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

EXPLORATORY STUDY OF MULTI-CRITERIA RECOMMENDATION ALGORITHMS OVER TECHNOLOGY ENHANCED LEARNING DATASETS

Nikos Manouselis[1], Giorgos Kyrgiazos[2], Giannis Stoitsis[1]

[1]Agro-Know Technologies, [2]Computer Technology Institute and Press - Diophantus
nikosm@ieee.org, kyrgiazos@cti.gr, stoitsis@ieee.org

Keywords: Multi-criteria recommender system, testing, web portal

Results of previous studies have indicated that the performance of recommendation algorithms depends on the characteristics of the application context. The same algorithms have shown to be performing in totally different ways when a new or evolved data set is considered, thus leading to a need for continuous monitoring of how they operate in a realistic setting. In this paper we investigate such a real life implementation of a multi-criteria recommender system and try to identify the needed adjustments that need to take place in order for it to better match the requirements of its operational environment. More specifically, we examine the case of a multi-attribute collaborative filtering algorithm that has been supporting the recommendation service within a Web portal for organic and sustainable education. Our study particularly explores the experimental performance of the already implemented algorithm, as well as an alternative one, using

for citations:
Manouselis N., Kyrgiazos G., Stoitsis G. (2014), *Exploratory study of multi-criteria recommendation algorithms over technology enhanced learning datasets,* Journal of e-Learning and Knowledge Society, v.10, n.1, 33-49. ISSN: 1826-6223, e-ISSN:1971-8829

data from the intended application, a simulated expansion of it, and from similar portals. The results of this study indicate the importance of the frequent experimental investigation of a recommender system's various design options, and the need for the exploration of adaptive implementations in real life recommender systems.

1 Introduction

The real world implementation of a recommender system is much more than a piece of software that is already complex since it is made by many interacting parts, from data integration to results' presentation (Picault et al., 2011). It is part of a much more complex environment, with lots of uncertainty and little control. Although there has been some very interesting and useful work in literature suggesting ways in which recommender systems may be adapted to specific domains (Burke & Ramezani, 2011), evaluated systematically and extensively (Herlocker et al., 2004; Shani, 2011), and deployed in real (and outside the lab) settings (Picault et al., 2011), monitoring the operation and performance of a recommender system in its actual environment is a challenging task. In the domain of education, recommender systems have been introduced more than a decade ago, with deployed and well-studied systems like Altered Vista (Recker et al., 2003) and CoFIND (Dron et al., 2000). Still, surveys of the systems that have been actually implemented in a real life setting still show that many of these systems are not extensively tested (Manouselis et al., 2013).

In this paper, we try to reflect on one of the main questions that the people responsible for an operational recommender system need to face: how can we monitor, test, and fine-tune the algorithms deployed in a real setting, by using data from its actual operation as well as from similar systems. More specifically, we focus on the case of an existing educational recommender system that collects data that educators and learners provide on digital content that may be used to support education and research on organic and sustainable agriculture, and uses them to provide recommendations about relevant resources (Manouselis et al., 2009). Our study particularly focuses on the collaborative filtering algorithm that has been chosen and parameterized to collect multi-criteria ratings on the content items in order to recommend new ones to the users.

2 Background & Theory

2.1 Multi-Criteria Recommendation

The problem of recommendation has been identified as the way to help individuals in a community to find information or items that are most likely to be interesting to them or to be relevant to their needs (Adomavicius & Tuzhilin, 2005). Typically, it assumes that there is set *Users* of all the users of a system and set *Items* of all possible items that can be recommended to them. Then, the utility function that measures the appropriateness of recommending item *iϵItems* to user *uϵUsers* is often defined as *R: Users × Items → R0*, where *R0* typically is represented by non-negative integers or real numbers within a certain range. It is assumed that this function is not known for the whole *Users × Items* space but is specified only on some subset of it. Therefore, in the context of recommendation, we want for each user *uϵUsers* to be able to (a) estimate (or approximate) the utility function *R(u,i)* for item *iϵItems* for which *R(u,i)* is not yet known, and (b) choose one or a set of items *i* that will maximize *R(u,i)*, i.e.,

$$\forall u \in Users\text{ , i= } \underset{i \in Items}{\arg\max}\, R(u,i) \tag{1}$$

