Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2013) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

2013 International Conference on Computational Science

Layered Evaluation of Multi-Criteria Collaborative Filtering for Scientific Paper Recommendation

Nikos Manouselis[a,*], Katrien Verbert[b]

[a]Agro-know Technologies, Athens, Greece
[b]Department of Computer Science, Eindhoven University of Technology, Eindhoven, the Netherlands

Abstract

Recommendation algorithms have been researched extensively to help people deal with abundance of information. In recent years, the incorporation of multiple relevance criteria has attracted increased interest. Such multi-criteria recommendation approaches are researched as a paradigm for building intelligent systems that can be tailored to multiple interest indicators of end-users – such as combinations of implicit and explicit interest indicators in the form of ratings or ratings on multiple relevance dimensions. Nevertheless, evaluation of these recommendation techniques in the context of real-life applications still remains rather limited. Previous studies dealing with the evaluation of recommender systems have outlined that the performance of such algorithms is often dependent on the dataset – and indicate the importance of carrying out careful testing and parameterization. Especially when looking at large scale datasets, it becomes very difficult to deploy evaluation methods that may help in assessing the effect that different system components have to the overall design. In this paper, we study how layered evaluation can be applied for the case of a multi-criteria recommendation service that we plan to deploy for paper recommendation using the Mendeley dataset. The paper introduces layered evaluation and suggests two experiments that may help assess the components of the envisaged system separately.

Selection and/or peer-review under responsibility of the organizers of the 2013 International Conference on Computational Science

Keywords: Recommender systems; Multi-Criteria Decision Making (MCDM); Evaluation

1. Introduction

In recent years, there is an increased interest in recommendations for scientific papers. This interest is reflected in the proliferation of the development and use of recommendation services in several scientific

* Corresponding author. Tel.: +30-210-6897905; fax: +30-210-6891961.
E-mail address: nikosm@agroknow.gr

portals such as Mendeley.com, CiteSeer (http://citeseerx.ist.psu.edu), citeulike.org and Google Scholar (http://scholar.google.be). Several approaches have been presented by researchers to generate relevant recommendations of scientific papers to end-users, in order to facilitate retrieval of relevant scientific articles from large collections of papers offered by these portals. These approaches rely on, or extend, traditional recommendation techniques – such as collaborative filtering, content-based filtering or hybrid techniques [3]. Bogers and van den Bosch [2] apply for instance three different collaborative filtering techniques to a dataset that is crawled from citeulike.org and that contains traditional user-item preferences. Several other systems rely on multiple data sources or criteria, or different combinations of recommendation techniques.

McNee et al. [13] presents a hybrid approach that combines content-based and collaborative filtering techniques to generate recommendations on the basis of citations in scientific papers. Similarly, TechLens+ [17] combines the citations of an article with its content in order to generate recommendations. The system relies on a combination of collaborative and content-based recommendation techniques. Papyres [14] is a multi-criteria hybrid recommender system. The first level is a content-based filter that operates on LOM metadata. The second level filters the output of this level with a multi-criteria collaborative filtering technique. This filter considers multiple ratings on various criteria as they are provided by end-users, including the *research contribution*, *originality*, quality of the *literature review*, *readability* and *organization*, *technical quality*, *testing procedure* and *quality of references.* The interest in considering multiple criteria for recommending scientific papers is discussed by several researchers such as Matsatsinis et al. [12] and Kapoor et al. [5]. To our knowledge, the approach has not yet been evaluated to assess the quality of recommendations.

Evaluation of such systems can be a challenging and difficult task. Recent studies have suggested the adoption of layered approaches in order to identify the components of a system that may affect its overall performance [16]. Layered evaluation (or decomposition) frameworks have attracted research attention for more than a decade, with several frameworks, methods and instruments being proposed and tested in relevant literature [15]. They try to decompose a recommender system in its constituent subsystems or layers and then apply particular evaluation methods that can assess the performance of each targeted layer. Nevertheless, we are not aware of any recommender system evaluation study that has explicitly applied some particular layered evaluation framework.

In this paper, we particularly focus on applying a layered evaluation framework to fit the interaction components of a recommender system. Earlier studies have indicated that the best performing technique is often dependent on the application context, and the dataset that is used to evaluate the approach. In our study, we focus on a system that is based on a large dataset in order to see how it may have an influence on the performance of the recommendation approach. More specifically, the Mendeley dataset [4] is used as the basis upon which a multi-criteria recommendation system may be designed. Then the multi-criteria recommendation system to work on top of the Mendeley data is examined under the prism of a layered evaluation framework.

