Je-LKS
PEER REVIEWED PAPERS
FOCUS ON RECOMMENDER SYSTEMS
Vol. 10, n.1, 2014
ISSN: 1826-6223 | eISSN: 1971-8829
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

LAYERED EVALUATION IN RECOMMENDER SYSTEMS: A RETROSPECTIVE ASSESSMENT

Nikos Manouselis[1], Charalampos Karagiannidis[2], Demetrios G. Sampson[3]

[1]Agro-Know Technologies, [2]University of Thessaly, [3]University of Piraeus & CERTH - Greece
nikosm@ieee.org, karagian@uth.gr, sampson@iti.gr

Keywords: Recommender systems, evaluation, adaptive systems, post-assessment

Evaluation of recommender systems has only lately started to become more systematic, since the emphasis has long been on the experimental evaluation of algorithmic performance. Recent studies have proposed adopting a layered evaluation approach, according to which recommender systems may be decomposed into several components, evaluating each of them separately. Nevertheless, there are still no evaluation studies of recommender systems that apply a layered evaluation framework to explore how all the different components or layers of such a system may be assessed. This paper introduces layered evaluation and examines how a previously proposed layered evaluation framework for adaptive systems can be applied in the case of recommender systems. It presents the possible adaptation of this layered framework that may fit the interaction components of recommender systems. Then, it focuses on a specific recommender system and carries out a retrospective analysis of its past evaluation results under the new

for citations:
Manouselis N., Karagiannidis C., Sampson D. G. (2014), *Layered Evaluation in Recommender Systems: A Retrospective Assessment,* Journal of e-Learning and Knowledge Society, v.10, n.1, 11-31. ISSN: 1826-6223, e-ISSN:1971-8829

prism that the layered evaluation approach brings. Our analysis indicates that implementing a layered-based evaluation of recommender systems has the potential to facilitate a more detailed and informed evaluation of such systems, allowing researchers and developers to better understand how to improve them.

1 Introduction

Adaptive Systems (AS) - also referred to as interactive adaptive systems (IAS) - analyse userinteractions with a system and modify the interface presentation or the system behaviour accordingly (Brusilovsky, 1996). According to Jameson (2001), a user-adaptive system is an interactive system which adapts its behaviour to each individual user on the basis of nontrivial inferences from information collected about that user. Research in AS begun as early as the 70s (when computers started to reach broader audiences), in an attempt to improve user interaction and stimulate acceptance through personalisation. Despite numerous intensive research efforts, many of the resulting AS systems remained at a prototype level and were not transformed into widely adopted commercial products. One of the main reasons for this under-exploitation relates partly to the difficulty of evaluating AS, and thus, generalising and reusing evaluation results across different applications (Brusilovsky *et al.*, 2004). Moreover, most of the early AS evaluation efforts followed a "with- and without-adaptation" approach (i.e. comparing the AS with its "non-adaptive" part), and could not provide reliable information about how the different components of an AS performed or how they should be improved.

Nevertheless, evaluation has always been considered of high importance in AS research. Starting with the early study of Brusilovksy & Eklund (1998) on the effects of adaptive link annotation, Paramythis *et al.* (2010) provide a comprehensive review of relevant work in the past fifteen years and explain why evaluation of AS is different to that of non-adaptive systems. As they explain, AS demonstrate a more complex behavior due to the nature of adaptivity and the fact that an AS is a highly interactive system. They point out that, in this capacity, adaptivity may require long-term, or even longitudinal studies, or be based on evaluation designs that explicitly account for that factor in order to avoid typical difficulties in comparing AS with their "static" counterparts - such as how to select the non-adaptive controls, how to select the appropriate equilibrium points, and how to take into consideration the dynamics of adaptive behaviour.

Recommender systems (RecSys) have early appeared as a type of system that "…*help(s) people make choices based on the opinions of other people*" (Goldberg *et al.*, 1992). The term initially covered systems mostly related to collaborative filtering but then was used in a more broader sense to refer to "...

any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options" (Burke, 2002; Burke & Ramezani, 2011). Even though this definition covers also the classic text-based filtering systems, Burke (2002) states that two criteria distinguish recommender systems from text-based ones: the criterion of 'individualisation' and the criterion of 'interesting and useful' content. This distinction reveals a direct link between AS and RecSys, since both of them are changing their behaviour to better match the individual user profile. The main difference is that AS usually take into account several contextual parameters collected from the system use (at runtime) before adaptations are implemented, whereas in typical RecSys, more rapid, data-driven personalisation approaches are used.