In most recommender systems, the utility function usually considers a single-criterion value, e.g., an overall evaluation or rating of an item by a user. In recent work, this assumption has been considered as limited (Adomavicius & Kwon, 2007; Adomavicius *et al.*, 2011), because the suitability of the recommended item for a particular user may depend on more than one utility-related aspect that the user takes into consideration when making the choice. Particularly in systems where recommendations are based on the opinion of others, the incorporation of multiple criteria that can affect the users' opinions may lead to more accurate recommendations. Thus, the additional information provided by multiple dimensions or criteria could help to improve the quality of recommendations because it would be able to represent more complex preferences of each user. Recommender systems have already adopted multiple criteria as relevant research indicates. A recent survey by Adomavicius *et al.* (2011) identified more than fifty (50) such systems that can be broadly classified as multi-criteria recommender ones.

2.2 Multi-Criteria Collaborative Filtering

Multi-criteria collaborative filtering is an extension of traditional collaborative filtering systems that is based on ratings expressed over multiple dimensions describing an item. They allow a user to specify her individual preferences by rating each item on multiple criteria and recommend to the user the items that can best reflect the user's individual preferences based on the multi-criteria ratings provided by this and other users. In single-attribute (or single-criterion) collaborative filtering, the problem space can be formulated as a matrix of users versus items (or user-rating matrix), with each cell storing a user's rating on a specific item. Then, the system may calculate and provide recommendations. First, multi-criteria collaborative filtering aims to predict the utility of items for a particular user (called active user). To do so, it is based on the items previously evaluated by other users. That is, the utility *R(a,i)* of item *i* for the active user *aϵUser* is estimated based on the utilities *R(u,i)* assigned to item *i* by those users *uϵUser* who are "similar" to user a.

The utility function *R(u,i)* is then referred to as the total utility of an item *s*, which is calculated by synthesizing the partial utilities of the item s on each one of the criteria. Assuming that there is no uncertainty during the decision making, the total utility of an item *iϵItems* for a user *uϵUser* is often expressed as an additive value function, such as:

$$R(u,i) = \sum_{1}^{k} g(u,i) \tag{2}$$

The collaborative filtering techniques that use multi-criteria ratings to predict an overall rating and/or individual criteria ratings can be classified by the formation of the utility function into two categories: heuristic-based (sometimes also referred to as memory-based) and model-based techniques (Adomavicius *et al.*, 2011). Heuristic-based techniques compute the utility of each item for a user on the fly based on the observed data of the user and are typically based on a certain heuristic assumption. In contrast, model-based techniques learn a predictive model, typically using statistical or machine-learning methods that can best explain the observed data, and then use the learned model to estimate the utility of unknown items for recommendations.

2.3 Case Study

In this paper, we focus on the particular case of a real life implementation of a multi-criteria recommender system in the context of an educational application. This is the case of the Organic.Edunet Web portal for agricultural and sustainable education (http://www.organic-edunet.eu) that was launched in 2010. Its aim has been to facilitate access, usage and exploitation of digital educational content related to Organic Agriculture (OA) and Agroecology (AE). In order to achieve this aim, it networked existing collections with educational content on relevant topics from various content providers, into a large federation where content resources are described according to standard-complying metadata. The recommendation service in Organic.Edunet is supported by two separate algorithms that are using different data as input and are currently running independently (Manouselis *et al.*, 2009): a content-based recommender using tags and textual reviews as input; and a multi-criteria collaborative filtering system that uses as input the ratings that users provide over three criteria: *Subject Relevance, Educational Value* and *Metadata Quality*.

This study focuses on the multi-criteria algorithm and the recommendations that it produces. This algorithm was proposed by Manouselis & Costopoulou (2007) as a multi-criteria extension to typical heuristic neighborhood-based algorithms that may be found in the collaborative filtering literature. It follows the generic steps of Herlocker *et al.* (2002) in order to calculate a prediction:

- *Stage A - Similarity Calculation*: similarity between the examined user (active user) and the rest users is calculated using some similarity measure;
- *Stage B - Feature Weighting*: to further weight similarity according to the characteristics of each examined user or some heuristic rules;
- *Stage C - Neighborhood Formation/Selection*: selection of the set of users to be considered for producing the prediction;
- *Stage D - Combining Ratings for Prediction*: normalizing the ratings that the users in the neighborhood have provided for the unknown item, and using some method to combine them in order to predict its utility for the active user.

