jan. 2004), 5-53.
4. Brosnan, K. (2005). “Developing and sustaining a national learning-object sharing network: A social capital theory perspective,” In: J.B. Williams, & M.A. Goldberg (Eds.), Proceedings of The ASCILITE 2005 Conference, pp. 105-114, Brisbane: Australia.
5. Monge, S., Ovelar, R. and Azpeitia, I. (2008). “Repository 2.0: Social Dynamics to Support Community Building in Learning Object Repositories”. Interdisciplinary Journal of E-Learning and Learning Objects, 4, 2008, http://ijklo.org/Volume4/IJELLOv4p191-204Monge.pdf
6. Han, P., Kortemeyer, G., Kramer, B.J. and Prummer, C. von. (2008). “Exposure and Support of Latent Social Networks Among Learning Object Repository Users. rs”. Journal of Universal Computer Science (J.UCS), 14, 2008.
7. Recker, M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative filtering of Web resources for education. Instructional Science, 31(4/5), 229-316.
8. Manouselis, N., Vuorikari, R., and Van Assche, F. (2007). "Simulated Analysis of MAUT Collaborative Filtering for Learning Object Recommendation", in Proc. of the Workshop on Social Information Retrieval for Technology-Enhanced Learning (SIRTEL 2007), 2nd European Conference on Technology Enhanced Learning (EC-TEL'07), Crete, Greece, September 2007.
9. Ochoa, X. and Duval, E. (2008). Quantitative Analysis of Learning Object Repositories. In: Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 6031-6048, Chesapeake, VA: AACE.
10. García-Barriocanal, E. and Sicilia, M.A. (2009). Preliminary Explorations on the Statistical Profiles of Highly-Rated Learning Objects. In: Proc. of the third Metadata and Semantics research conference (MTSR 2009), Springer Communications in Computer and Information Science 46, pp. 108-117.