For instance, classic AS have a well-defined user model that allows distinguishing at least two separate layers - how well the system processes the collected data into a user model and then how well it adapts using this model. On the other hand, specific types of RecSys, such as collaborative filtering ones that are based on explicit ratings, have no user model in the classic sense. What they do store is raw user item preferences indicated as ratings, which can be considered as an input for user modelling. Still, reviews of research work in RecSys (Adomavicius & Tuzhilin, 2005; Manouselis & Costopoulou, 2007a) have revealed that rather complex user and domain models are being introduced and used in recommendation approaches, in a way very similar to the way classic AS represent and treat them. In this sense, RecSys seem to evolve into a special case of the broader category of AS, since both technologies aim to improve user experience through personalization; RecSys focus mainly on richer domain representations, whereas AS address more widely the aspects of user modelling and interaction (Paramythis *et al.*, 2010).

This explains why recent studies on the state-of-art of recommender systems' evaluation have either explicitly suggested the adoption of layered approaches (Pu *et al.*, 2012) or have developed evaluation frameworks that have a decomposition logic (Knijnenburg *et al.*, 2011; 2012). Layered evaluation (or decomposition) frameworks have attracted research attention in AS for more than a decade, with several frameworks, methods and instruments being proposed and tested in relevant literature (Paramythis *et al.*, 2010). They try to decompose an AS 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 RecSys evaluation study that has explicitly applied some particular layered evaluation framework and exploring how the different components or layers of a RecSys can be assessed.

In this paper, we aim to contribute to this direction, by examining how a previously proposed layered evaluation framework for AS can be applied to

RecSys. More specifically, we present and examine the potential adaptation of a specific layered evaluation framework, developed and tested in our previous studies, to fit the interaction components of RecSys. Then we focus on a particular case study of a RecSys and discuss the interpretation of its evaluation results from the perspective of this layered evaluation framework, aiming to demonstrate the added-value of this approach for the effective evaluation of RecSys. A number of conclusions are being made, also informing the further development, refinement and application of layered approaches for the evaluation of RecSys.

2 Background

2.1 Layered Evaluation for Adaptive Systems

As noted earlier, the evaluation of AS is a challenging task, since AS adapt dynamically during the same "interaction session" and can therefore break some of the most fundamental rules of usability and human-computer interaction, such as consistency, predictability and user control. Most early attempts to evaluate AS followed a "with- and without-adaptation" approach; that is, the "adaptive component" was "separated" from the system, and the two versions of the system (the one with adaptive features and the one without adaptive features) were compared to investigate whether adaptation brought significant benefits. This approach has a fundamental problem: the "non-adaptive" 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 adaptation component. Moreover, this approach is not useful when the adaptive system is found to be ineffective, since there is no way to understand why the AS (or which specific component of the AS) was not successful so as to improve it.

In this context, a series of layered evaluation frameworks have been proposed in the AS literature, advocating that each different AS component should be evaluated separately in order to get detailed information on the pros and cons of each part of the system. The idea can be traced back to the early 90s, when Totterdell & Boyle (1990) proposed that (i) the accuracy of the user model and (ii) the effectiveness of the changes (adaptations) made by the AS should be evaluated separately. Ten years later, Karagiannidis & Sampson (2000) proposed the term "layered evaluation", and suggested that AS evaluation should address the main components of each system separately.

Weibelzahl (2001) has proposed a similar layered framework, suggesting the decomposition of adaptation into the following three layers: (i) evaluation of input data, (ii) evaluation of the inference mechanism, and (iii) evaluation

of the adaptation decisions. Paramythis *et al.* (2010) further elaborated their decomposition by proposing five layers (or modules): (i) interaction monitoring, (ii) interpretation and interface, (iii) modelling, (iv) adaptation decision making, and (v) applying adaptations. Additional approaches that adopted to some extent a layered- or component-based approach were also proposed by Herder (2003), Magoulas *et al.* (2003), and Tobar (2003).

Reviewing the state of the art in related work, Paramythis *et al.* (2010) grouped together the main approaches and suggested the following main layers of adaptation: collection of input data; interpretation of the collected data; modelling the current state of the "world"; deciding upon adaptation; and applying (or instantiating) adaptation. They argued that these adaptation layers serve as the core components upon which evaluation can take place, aiming to "isolate" and evaluate separately as many as possible given the particularities of a given system.

2.2 Mapping Adaptive System Layers to the Interaction Model of Recommender Systems

Evaluation of RecSys has only lately started to become more systematic, since the emphasis has long been on the experimental evaluation of algorithmic performance (Herlocker *et al.*, 2004; Schroder *et al.*, 2011). Recent studies have proposed a more systematic approach to RecSys evaluation, separating the interactive from the non-interactive components of a recommender system (del Olmo & Gaudioso, 2008), distinguishing among types of evaluation studies and suggesting appropriate protocols (Shani & Gunawardana, 2011), and emphasizing on the user-perceived criteria to assess the success of a RecSys (Pu *et al.*, 2012; Knijnenburg *et al.*, 2012). All these approaches include to some extent a degree of decomposition of the RecSys into several components, bringing them close to the logic of typical layered evaluation approaches. Especially the evaluation framework of Knijnenburg *et al.* (2011; 2012) de-composes the user experience into different objective system objects that can then be separately evaluated using different experiments, protocols and metrics. This approach is very similar to the way layered approaches de-compose AS, although it takes the user perspective in consideration and does not cover system-level components that a typical recommender system analysis may include (Manouselis & Costopoulou, 2007a).