A variety of parameters and options for defining and fine-tuning them have been investigated by Manouselis & Costopoulou (2007). The multi-criteria extension of Manouselis & Costopoulou (*op cit.*) used in Organic.Edunet, was called the *Similarity per evaluation* (PG). Since the implementation of the recommendation algorithm took place during the design stage of the portal, we based our selection on the experience from a lab testing experiment that took using an existing data set from another learning portal (Manouselis *et al.*, 2010). Results of the simulated execution of more than 360 variations of the PG algorithm over this data set (selecting and fine-tuning various parameters) indicated that it would make sense to implement a version that uses a Cosine/Vector distance function to measure similarity between users, a Correlation Weight Threshold (CWT) of users with similarity slightly more than 0.5 for the neighborhood, and calculates predicted ratings as a weighted mean of the ratings that the neighbors have given over an unknown item. This variation has shown to achieve a Mean Absolute Error (MAE) over the prediction of less than 0.7 (in a scale 1-5) and a coverage close to 70% of the items.

Nevertheless, the fact that the specific algorithm variation performs well over a data set coming from a similar application context (that is, of a portal with learning resources) does not mean that it will also perform well during the operation of the Organic.Edunet portal. There are several reasons for this:

- The properties of the users vs. items matrix of Organic.Edunet may be different than the ones of the data set of the other application.
- The properties of the Organic.Edunet matrix may change/evolve with time.

- Alternative algorithms (e.g. new ones proposed in literature) that were not included in the initial experimentation may prove to perform better than the one selected.

To this end, we decided to repeat the experimental investigation of candidate algorithms for the Organic.Edunet portal, using additional algorithms that support multi-criteria recommendation, as well as different data sets that include multi-criteria ratings over educational content. Preliminary results from this experiment have been presented and discussed with the community in a relevant workshop (Manouselis *et al.*, 2012) are presented and discussed here in their full extend.

3 Methodology

3.1 Experimental Setting

The main goal of the experimental testing has been to investigate the experimental performance of different variations of both the algorithm currently implemented in Organic.Edunet as well as alternative multi-criteria recommendation algorithms. The specific objectives have been:

- To use a current instance of the users vs. items matrix of Organic.Edunet in order to execute all candidate variations and measure their expected performance.
- To use other multi-criteria rating data sets from similar educational portals, in order to execute the candidate algorithms and see if estimated performance changes.
- To generate a synthetic data set that will mimic an instance of the Organic.Edunet community in the future, in order to explore if performance of the candidate algorithms would be expected to change in the future.

The evaluation protocol follows the typical steps of offline experiments with pre-collected or simulated data that Shani & Gunawardana (2011) also described for testing the performance of candidate algorithms. Generally speaking, our experiment follows the approach of similar experiments in other domains (Herlocker *et al.*, 2004) or education (Sicilia *et al.*, 2010). The following paragraphs describe the settings, methods and tools of the experimental investigation.

3.1.1 Simulation tool/environment

The offline experiment took place using a software environment that has been specifically developed and used for the simulation of multi-criteria recommender systems, called the Collaborative Filtering Simulator (CollaFiS) (Manouselis & Costopoulou, 2006). CollaFiS is a software environment that provides both a graphical interface for researchers, as well as handles multi-criteria rating datasets. It is an integrated web application for performing experiments with ratings data sets for recommender systems. The graphical interface can be used for parameterising the algorithms, implementing a number of basic but also multi-criteria algorithms that can be tested.

This environment allows for importing various data sets, parameterizing candidate algorithms, executing them and measuring expected performance using multiple performance metrics. The CollaFiS environment has been extended to support the algorithms and metrics that are particularly studied in this experiment, as described later. The simulation took place in a personal computer with an Intel Core i7-2720QM (2.2 GHz, 8GB RAM; Intel Corporation, Santa Clara, CA) running LUbuntu 10.10, Apache2 server, PHP 5.3.3-1, and MySQL Server 5.1.61.