2. Multi-criteria collaborative filtering using the Mendeley dataset

In related research, the problem of recommendation has been identified as the way to help individuals in a community to find the information or products that are most likely to be interesting to them or to be relevant to their needs [7]. It has been further refined to the problem (i) of predicting whether a particular user will like a particular item (prediction problem), or (ii) of identifying a set of N items that will be of interest to a certain user (top-N recommendation problem) [1]. Therefore, the general recommendation problem can be formulated as follows: let C be the set of all users and S the set of all possible items that can be recommended. We define

as $U^c(s)$ a utility function $U^c(s): C \times S \rightarrow \Re^+$ that measures the appropriateness of recommending an item *s* to user *c*. It is assumed that this function is not known for the whole *C* x *S* space but only on some subset of it. Therefore, in the context of recommendation, we want for each user *c* to be able to:

(i) estimate (or approach) the utility function $U^c(s)$ for an item *s* of the space *S* for which $U^c(s)$ is not yet known; or,
(ii) choose a set of *N* items $s \in S$ that will maximize $U^c(s)$:

$$\forall c \in C, s = \arg\max_{s \in S} U^c(s) \quad (1)$$

In most recommender systems, the utility function $U^c(s)$ usually considers one attribute of an item, e.g. its overall evaluation or *rating*. Nevertheless, utility may also involve more than one attribute of an item. The recommendation problem therefore becomes a multi-attribute one. We want to explore how recommendation of scientific papers can be better defined if such a multi-attribute collaborative filtering approach is adopted.

To achieve this, we have taken into consideration the dataset that the Mendeley.com platform has published as part of the RecSys 2012 challenge by Mendeley [4]. We have treated this dataset as a multi-criteria rating one, considering that the *existence of a paper in a user library*, the *flag of whether a paper was read*, and the *existence of a vote/star next to a paper* can be three different dimensions upon which a user may express his/her preferences over a publication, with a value range for each dimension 0 to 1. The Mendeley dataset includes 50.000 user libraries that contain a total of 4.848.724 articles. In this dataset 615.308 of the 4.848.724 library entries has been starred by users. The general characteristics for the dataset are presented in Table 1.

Table 1: characteristics Mendeley dataset

Characteristic	Value
Number of users	50.000
Number of unique Items	3.652.285
Number of ratings	4.848.725
Rating ranges	0 to 1
Overall rating density (%)	0,003*

* estimated using the starred dimension.

The multi-criteria recommender system that we would like to test for the Mendeley case, is a multi-attribute utility (MAUT) collaborative filtering system that was introduced by Manouselis & Costopoulou [8], and has been experimentally tested in various occasions and contexts during the past few years (e.g. [9][10][11]). The studied approach is a multi-attribute extension of related single-attribute algorithms. It considers each attribute in separate, first trying to predict how the active user would evaluate item *s* upon each attribute, and then synthesizing these attribute-based predictions into a total utility value. A variety of design options can be considered for the studied algorithm, leading to several versions and parameterizations. A detailed explanation of these options can be found in Manouselis & Costopoulou [8].

3. Layered evaluation for MAUT recommender systems

Layered evaluation (or decomposition) frameworks have attracted research attention in adaptive systems'

research for more than a decade, with several frameworks, methods and instruments being proposed and tested in relevant literature [15]. They try to decompose a system in its constituent subsystems or layers and then apply particular evaluation methods that can assess the performance of each targeted layer. The rationale behind layered approaches is straightforward: most early attempts to evaluate such systems followed a "with- and without-personalization" approach; that is, the "personalization component" was "separated" from the system, and the two versions of the system (the one with personalization features and the one without) were compared to investigate whether personalization brought significant benefits. This approach has a fundamental problem: the "non-personalized" system used for evaluation is not an application which has been developed according to certain design considerations, but rather a "bi-product" resulting when removing the personalization component. Moreover, this approach is not useful when the personalized system is found to be ineffective, since there is no way to understand why the system (or which specific component of the system) was not successful so as to improve it.

Several layered evaluation frameworks have been proposed in the literature. The idea can be traced back to the early 90s, when Totterdell & Boyle [18] proposed that the accuracy of the user model and the effectiveness of the changes made by the system should be evaluated separately. Ten years later, Karagiannidis & Sampson [6] proposed the term "layered evaluation", and suggested that layered evaluation should address the main components of each system separately. Weibelzahl [19] has proposed a similar layered framework, suggesting the decomposition of personalization into the following three layers: (i) evaluation of input data, (ii) evaluation of the inference mechanism, and (iii) evaluation of the personalization decisions. Paramythis et al. [15] further elaborated their decomposition by proposing five layers (or modules): (i) interaction monitoring, (ii) interpretation and interface, (iii) modeling, (iv) personalization decision making, and (v) applying personalizations.