An extensive survey of evaluation issues in recommender systems was carried out by Pu *et al.* (2012). 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 pre-

sentation of the system's recommendation results. This decomposition is very close to the way that AS layered evaluation frameworks are decomposing an AS in separate components that can be evaluated one by one. Pu *et al.* (2012) have suggested that layered evaluation can be used in RecSys research as a powerful technique in identifying areas of a system that require further improvements.

More specifically, the three interaction steps that Pu *et al.* (2012) 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 behaviours (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.

In Table 1, a connection is made between the interaction steps of this generic model for recommender systems to the layers of the various evaluation frameworks proposed in AS literature.

Table 1
CONTRASTING PU *ET AL.* (2012) INTERACTION STEPS TO DECOMPOSITION LAYERS OF REPRESENTATIVE LAYERED EVALUATION FRAMEWORKS

<table>
<tr><th></th><th>Karagiannidis & Sampson (2000)</th><th>Weibelzahl (2001)</th><th>Paramythis et al. (2001)</th><th>Integrated framework of Paramythis et al. (2010)</th><th>Pu et al. (2012)</th></tr>
<tr><td rowspan="5">Layers</td><td rowspan="3">interaction assessment</td><td>evaluation of input data</td><td>interaction monitoring</td><td>collection of input data</td><td rowspan="3">elicit user preferences & revise user preferences</td></tr>
<tr><td rowspan="2">evaluation of the interface mechanism</td><td>interpretation / inferences</td><td>interpretation of collected data</td></tr>
<tr><td>modeling</td><td>modeling current state of world</td></tr>
<tr><td rowspan="2">adaptation decision making</td><td rowspan="2">evaluation of the adaptation decisions</td><td>adaptation decision making</td><td>deciding about adaptation</td><td rowspan="2">display recommendations</td></tr>
<tr><td>applying adaptation</td><td>applying adaptation</td></tr>
</table>

3 Adapting a Layered Evaluation Framework

As discussed in the previous section, the layered evaluation framework of

Karagiannidis & Sampson (2000) suggested that AS evaluation should address the two main components of each system separately:

- *interaction assessment* (i.e. whether the user model maintains an accurate profile of each user's characteristics), and
- *adaptation decision making* (i.e. whether the adaptation rules applied are effective).

The authors demonstrated the benefits of adopting such a layered evaluation framework, by re-visiting past evaluations of two AS, namely InterBook and KOD (Brusilovsky et al., 2004).

3.1 Mapping the layered framework to RecSys components

To investigate how this layered framework could be applied to RecSys evaluation, we focus on the interaction steps that Pu *et al.* (2012) have identified (the initial preference elicitation process, the preference refinement process, and the presentation of the system's recommendation results). More specifically, we study the basic decomposition layered that the framework suggests and try to map them to the interaction steps that Pu et al. have found as being generally applicable for RecSys. Viewing these steps as parts of the layered decomposition shown in Figure 1, we can suggest that:

All interactions related to the user preference profiles, i.e. the step of eliciting user preferences and the step of revising user preferences, correspond to the "assessment of interaction" component of the diagram, 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, correspond to the "adaptation decision making" component of the diagram, 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 the RecSys interaction activities based on the adaptation decomposition components of Karagiannidis & Sampson (2000), it could be argued that a layered approach would also apply for the evaluation of RecSys as follows:

- *Layer 1 - evaluation of user modelling*: at this layer the user modelling 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 modelling approaches, different model representation formats, as well as the evaluation of techniques to boost performance, such as the use of stereotypes to create an initial user model and to 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).

This analysis allows us to argue that the application of this layered framework could be possible for the case of recommender systems, since the majority of the interaction components of Pu *et al.* (2012) are covered by the framework.

Nevertheless, although the application of an AS layered decomposition in a RecSys may seem to be rather straightforward, the detail of decomposition is rather low compared to typical decompositions of information filtering systems (Hanani *et al.*, 2001) and recommendation systems (Manouselis & Costopoulou, 2007a; Knijnenburg *et al.*, 2012). We suspect that such a layered approach offers the opportunity to focus on various separate aspects of a RecSys in order to improve their individual performance. The large number of RecSys components to be considered, as well as the variety of evaluation techniques that can be used (Paramythis *et al.*, 2010; Shani & Gunawardana, 2011), still make this a complex and challenging task.