3.1.2 MAUT algorithm variations

CollaFiS provides the option for experimentally testing the multi-criteria algorithms proposed by Manouselis & Costopoulou (2007). We have extended the previous implementation of CollaFiS in order to also include the algorithms proposed by Adomavicius & Kwon (2007). Overall, the studied algorithms included:

- The *Similarity per evaluation* (PG) algorithm (currently implemented in Organic.Edunet) that calculates similarity separately upon each criterion, predicts the rating also separately upon each criterion, and then is synthesizing the predictions into a total predicted utility;
- The *Average Similarity* (AS) and *Minimum or Worst-case Similarity* (WS) algorithm versions proposed by Adomavicius & Kwon (2007) that use either the average or the minimum of the similarities over each criterion in order to calculate the total predicted utility;
- Some *Non-personalised algorithms* as a basic comparison, such as giving random values as predictions or calculating an arithmetic, geometrical or deviation-from-mean weighted sum of all past evaluations.

For the personalized algorithms (PG, AS, WS) we have considered the following design options in order to study different variations:

- During *Stage A - Similarity Calculation*: examined the calculation of the similarity using the Euclidian, Vector/Cosine, and Pearson distance functions as options.
- During *Stage C - Neighborhood Formation/Selection*: examined both the

use of a Correlation Weight Threshold (CWT) for the similarity value as a selector of potential neighbors, as well as of an absolute value for the Maximum Number of Neighbors (MNN).

- *During Stage D - Combining Ratings for Prediction*: examined three different options for synthesizing partial utilities, i.e. calculating the prediction as a simple arithmetic mean, as a mean weighted by the similarity value, as well as a normalized weighted mean that takes into consideration also the deviation from the arithmetic mean as (Herlocker *et al.*, 2002).

This led to 18 variations of each examined algorithm. By also experimenting with various values for the CWT (20 variations between '0' and '1' as a threshold) and MNN (20 variations using '1' to '20' maximum neighbors) parameters, the number grew to more than 1,080 algorithmic variations explored in total.

3.1.3 Metrics

There are several performance metrics used in the literature. In this experiment we examined the following evaluation metrics that CollaFiS incorporates:

- *Accuracy*: to measure the predictive accuracy of the multi-criteria algorithms, we calculated the mean-absolute error (MAE). MAE is the most frequently used metric when evaluating recommender systems. Herlocker *et al.* (2004) have demonstrated that since it is strongly correlated with many other proposed metrics for recommender systems, it can be preferred as easier to measure, having also well understood significance measures.
- *Coverage*: to measure the coverage of the multi-criteria algorithms, we calculated the items for which an algorithm could produce a recommendation, as a percentage of the total number of items. Herlocker *et al.* (*op. cit.*) recommend the measurement of coverage in combination with accuracy.

Furthermore, a new *combined* metric was also calculated. The idea has been to introduce a single indicator, which would have a heuristic but practical value, and that would combine and normalize measurements using the two metrics in order to allow the straightforward comparison of results coming from the execution of algorithms over different datasets. In addition, we wanted to give to system designers the ability to easily modify and fine-tune the importance that each individual metric had in this combined indicator, by using an arbitrarily specified weight factor. This combined metric is defined as:

$$Combo = (a \cdot \text{cov}) + \left(b \cdot \frac{1}{MAEn}\right) \tag{3}$$

where *cov* is the coverage of the algorithm, a and b two weighting factors that allow designers to define the weight of each metric in the combined one (with *a*+*b*=1).

MAEn is a normalized version of the error so that it is depicted in the [0,..,1] space, calculated as:

$$MAEn = \frac{MAE - d_{zero}}{scale} \tag{4}$$

where d_{zero} is the distance of the lower value of the rating scale from "0" and scale is the number of rating values.