An extensive survey of evaluation state-of-art and issues in recommender systems was carried out by Pu et al. [16]. This survey identified a generic interaction model for such systems that includes three crucial components that corresponded to groups of interaction activities between the user and the system: the initial preference elicitation process, the preference refinement process, and the presentation of the system's recommendation results. This decomposition is very close to the way that layered evaluation frameworks are decomposing a system in separate components that can be evaluated one by one. Pu et al. have suggested that layered evaluation can be used in recommender systems' research as a powerful technique in identifying areas of a system that require further improvements. More specifically, the three interaction steps that the authors have identified are described as such:

- Elicit user preferences: the initial user preference profile can be established by users' stated preferences (explicit elicitation) or their objective behaviors (implicit elicitation).
- Display recommendations: the system uses the above information to decide what to suggest to a user, and is concerned with methods and strategies for effectively selecting and presenting results to its users.
- Revise user preferences: users' interaction with the system can lead to changes into the information stored as preferences, thus resulting into a revision of the user preference profile.

To investigate how this layered framework could be applied to the MAUT recommender system evaluation, we focus on these interaction steps and how they relate to the adaptation components/layers of a typical evaluation framework (e.g. [6]):

- All interactions related to the user preference profiles, i.e. the step of eliciting user preferences and the step of revising user preferences, are corresponding to the "assessment of interaction" component, since they deal with the way that the user model is being constructed and updated, and their evaluation should take place in similar ways.

- All interactions related to the recommendation provision itself, i.e. the step of displaying recommendations, is corresponding to the "adaptation decision making" component, since it deals with the way that the recommendation is being created and presented, and its evaluation should take place at this level.

According to this classification of interaction activities to such decomposition layers, it could be argued that a layered approach would also apply for the evaluation of a recommender system as it follows:

- *Layer 1 - evaluation of user modeling*: at this layer the user modeling process is being evaluated, focusing mostly on whether the user characteristics are being successfully represented, recorded and stored in the user model. This can include evaluation of the accuracy and completeness of the user model (e.g. self-assessment by users) but also of the granularity of the user model. It can also include experimentation with different modeling approaches, different model representation formats, as well as the evaluation of techniques to boost performance such the use of stereotypes to create an initial user model and avoid the cold-start problem.
- *Layer 2 – evaluation of adaptation decision making*: at this layer the adaptation process, logic and results are being evaluated, focusing mostly on whether the personalization actions are valid and meaningful for the given state of the user model. This phase can be evaluated through user testing (e.g. via usage scenarios) or by studying how the provided information leads to some desired result (e.g. buying a particular product or viewing a particular item). It can also separate the evaluation of how the recommendation is generated (testing different techniques or algorithms) from the evaluation of the way recommendation is presented (testing alternative interface design options).

4. Setting up evaluation experiments adopting a layered approach

According to the analysis carried out in the previous section, the MAUT system designed on top of the Mendeley dataset can be decomposed into:

i. All interactions related to the user preference profiles: in this system the user preferences are represented as ratings over items and the representation method is a user-item matrix. The rating types are numeric (measurable) and they are multi-criteria or multi-attribute ones, that is, ratings upon multiple dimensions are being provided by the user in order to express preferences over an item.
ii. All interactions related to the recommendations: in the system recommendations are provided in the form of predicted ratings for unknown values, that is, a collaborative filtering algorithm is used to predict how a user would rate an unknown item upon each dimension, according to how other people with similar user models have rated it. This is a memory-based approach since it uses all history of stored ratings for all users. It is also a personalized approach since the prediction is different for each user, depending on his/her past ratings as well as the ratings of people that are found as similar-minded.

In the next paragraphs we propose two possible evaluation experiments that could examine separately each component and reveal useful insight on whether the MAUT approach may bring added value into such an application context.

4.1. Evaluating the multi-criteria user model

The user model that we inferred from the Mendeley dataset is a typical one, since all collaborative filtering systems are using (explicit or implicit) ratings to represent user preferences over items. In the MAUT system, the particularity is that a multi-dimensional approach is used, which is argued to bring more accurate preference

modeling and therefore better recommendation results.