3.2 Towards an initial set of principles for constructing & using evaluation layers

To illustrate how a layered de-composition can serve as a starting point for the development of a more concrete and practical evaluation framework, we elaborate on the mapping of the layers proposed by Karagiannidis & Sampson to the RecSys components proposed by Pu et al., in order to provide some generic principles and guidelines that a RecSys researcher could use. The aim of this exercise is to demonstrate how such a generic framework can be built using the specific evaluation layers (user model and recommendation system),

rather than proposing a complete and detailed generic framework. It will help us inform the analysis of a specific case study in the section that follows.

More specifically, we focus on each layer and breakdown the interaction components to distinguishable elements – for example, separating the way that user preferences are elicited from the way that they are revised. Then, using an existing analysis of RecSys to various dimensions, we further analyse the interaction components to more fine-grained sub-components such as the way that the user model is being represented and the way that it is generated. This analysis can down to the level of granularity that the evaluation framework designers believe that it will provide meaningful results to the researchers. In Table 2, this is illustrated with a handful of RecSys dimensions from the many that the analysis framework of Manouselis & Costopoulou (2007a) identifies.

After the specific sub-components have been identified, what is needed is a suggestion of appropriate evaluation methods, protocols, metrics and instruments. There are several RecSys evaluation studies that may be used as a source for this information. In Table 2 we present two specific attributes: the suggestion of an appropriate evaluation method and a corresponding metric that may be used based on Shani & Gunawardana (2011).

Table 2
AN EXAMPLE OF HOW EVALUATION LAYERS AND COMPONENTS MAY BE ELABORATED TO SPECIFIC EVALUATION GUIDELINES THAT RESEARCHERS COULD APPLY

Evaluation layers (Karagiannidis & Sampson, 2000)	Interaction components (Pu *et al.*, 2012)	**RecSys dimension** (Manouselis & Costopoulou, 2007a)	Evaluation Method (Shani & Gunawardana, 2011)	Evaluation Metric (Shani & Gunawardana, 2011)
interaction assessment	Elicit user preferences	User Model Representation	Offline experiment, User study, Online evaluation	User Preference, Utility
	Revise user preferences	User Model Generation	Offline experiment, User study	User Preference, Utility
		User Model Update	Offline experiment, User study	Adaptivity
adaptation decision making	Display recommendations	Personalisation algorithm	Offline experiment	Prediction Accuracy, Coverage, Robustness, Scalability
		Personalisation output	User study, Online evaluation	User Preference, Trust, Novelty, Serendipity

Such a suggestion indicates that the representation of the user model may

be evaluated by:

- running an offline experiment (as we demonstrate later in the case study) that will try to measure e.g. the utility of the engaged user model,
- a focused user trial that e.g. may ask users about their desired way of representing their preferences,
- or an online evaluation where e.g. two different methods for representing the user model may be compared through an A/B test.

In a similar way, the sub-components responsible for updating the user model, for generating the recommendations, or for presenting the recommendations, may be tested using different methods and several possible metrics that Shani & Gunawardana (2011) propose.

4 Layered Interpretation of Evaluation Results

In this section, we want to present an actual case study of how such a layered decomposition may be applied in the case of an existing RecSys. We chose to analyse the existing results of an already evaluated system and to examine whether the use of the layered approach may help us interpret these results in a way that will provide useful insight for each studied component. More specifically, we present and discuss results from the evaluation of the system that serves as our case study according to the two layers proposed by Karagiannidis & Sampson (2000), in order to potentially:

- reach conclusions for each layer that may provide insight on how to improve the specific RecSys, and
- investigate the generalisation of the evaluation results for each different layer so that they can be used across different applications.

This is similar to the study of Brusilovksy *et al.* (2004) where the same framework was used to re-visit and re-process data from a previous study under the prism of layered evaluation. The RecSys into consideration is a multi-attribute utility (MAUT) collaborative filtering system that was introduced by Manouselis & Costopoulou (2007b), and has been experimentally tested in various occasions and contexts during the past few years (e.g. Manouselis *et al.*, 2007; Manouselis & Costopoulou, 2008; Manouselis *et al.*, 2012). This experimental investigation has produced several evaluation results which are revisited in this section.

4.1 Decomposing the MAUT Collaborative Filtering System

According to the analysis carried out in the previous section, the MAUT

RecSys of Manouselis & Costopoulou (2007b) can be decomposed into:

- 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.
- 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 personalised 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 focus on two specific sub-components for which we had relevant results from the previous experiments: the User Model Representation which is using multi-criteria ratings for the expression of user preferences; and the Personalisation Algorithm which is a multi-criteria extension of typical neighbourhood-based collaborative filtering ones.

4.2 Evaluating the MAUT User Model

For this sub-component, we focused on an offline experiment that used existing synthetic rating data. The user model used 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 modelling and, therefore, better recommendation results.