For the needs of our experiment, we assumed that all criteria are evaluated using a similar rating scale with discreet values. This was the case for the data sets examined. The weighting values used for the Combo metric have been a=0.25 and b=0.75, illustrating the importance that coverage has for a real life application (thus being able to produce recommendations for a high number of unknown items).

3.2 Data Sets

Four different data sets have been used to support the simulated execution of the algorithms. All data sets have been imported into CollaFiS and appropriately processed. To facilitate the execution of the experiments, they have been split into one training and one testing component (using an 80%–20% split).

The first data set (*OEreal*) has been a recent export/instance of the users vs. items matrix of Organic.Edunet, with the collected multi-criteria ratings that the users of the portal have provided over the content items. As mentioned before, Organic.Edunet collects user evaluations over three criteria that are all rated using a discreet scale from 1 to 5. In this real data sets, 99 users have provided 477 multi-criteria ratings over 345 items.

The second data set (*EUN*) has been the one that has served as experimentation input during the initial studies (Manouselis *et al.*, 2010). It comes from a teacher portal of the European Schoolnet, a network of European Ministries of Education. This portal collected user ratings over six criteria, such as ease of integration in classroom, relevance to teaching topics, ability to help students learn etc. (detailed in Manouselis *et al.*, 2010), evaluated in a scale of 1 to 5.

This data set includes 2,554 multi-criteria ratings over 899 learning resources, which have been provided by 228 users.

The third data set (*MERLOT*) comes from the Multimedia Online Resource for Learning and Online Teaching (www.merlot.org), a very popular US portal for education. The portal collects peer-review evaluations that come from expert committees, specialized for each thematic area (by eighteen specialized Editorial Boards). Ratings over three criteria (quality of content, potential effectiveness as a teaching-learning tool, ease of use) are collected using a scale from 0 to 5. Since the evaluations cannot be distinguished on an expert level but only on an Editorial Board level, when analyzing this data set we make the assumption that each thematic Editorial Board will be treated as a separate user that can be recommended relevant resources for this thematic area. Thus, this data set includes 2,626 multi-criteria ratings from the 18 “users” over 2,603 resources (since almost all considered items have been peer reviewed).

Finally, to be able to study a state of the users vs. items matrix of Organic. Edunet in a future setting, a simulated data set (*OEsim*) was also generated. More specifically, the distributions of the ratings of the OAreal data set were taken as input to a Monte Carlo generator of random multi-criteria ratings of the same users. The considered scenario is that the current users that have been rating a sample of the Organic.Edunet items provide more ratings on this specific sample of already rated items in order to make it more dense. The produced synthetic data set incorporates the original real one, has the same number of users and items, and includes a total number of 1,280 multi-criteria ratings. In a similar way, alternative scenarios could be considered, with more users and/or items, and with more dense or sparse data sets.

4 Results

In this section we will present the results that have been produced by the CollaFiS tool, after executing all the studied variations of the algorithms.

4.1 Organic.Edunet Real

The execution of the candidate algorithms over the OEreal data set did not provide very good results. It seems that the majority of the tested variations performed a bit better than the non-personalised algorithms but still with a very low coverage that was in the vicinity of 16%-18% of the items for which a prediction needed to be made. As Table 1 shows, there are some variations (like the PG Cosine Deviation-from-Mean with both MNN and CWT parameters) that had an acceptable MAE that is below ‘1’. Still, we consider this error to be rather high for an operational recommender system. These results imply that the performance of any algorithm would be judged not satisfactory if only the current data set of Organic.Edunet was used for experimentation. The PG variations seemed to be generally performing better than the AS and WS ones. Nevertheless, this performance seems to be rather low over the OEreal data set.