To evaluate how this user modeling approach performs in the context of the MAUT system in comparison to a single-attribute approach where only one overall rating is being provided for the item by the user, we can compare how the number of dimensions (criteria) affected the performance of the system. In particular:

- A large number of datasets may be created, by breaking down the very large Mendeley dataset into many smaller ones or by producing a variety of synthetic (simulated) datasets, which will have various properties, e.g. ranging from single-criterion to multi-criteria datasets and from very sparse to very dense ones.
- The algorithm of the MAUT collaborative filtering system can be then executed upon all these datasets, and its performance can be measured using a variety of metrics (such as MAE and coverage).
- The correlation can be then calculated between the properties of the dataset that relate to the user model (and especially the number and scale of the rating dimensions) and the values of the measured performance metrics. The aim is to explore whether using multiple dimensions seems to be connected with better or worse performance results for the collaborative filtering algorithms.

Table 2 presents an example/dummy version of how such results could look like, in order to be able to understand whether adding multiple dimensions to the user model may result into different performance measures.

Table 2: Example of the study of correlation between examples of dataset properties and examples of performance metrics

Dataset Property	**Metric**	**Correlation**
# of criteria	MAE	*value*
# of scales	MAE	…
# of criteria	Coverage	…
# of scales	Coverage	…
…	…	…

Table 3: Example of the comparison of various algorithms' performance over datasets from the same application domain

Algorithm	**Dataset I**	**Dataset II**	**Dataset …**
Basic algorithm A	*performance metric value*	…	…
Basic algorithm B	…	…	…
Basic algorithm …	…	…	…
Proposed MAUT algorithm	…	…	…

4.2. Evaluation of the collaborative filtering algorithm

The recommendation algorithms that have been used within the proposed MAUT Mendeley recommender are rather typical, and the only difference is that the user model is a multi-attribute one. Trying to disconnect the

evaluation of the user model from the evaluation of the recommendation method and algorithm, we would like to carry out an evaluation experiment that will compare the MAUT collaborative filtering algorithms versus some non-personalized basic algorithms. In this comparison, we would study how the collaborative filtering algorithm performed over a number of smaller datasets that have been created by the original larger one, or (even better) by studying their performance over additional datasets that come from paper recommendation settings except from the Mendeley one. Table 3 presents an example/dummy version of how such results could look like, comparing the performance of various algorithms and their variations over different datasets from this application domain.

5. Discussion

The benefit of using such a layered evaluation approach is first identified in the way it may reveal whether the choice of a data property value (such as the number of criteria or their rating scales) may affect the performance of the algorithms and therefore be treated carefully during experimentation. It may also help justify whether multi-criteria user modelling may improve the performance of recommender systems and in which ways, by studying the way that performance measures are correlated with the multiple rating attributes.

Furthermore, an additional benefit of such an experimental testing approach is the fact that it illustrates the importance of using multiple datasets (by creating small random samples of a bigger one, by generated simulated/synthetic ones, or by using different datasets available in the same application domain). This is the most widely used experiment type in recommender system research, but sometimes the importance of using multiple but comparable datasets is underestimated. The layered approach has indicated how important this perspective is when trying to assess the difference that a particular algorithm (or variation of an algorithm) makes for a given context.

Nevertheless, layered frameworks would benefit recommendation systems' research if they can further provide them with:

- Coherent and systematic evaluation methods ready for application in testing.
- Recommender system decompositions that can apply to more systems and algorithms.
- Ways in which evaluation hints may be translated into concrete and measurable indicators for the system implementor/operator.
- More detailed evaluation guidelines and recommendations, such as suggested methods, tools and instruments that would fit each component.

6. Conclusions

This paper introduces the case of a multi-criteria recommender system for the Mendeley platform using an existing dataset that Mendeley has published, which may be considered to be a multi-attribute one. It then studies how evaluation of this approach can take place by adopting a layered evaluation framework. The paper assesses the applicability of layered evaluation for such systems and examines the types of experiments that may be carried out to assess the various components of multi-criteria recommender systems. Our analysis indicates that implementing a layered-based recommender system evaluation has the potential to facilitate a more detailed and informed evaluation of such systems, allowing researchers and developers to better understand how they may improve them.

Acknowledgements

The work of Nikos Manouselis has been funded with the support by European Commission, and more specifically the FP7 project SemaGrow "Data Intensive Techniques to Boost the Real-Time Performance of Global Agricultural Data Infrastructures" (http://semagrow.eu). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Katrien Verbert is a Post-doctoral Fellow of the Research Foundation – Flanders (FWO).