To evaluate how this user modelling 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 referred to an analysis that took place in Manouselis & Costopoulou (2008) and which compared how the number of dimensions (criteria) affected the performance of the system. In particular:

- A variety of synthetic (simulated) data sets have been created, including ratings of various properties, e.g. ranging from single-criterion to multi-criteria data sets and from very sparse to very dense ones.
- The algorithm of the MAUT collaborative filtering system has been

executed upon all data sets, and its performance has been measured using two metrics: the Mean Absolute Error (MAE) in the predictions and the number of items out of the data set for which a prediction was possible (coverage).

- The correlation between a number of data properties and the values of the measured performance results has been examined. The aim was to explore whether some of these data properties resulted in better or worse performance results.

In total, 243 synthetic data sets have been produced with characteristics ranging as illustrated in Table 3.The rows in the table show the minimum and the maximum values that each variable took, in order to give an idea of the experiment's scope. For example, the ratings in the 247 datasets ranged from 5 in total (very scarce dataset) to 2,500 ones (very dense one). Interested readers should refer to Manouselis & Costopoulou (2008) for more information on the experimental setup.

Table 3
RANGE OF PROPERTIES OF THE SYNTHETIC DATA SETS USED IN THE EXPERIMENT (FROM MANOUSELIS & COSTOPOULOU, 2008)

	Min	Max
Criteria	1 criterion	10 criteria
Evaluation scales	2-scale (binary)	10-scale
Items	100 items	1,000 items
Users	50 users	250 users
Evaluations/Ratings	5 evaluations	2,500 evaluations

Since at this stage of the layered approach we would focus only on the component related to the user preference profiles (level *i* in section 4.1), this post-analysis will study the particular data set properties that are related to the user model; that is, the number of criteria that the user may engage in rating an item and the number of rating scales used upon each criterion.

Tables 4 presents the Pearson correlation values of these two user model related properties to the examined performance metrics. From this type of analysis, the following type of hypotheses may be investigated for each pair of variables:

- *Number of criteria in user model and MAE*. A negative correlation between these two variables would mean that using multiple criteria provides better prediction accuracy (less MAE).
- *Number of evaluation scales and MAE*. A negative correlation would

mean that using a larger number of scales to represent the ratings may improve prediction accuracy.

- *Number of criteria and coverage*. A positive correlation would mean that the number of criteria improve the coverage of the algorithm.
- *Number of evaluation scales and coverage*. A positive correlation would mean that using a larger number of scales to represent the ratings may improve recommendation coverage.

Table 4
PEARSON CORRELATION OF EACH PROPERTY WITH EXAMINED METRICS (**SIGNIFICANT AT ALPHA LEVEL OF 0.01)

Data Set Property	Metric	Pearson	Sign. (2-tailed)
# of criteria	MAE	-0.139(**)	0.000
# of scales	MAE	0.289(**)	0.000
# of criteria	Coverage	-0.017	0.127
# of scales	Coverage	0.009	0.424
Evaluations/Ratings	5 evaluations	2,500 evaluations	2,500 evaluations

Unfortunately, in the results of the specific experiment that we analysed, the significance level itself has not been big and the effect size was rather small considering only two variables. For instance, the correlation between number of criteria and MAE is = -0.139, which is an effect size of *pho^2* = 0.019, i.e. we can say that the "number of criteria" explains 1.9% of the variance of the variable "MAE". This is an inherent weakness of the previous experiment that we revisit here and does not allow us to reach safe conclusions for the MAUT model. In addition, since in this experiment the use of artificially generated datasets cannot guarantee that they reflect "real" user data, this conclusion is only indicative of the type of observations that a layered analysis of the results allows – rather than a well-founded fact on the relations between the examined variables.

Still, our aim in this section is to view such past experimental results through the lens of layered decomposition. We believe that this is the type of experiment that will help RecSys researchers investigate the effect that the user model attributes have to the performance of the algorithms.

4.3 Evaluating the MAUT Recommendation Model

For this sub-component, we focused on results from a series of offline experiments that use available data from various application contexts. More specifically, we revisit some experimentation results that come from various previous studies (Manouselis & Costopoulou, 2007b; Manouselis *et al.*, 2007;

Manouselis *et al.*, 2012) and compare the MAUT collaborative filtering algorithms with some non-personalised basic ones. In this comparison, we study how the collaborative filtering algorithm performed over the following four data sets with multi-attribute ratings that come from real users (and within contexts that are different among them) in comparison to the non-personalised ones:

- A data set with evaluations that users give over agricultural e-markets and e-shops (557 ratings over 30 items from 255 users).
- A data set with evaluations that teachers give over the digital learning resources of the European Schoolnet portal (2,554 ratings over 899 items from 228 users).
- A data set with evaluations that agricultural educators and researchers provide over the Organic.Edunet web portal resources (477 ratings over 345 items from 99 users).
- A data set with evaluations that editorial peer-reviewing groups provide for selected resources of the MERLOT learning portal (2,626 ratings over 2,603 items from 18 "users").