Table 1
TOP-5 CWT AND TOP-5 MNN VARIATIONS OVER THE OEREAL DATA SET

Alg.	Similarity	Norm/tion method	AVG Cov.	AVG MAE	AVG Combo
MNN variations					
PG	Cosine	Deviation-from-Mean	18.95%	0.9928	0.1425
PG	Euclidian	Simple Mean	18.95%	1.3194	0.1300
PG	Cosine	Weighted Mean	18.95%	1.3337	0.1296
AS	Euclidian	Deviation-from-Mean	18.95%	1.5008	0.1254
WS	Cosine	Deviation-from-Mean	18.95%	1.6886	0.1217
CWT variations					
PG	Cosine	Deviation-from-Mean	16.32%	0.8650	0.1322
PG	Cosine	Simple Mean	16.32%	1.1831	0.1157
AS	Cosine	Deviation-from-Mean	16.95%	1.5202	0.1100
WS	Cosine	Deviation-from-Mean	16.37%	1.7316	0.1017
AS	Cosine	Simple Mean	16.95%	2.1074	0.1009

4.2 EUN

The execution of the candidate algorithms over the *EUN* data gave some better performance results. The coverage of the non-personalized algorithms was still higher than the examined ones, but there were some candidate variations that perform very well. More specifically, most of the variations had more than 64% coverage; some had very good MAE measurements. PG Cosine Deviation-from-Mean is again performing very well, with a low MAE of around 0.57 for both CWT and MNN parameters.

If the selection was made only based on the *EUN* data set, this one would probably be the variation of the algorithm that we would chose to implement, most probably in its MNN variation that seems to be rather stable in its performance.

Table 2
TOP-5 CWT AND TOP-5 MNN VARIATIONS OVER THE EUN DATA SET

Alg.	Similarity	Norm/tion method	AVG Cov.	AVG MAE	AVG Combo
MNN variations					
PG	Cosine	Deviation-from-Mean	69.08%	0.5721	0.5555
PG	Euclidian	Simple Mean	69.08%	0.6802	0.5416
AS	Cosine	Deviation-from-Mean	69.08%	2.6172	0.4872
AS	Euclidian	Deviation-from-Mean	69.08%	2.8330	0.4858
WS	Cosine	Weighted Mean	69.08%	3.2138	0.4837
CWT variations					
PG	Cosine	Deviation-from-Mean	64.02%	0.5721	0.5177
PG	Cosine	Weighted Mean	64.02%	0.6782	0.5039
AS	Cosine	Deviation-from-Mean	65.84%	2.6137	0.4629
AS	Cosine	Simple Mean	65.84%	3.2102	0.4594
WS	Cosine	Weighted Mean	64.09%	3.3173	0.4457

4.3 MERLOT

The case of the *MERLOT* data set has been a bit special, due to the assumption that we made about the grouping of expert users into 18 “editorial” users. The results of the execution over this very dense data set have been rather disappointing. All variations (including the non-personalised ones) seemed to produce a very low coverage that was not more than 0.95%. Practically, this means that the algorithms cannot make a prediction about any item in the data set. The reason is that most of the items have received only one rating by one “user” (i.e. Editorial Board), thus making it impossible to predict how another “user” would rate them.

4.4 Organic.Edunet Synthetic

The execution of the candidate algorithms over the synthetic *OEsim* data set seemed to perform much better than the original *OEreal* one, as one would have expected (since a more dense version of the data set has been created). The majority of the outstanding variations produced a rather good coverage that is close to (for CWT) and more than (for MNN) 60%.

Table 3
TOP-5 CWT AND TOP-5 MNN OVER THE OESIM DATA SET

Alg.	Similarity	Norm/tion method	AVG Cov.	AVG MAE	AVG Combo
MNN variations					
PG	Euclidian	Simple Mean	0.6133	0.8626	0.4679
PG	Cosine	Simple Mean	0.6133	0.8653	0.4678
PG	Cosine	Deviation-from-Mean	0.6133	0.8855	0.4664
AS	Euclidian	Deviation-from-Mean	0.6133	1.2972	0.4485
WS	Euclidian	Deviation-from-Mean	0.6133	1.8486	0.4370
CWT variations					
PG	Cosine	Simple Mean	0.5791	0.8673	0.4420
PG	Cosine	Weighted Mean	0.5791	0.8681	0.4419
PG	Cosine	Deviation-from-Mean	0.5791	0.8908	0.4405
AS	Cosine	Deviation-from-Mean	0.5934	1.2983	0.4335
AS	Cosine	Weighted Mean	0.5934	2.2086	0.4177

Surprisingly, the MAE results seem to be at the level of the non-personalised algorithms and around 0.86 for the PG MNN Euclidian Simple Mean and the PG CWT Cosine Simple Mean. It seems that very simple algorithms that create weighted sums of the past ratings, such as the Arithmetic Mean and the Geometrical Mean have been found to provide predictions that have less MAE than the collaborative filtering variations. This could be due to the fact that the simulated users have provided additional ratings with similar distributions but still such a simplistic interpretation of this observation is not enough. In most cases the PG variations seem to be performing better than the AS and WS ones, although the differences are small.