References

[1] Adomavicius, G., Manouselis, N., & Kwon, Y., 2011. Multi-Criteria Recommender Systems. (F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor, Eds.) *Recommender Systems Handbook*, *1*(Mcdm), Springer; p. 769–803.

[2] Bogers, T., Van Den Bosch, A., 2008. Recommending scientific articles using citeulike. *Proceedings of the 2008 ACM conference on Recommender systems RecSys 08*. ACM Press; p. 287-290.

[3] Burke, R., 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and UserAdapted Interaction*, *12*(4):331–370. Springer

[4] Jack, K., Hammerton, J., Harvey, D., Hoyt, J. J., Reichelt, J., Henning, V., 2010. Mendeley's Reply to the DataTEL Challenge. *Procedia Computer Science*, *1*(2):1-3. Retrieved from http://www.mendeley.com/research/sei-whale/

[5] Kapoor, N., Chen, J., Butler, J. T., Fouty, G. C., Stemper, J. A., Riedl, J., Konstan, J. A., 2007. Techlens: a researcher's desktop. *Proceedings of the 2007 ACM conference on Recommender systems*. ACM; p. 287-290. Retrieved from http://portal.acm.org/citation.cfm?id=1297268

[6] Karagiannidis C., Sampson D., 2000. Layered Evaluation of Adaptive Applications and Services. In Brusilovsky P., Stock O., Strapparava C. (Eds.): Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000 Conference Proceedings). Springer LNCS 1892; p. 343-346.

[7] Konstan, J. A., 2004. Introduction To Recommender Systems: Algorithms and Evaluation. *ACM Trans. Inf. Syst.*, 22(1):1-4.

[8] Manouselis, N., Costopoulou, C., 2007. Analysis and Classification of Multi-Criteria Recommender Systems. *World Wide Web Internet And Web Information Systems*, *10*(4):415-441. Springer Netherlands. Retrieved from http://www.springerlink.com/index/10.1007/s11280-007-0019-8

[9] Manouselis N., Costopoulou C., 2008. Preliminary Study of the Expected Performance of MAUT Collaborative Filtering Algorithms, in Proc. of the First World Summit on "Emerging Technologies and Information Systems for the Knowledge Society (WSKS 2008)", Springer, CCIS 19, p. 527-536

[10] Manouselis, N., Vuorikari, R., 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), CEUR-WS Series, ISSN 1613-0073, Vol. 307, p. 27-35.

[11] Manouselis, N., Kyrgiazos, G., Stoitsis, G., 2012. Revisiting the Multi-Criteria Recommender System of a Learning Portal, in Proc. of the 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2012), 7th European Conference on Technology Enhanced Learning (EC-TEL 2012), Saarbrucken (Germany), CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 896, p. 35-48.

[12] Matsatsinis, N. F., Lakiotaki, K., Delias, P., 2007. A System based on Multiple Criteria Analysis for Scientific Paper Recommendation. *Decision Support Systems*. 11th Panhellenic Conference in Informatics. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.5021

[13] McNee, S., 2006. Meeting User Information Needs in Recommender Systems. PhD thesis, University of Minnesota, June 2006.

[14] Naak, A., Hage, H., Aimeur, E., 2008. Papyres: A Research Paper Management System. *2008 10th IEEE Conference on ECommerce Technology and the Fifth IEEE Conference on Enterprise Computing ECommerce and EServices*. IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4785064

[15] Paramythis A., Weibelzahl S., Masthoff J., 2010. Layered evaluation of interactive adaptive systems: framework and formative methods. User Modeling and User-Adapted Interaction. 20:383-453.

[16] Pu P., Chen L., Hu R., 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. User Modeling and User-Adapted Interaction, 22:317-355.

[17] Torres, R., McNee, S. M., Abel, M., Konstan, J. A., Riedl, J., 2004. Enhancing digital libraries with TechLens. *Proceedings of the 2004 joint ACMIEEE conference on Digital libraries JCDL 04*, *pp*, 228. ACM Press. Retrieved from http://portal.acm.org/citation.cfm?doid=996350.996402

[18] Totterdell, P., Boyle, E., 1990. The evaluation of adaptive systems. In Browne D., Totterdell P., Norman M. (Eds.): Adaptive User Interfaces. Academic Press; p. 161-194.
[19] Weibelzahl, S., 2001. Evaluation of adaptive systems. In Proc. 8th International Conference on User Modeling. Springer LNCS 2109; p. 292–294.