Table 5 indicates that the personalized algorithms seem to be generally providing more accurate predictions of unknown ratings compared to the "predictions" made by (some even very naive) non-personalised algorithms, such as one that is providing a random number as a prediction or one that is randomly selecting a rating that some other user gave on the same item. One the other hand, the results presented in Table 6 indicate that some of the simplest algorithms can make a prediction in many more cases than the ones that the personalized algorithms may support – mostly because in the case of users with few ratings or very sparse data sets, there are not enough past ratings in order for predictions to be calculated.

Table 5
MAE RESULTS FOR THE COMPARED DATA SETS

Algorithm	Agricultural e-markets	European Schoolnet	Organic.Edunet	MERLOT
Pure Random	2.06	1.48	1.59	1.69
Random Exists	0.86	0.81	1.33	0.76
Arithmetic Mean	0.72	0.74	1.28	0.78
Geometric Mean	0.74	0.75	1.27	0.78
Deviation from Mean	0.76	0.74	1.03	0.45
Best MAUT collaborative filtering variation	0.24	0.57	0.99	0.22

Table 6
COVERAGE RESULTS FOR THE COMPARED DATA SETS

Algorithm	Agricultural e-markets	European Schoolnet	Organic.Edunet	MERLOT
Pure Random	100%	100%	100%	100%
Random Exists	100%	100%	100%	100%
Arithmetic Mean	100%	83.56%	37.89%	1.71%
Geometric Mean	100%	83.56%	37.89%	1.71%
Deviation from Mean	94.59%	81.41%	32.63%	1.71%
Top MAUT collaborative filtering variation	92.79%	69.08%	18.95%	0.95%

Viewing this under the prism of the layered decomposition, it could be an indication that the recommendation method and algorithm that has been chosen (a pure collaborative filtering one) may not be the best choice when the application context has sparse data. As it has happened with the revisited results of the previous section, this is an issue that was not identified in the first experimental evaluation of the MAUT system (Manouselis & Costopoulou, 2007b) where the data set was dense enough to provide very good performance results. It was rather revealed by this post-processing analysis that particularly examined the performance of the personalized method vs. the non-personalised ones, across various data sets. This could be an indication that an alternative recommendation model, such as a hybrid approach that will use a simple algorithm whenever the personalised one cannot produce a recommendation, might be a good solution to improve this shortcoming of the studied RecSys.

5 Discussion

In this section we try to reflect on the outcomes of this case study analysis and the type of conclusions that we managed (or did not manage) to reach by using the layered approach. The first layer of decomposition focused on the user preference profile, which has been constructed using a multi-attribute rating model. Evaluating this layer gave a very slight but quite interesting type of indication about something that was not discovered during the original study: that one particular parameter (e.g. the granularity of the rating scale) may affect the performance of the algorithms, thus putting possible bias into the experiments. This effect has been identified in relevant recommender system research (Sparling & Sen, 2011; Gena *et al.*, 2011), but still typical experimentation protocols of new algorithms often neglect this kind of investigation. In addition, the results of the evaluation of introducing a MAUT preference model provide indications towards the use of multi-criteria recommendation approaches.

The benefit of using the decomposition approach for the second layer of decomposition (the MAUT algorithm) has been the indication that in such offline experimentation it is important to go beyond the execution of candidate algorithms over a single or a couple of similar datasets, and to perform experiments that include and cover datasets from a variety of application contexts and users. Since this type of offline algorithmic testing is one of the most typical and widely used experiments found in recommender systems research (Herlocker *et al.*, 2004), insight provided by the layered evaluation is useful to drive and inform further experimentation.

It is important to state that in this paper we attempt a retrospective assessment of the results of previous experiments in order to show the potential of a layered approach. The fact that the specific results do not help us reach strong conclusions does not mean that the evaluation approach is weak. On the contrary, our retrospective analysis shows two specific examples of the experiments that may be performed in order to investigate the various sub-components, and by no means is an exhaustive one. For instance, if the scope of the algorithmic experiment was on the item characteristics that should be taken into consideration, the offline experiments could compare the performance of standard/base CF algorithms and standard/based content-based ones, using the rating info vs. the item metadata for the recommendation of the same items.

In this sense, our study is very similar to the study of Brusilovsky *et al.* (2004) who tried to revisit past evaluation results under the prism of their layered evaluation framework. To this end, the results are more informative and descriptive than prescriptive. As RecSys have differences compared to traditional AS, it would make sense to actually re-define (and possibly standardize) a layered evaluation framework instead of trying to adapt one of the existing frameworks since the differences (such as the lack of complex user models) make it wise to do analogy and not copy. Nevertheless, existing frameworks can be of extremely high value since:

They may guide the de-composition process in order to bring closer the layers identified in recent AS work (such as Paramythis *et al.*, 2010) with the components identified in RecSys analyses (such as Manouselis & Costopoulou, 2007a).