To further investigate which would be the more appropriate algorithm variations to support recommendation in Organic.Edunet in such a future scenario, we did an additional experimental analysis. More specifically, we investigated the performance that the algorithms that performed well on all three real data sets (i.e. *OEreal*, EUN, MERLOT) had over the synthetic OEsim.

Table 4
TOP-5 CWT AND TOP-5 MNN OVER THE OESIM DATA SET FROM VARIATIONS PERFORMING BETTER OVER THE THREE REAL DATA SETS

Alg.	Similarity	Norm/tion method	AVG Cov.	AVG MAE	AVG Combo
MNN variations					
PG	Cosine	Deviation-from-Mean	61.33%	0.8855	0.4664
PG	Euclidian	Simple Mean	61.33%	0.8626	0.4679
PG	Pearson	Simple Mean	61.33%	0.8646	0.4678
AS	Euclidian	Deviation-from-Mean	61.33%	1.2972	0.4485
WS	Cosine	Deviation-from-Mean	61.33%	1.8514	0.4370
CWT variations					
PG	Cosine	Deviation-from-Mean	57.91%	0.8908	0.4405
PG	Cosine	Simple Mean	57.91%	0.8673	0.4420
AS	Cosine	Deviation-from-Mean	59.34%	1.2983	0.4335
AS	Cosine	Simple Mean	59.34%	2.2090	0.4177
WS	Cosine	Weighted Mean	57.91%	2.5190	0.4042

In this way we tried to see if some of the algorithm variations that performed in a good way through such diverse applications like the ones from which we collected the testing data, would also perform in a similar way over the synthetic data set for the Organic.Edunet application. As illustrated in Table 3, these algorithm variations seemed to also perform in a satisfactory way over the *OEsim* data. Some of them seem to be common across all data sets, with most prominent the PG Cosine Deviation-from-Mean variation.

Conclusions

In this paper we investigate how the recommendation algorithm used in a real life implementation of a multi-criteria recommender system performed under various experimental conditions, by using as input different data sets with multi-criteria ratings. The case study has been a portal for organic and sustainable education, and the experimentation tested a wide number of variations with three real data sets ratings and one synthetic one. The results indicated that some particular variations seem to perform in a satisfactory way over all four data sets. This was an interesting observation, considering that in related work we have witnessed significant alterations in the performance of the same algorithms over different data sets (Manouselis *et al.*, 2007; Manouselis *et al.*, 2010). It gave useful input regarding the improvements that need to be made in the algorithm currently implemented in Organic.Edunet.

The experimental analysis that we carried out clearly indicated that the PG algorithm that is currently implemented in the Organic.Edunet portal, is a good choice. Still, its exact parameterization and fine-tuning so that the right values are chosen that will give better results, is an exercise that needs to be taking place quite often in such a changing environment. As the community of users grows, the properties of the *Users x Items* matrix (that is, of the data set) will be dynamically changing. For instance, during the past year only, more than 1,000 new users have registered in the portal. In addition, the content collections to which the portal gives access to, is ready to expand from about 11,000 items to some 30,000.

This calls for careful consideration and planning from the perspective of the designer and operator of the recommendation service. One option would be to run frequent offline experiments with most recent updates of the data set, in order to find which algorithm variations is more appropriate every time for the application. Another approach would be the investigation of adaptive algorithms that will automatically measure their performance (e.g. the accuracy and coverage of their predictions) and then adapt their parameters in order to achieve better results. Such an approach can be a rather computationally-demanding task that calls for a re-engineering of the existing recommendation service of the portal and maybe an investigation of new multi-criteria recommendation algorithms.