They may provide a pool of candidate methods, tools and criteria that RecSys evaluation researchers may use in order to select the ones appropriate for this context.

The layered evaluation framework that we used does not originally provide specific hints on the types of experiments that can be carried out for the different types of systems or the experimentation methods, tools and criteria.

Elaborating a layered framework for RecSys to this direction would make its application more straightforward and comparisons and generalizations easier; for example, suggesting such simulated experiments where the parameters of the user model are controlled by the researchers in order to measure the effect on algorithmic performance. The framework could include specific guidelines and protocols on the way specific experiments should be carried out, adopting and synthesizing suggestions from work such as the one of Shani & Gunawardana (2011) and Pu *et al.* (2012). In section 3 we gave an example of how such a framework may be built.

On the other hand, as our study has indicated, there are several challenges to be overcome. By revisiting the past evaluation results of a RecSys, we have observed that workstill needs to be carried out to make layered evaluation frameworks more detailed and specific, in order to become more specific and relevant to RecSys researchers and developers. The reason is that each RecSys component needs to be mapped to the right evaluation technique whichcan be an elaborate and time-consuming process. Having a framework that has a clear mapping between fine-grained system components and the various evaluation layers would allow the straightforward selection and application of the right techniques for each component. An initial effort has been made in Knijnenburg *et al.* (2012); However, an elaboration on the various system components and their connection to a variety of different evaluation techniques, protocols, instruments, and metrics that researchers may use still needs to be developed.

In addition, there is a need for empirical studies that will inform the enhancement of existing frameworks. Although layered approaches have been used for over a decade in AS evaluation, a recent note by Paramythis et al. highlighted that[1]:

- Most published works report summative evaluations, aiming to establish the extent to which the use of an adaptation method has improved the system. Often more scientific insight can be gained from formative evaluations that inform and guide the development process of adaptive systems.
- Most published works report on a single evaluation activity, often assessing the system as a whole only. More principled and rigorous forms of evaluation are possible, in which different system layers or components are evaluated separately, and more is learned about what causes success (or, more importantly in some cases, failure).
- Certain success criteria have received much more attention than others. For example, for recommender systems, the focus has often been on recall and precision, rather than serendipity, privacy and trust. A more holistic approach to evaluation is needed, including the consideration

[1] http://www.easy-hub.org/umuai/cfp.dot

of trade-offs between criteria. Metrics and methods for evaluating new criteria are also needed.

Conclusions

RecSys constitute one of the most successful cases in the effort towards improving user experience and satisfaction through personalisation. In this paper, we discuss the application of a layered evaluation framework for RecSys, revisiting past evaluation results through a layered evaluation view. Our analysis indicates that implementing a layered-based RecSys evaluation has the potential to facilitate a more detailed and informed evaluation of such systems, allowing researchers and developers to better understand how to improve them.

Building upon this analysis we can suggest some directions of future work in the direction of developing more concrete layered evaluation approaches for RecSys:

- *Need for coherent and systematic method ready for application*: current layered frameworks give suggestions of useful methods and instruments, but they are not specific enough to guide practical applications. It would be useful to have an out-of-the-box approach with pre-defined protocols and suggested experiments to carry out for each type of recommender system. Any additional support (such as decision trees) to help adopters chose the most appropriate tool for their system and setting would be of high value.
- *Decomposition needs to get better*: the five layers of Paramythis *et al.* (2010) need to be better connected to the components identified by information filtering and RecSys studies, such as Hanani *et al.* (2001) and Manouselis & Costopoulou (2007a). This will help the layers become more specific in terms of RecSys aspects and dimensions that they focus on, also allowing the suggestion of various techniques for each dimension and type.
- *Translate evaluation hints to concrete indicators*: even if a good layered approach is applied to evaluate different components of a RecSys, the way that the results can be combined into a set of (possibly measurable) indicators has not been yet provided. Such a set of indicators could help the RecSys researcher or developer to decide on the trade-offs of devoting resources to improve one dimension over the other, and to have an overall and comparable view of the outcomes of an evaluation compared to a similar one.
- *Connect with evaluation guidelines and recommendations*: there are several existing frameworks in AS that suggest relevant methods and techniques for the evaluation of each layer. In a similar sense, there are

also surveys and studies of relevant evaluation approaches for recommender systems, including practical guidelines and recommendations. This existing body of knowledge should be carefully studied and combined in order to equip new frameworks with suggested methods, tools and instruments that would fit each component.

Overall, further work towards the standardization of the layered evaluation frameworks applied to RecSys should be expected to make it possible to facilitate the comparison and generalization of research results, and their reuse across different application domains.

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.

REFERENCES

Adomavicius G., Tuzhilin A. (2005), *Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*, IEEE Trans. Knowl. Data Engin. 17(6), 734-749.

Brusilovsky P. (1996), *Methods and Techniques of Adaptive Hypermedia*. User Modeling and User Adapted Interaction. 6 (2-3), 87-129.

Brusilovsky P., Karagiannidis C., Sampson D. (2004), *Layered Evaluation of Adaptive Learning Systems*. International Journal of Continuous Engineering Education and Lifelong Learning. 14 (4-5), 402-421.

Brusilovsky, P. & and Eklund, J. (1998), *A Study of User Model Based Link Annotation in Educational Hypermedia*. Journal of Universal Computer Science. 4 (4), 429-448. Springer Science Online.

Burke R. (2002), *Hybrid Recommender Systems: Survey and Experiments*. User Modeling and User-Adapted Interaction, 12, 331-370.

Burke R., Ramezani M. (2011), *Matching Recommendation Technologies and Domains*. in Ricci F., Rokach L., Shapira B. (Eds.), Recommender Systems Handbook, Springer. 367-386, US.

del Olmo F. H., Gaudioso E. (2008), *Evaluation of recommender systems: A new approach*. Expert Systems with Applications, 35, 790-804.

Gena, Cristina, Brogi, Roberto, Cena, Federica and Vernero, Fabiana (2011), *The Impact of Rating Scales on User's Rating Behavior*. In: Proceedings of the 2011 Conference on User Modeling, Adaptation and Personalization 2011. 123-134.

Goldberg D., Nichols D., Oki B.M., Terry D. (1992), *Using Collaborative Filtering to*

Weave an Information Tapestry. Communications of the ACM, 35(12), 61-70

Hanani U., Shapira B., Shoval P. (2001), *Information Filtering: Overview of Issues, Research and Systems*. User Modeling and User Adapted Interaction, 11, 203-259.

Herder, E. (2003), *Utility-Based Evaluation of Adaptive Systems*. In: Weibelzahl, S. and Paramythis, A. (eds.), *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems*, held at the 9th International Conference on User Modeling UM2003, Pittsburgh, 25-30.

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

Jameson, A. (2001), *Systems That Adapt to Their Users: An Integrative Perspective*, Saarbrücken: Saarland University.

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 (343-346).

Knijnenburg, B.P., Willemsen, M.C., Kobsa, A. (2011), *A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems*. Short paper at the ACM Conference on Recommender Systems (RecSys).

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C. (2012), *Explaining the User Experience of Recommender Systems*. User Modeling and User-Adapted Interaction (UMUAI).

Magoulas, G. D., Chen, S. Y. and Papanikolaou, K. A. (2003), *Integrating Layered and Heuristic Evaluation for Adaptive Learning Environments*. In: Weibelzahl, S. and Paramythis, A. (eds.). Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh, 5-14.

Manouselis N., Costopoulou C. (2007a), *Analysis and Classification of Multi-Criteria Recommender Systems*. World Wide Web. 10, 415-441.

Manouselis N., Costopoulou C. (2007b), *Experimental Analysis of Design Choices in multiattribute Utility Collaborative Filtering*. International Journal of Pattern Recognition and Artificial Intelligence, 21, 311-331.

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, 527-536

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, 896, 35-48.

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, 307, 27-35.

Paramythis A., Totter A., Stephanidis C. (2001), *A modular approach to the evaluation of adaptive user interfaces*. In Proc. 1st Workshop on Empirical Evaluation of Adaptive Systems (UM2001), Sonthofen, Germany, 9-24.

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.

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.

Schroder G., Thiele M., Lehner W. (2011), *Setting Goals and Choosing Metrics for Recommender System Evaluations*. In Joint proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems (Decisions@ RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces-2 (UCERSTI 2), CEUR Workshop Proceedings, ISSN 1613-0073, 811, 78-85.

Shani G., Gunawardana A. (2011), *Evaluating Recommendation Systems*. in Ricci F., Rokach L., Shapira B. (Eds.), Recommender Systems Handbook, Springer, 257-297.

Sparling E.I., Sen Sh. (2011). Rating: how difficult is it?. In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11). ACM, New York, NY, USA, 149-156.

Tobar, C. M. (2003), *Yet Another Evaluation Framework*. In: Weibelzahl, S. and Paramythis, A. (eds.). Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh, 15-24.

Totterdell P., Boyle E. (1990), *The evaluation of adaptive systems*. In Browne D., Totterdell P., Norman M. (Eds.): Adaptive User Interfaces. 161-194, Academic Press.

Weibelzahl S. (2001). Evaluation of adaptive systems. In Proc. 8th International Conference on User Modeling. Springer LNCS 2109 292–294