Our future work includes a more extensive experiment where the correlation between the various algorithmic parameters and options and the properties of the data sets will be explored. The currently available real data sets will be used as generators of a large number of synthetic data sets with varying properties. Then the CollaFiS simulator can be used to execute a large number of variations and measure how they perform over the various data sets.

Acknowledgements

The work presented in this paper has been funded with support by the European Commission, and more specifically the FP7 project "agINFRA: A data infrastructure to support agricultural scientific communities" of the Research Infrastructures, Capacities Programme, Objective INFRA 2011 1.2.2: Data infrastructures for e-Science. Giorgos Kyrgiazos was affiliated with (and funded by) Agro-Know Technologies at the time that this study took place.

REFERENCES

Adomavicius G., Kwon Y. (2007), *New Recommendation Techniques for Multi-Criteria Rating Systems*, IEEE Intelligent Systems, 22(3), 48-55, May/June 2007.

Adomavicius G., Manouselis N., Kwon Y. (2011), *Multi-Criteria Recommender Systems*, in Kantor P., Ricci F., Rokach L., Shapira, B. (Eds.), Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners, Springer, 769-803, 2011.

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

Dron J., Mitchell R., Siviter P., Boyne C. (2000), *CoFIND-an experiment in n-dimensional collaborative filtering*, Journal of Network and Computer Applications, 23(2), 131-142, 2000.

Herlocker J., Konstan J., Riedl J. (2002), *An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms*, Information Retrieval, 5, 287–310, 2002.

Herlocker J.L., Konstan J.A., Terveen L.G., Riedl J.T. (2004), *Evaluating Collaborative Filtering Recommender Systems*, ACM Transactions on Information Systems, 22(1), 553, 2004.

Manouselis N., Costopoulou C. (2006), *Designing a Web-based Testing Tool for Multi Criteria Recommender Systems*, Engineering Letters, Special Issue on Web Engineering, 13(3), November 2006.

Manouselis N., Costopoulou C. (2007(, *Experimental Analysis of Design Choices in Multi-Attribute Utility Collaborative Filtering*, International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), 21(2), 311-331, 2007.

Manouselis N., Drachsler H., Verbert K., Duval E. (2013), *Recommender Systems for Learning*, Springer Briefs.

Manouselis N., Drachsler H., Vuorikari R., Hummel H., Koper R. (2011), *Recommender Systems in Technology Enhanced Learning*, in Kantor P, Ricci F, Rokach L, Shapira B (eds), Recommender Systems Handbook, pp. 387-415, Springer US, 2011.

Manouselis N., Kosmopoulos T., Kastrantas K. (2009), *Developing a Recommendation Web Service for a Federation of Learning Repositories*, in Proc. of the International Conference on Intelligent Networking and Collaborative Systems (INCoS), IEEE Press, 2009.

Manouselis N., Kyrgiazos G., Stoitsis G. (2012), *Revisiting the Multi-Criteria Recommender System of a Learning Portal*, in Proceedings of the 2nd Workshop on Recommender Systems in Technology Enhanced Learning 2012 (RecSysTEL'12), in conjunction with the 7th European Conference on Technology Enhanced Learning (EC-TEL 2012), Saarbrücken, Germany, September 18-19, 2012.

Manouselis N., Vuorikari R., Van Assche F. (2010), *Collaborative Recommendation of e-Learning Resources: An Experimental Investigation*, Journal of Computer Assisted Learning, 26(4), 227-242, 2010.

Picault, J., Ribière, M., Bonnefoy, D.,& Mercer, K. (2011), *How to Get the Recommender Out of the Lab?* In Ricci, F., Rokach, L., Shapira, B., Kantor, P. B. (Eds.) Recommender Systems Handbook 333-365.

Recker M.M., Walker A., Lawless K. (2003), *What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education*, Instructional Science, 31(4/5), 299–316, 2003.

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

Sicilia M.-A., Garca-Barriocanal E., Sanchez-Alonso S., Cechinel C. (2010), *Exploring user-based recommender results in large learning object repositories: the case of MERLOT*, in Manouselis N, Drachsler H, Verbert K, Santos OC (eds) Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010), Procedia Computer Science, 1(2):2859-2864, 2010.