Dataset-driven Research for Improving Recommender Systems for Learning

Katrien Verbert
Department of Computer Science, K.U.Leuven
Celestijnenlaan 200A
B-3001 Leuven, Belgium
katrien@cs.kuleuven.be

Hendrik Drachsler
Open University of the Netherlands (OUNL)
P.O. Box 2960, 6401 DL
Heerlen, The Netherlands
hendrik.drachsler@ou.nl

Nikos Manouselis
Agro-Know Technologies, Athens, Greece
and
University of Alcala, Spain
nikosm@ieee.org

Martin Wolpers
Fraunhofer Institute for Applied Information Technology (FIT)
Schloss Birlinghoven, 53754
Sankt Augustin, Germany
martin.wolpers@fit.fraunhofer.de

Riina Vuorikari
European Schoolnet (EUN)
Rue de Trèves, 61
1040 Brussels, Belgium
riina.vuorikari@eun.org

Erik Duval
Department of Computer Science, K.U.Leuven
Celestijnenlaan 200A
B-3001 Leuven, Belgium
erik.duval@cs.kuleuven.be

ABSTRACT

In the world of recommender systems, it is a common practice to use public available datasets from different application environments (e.g. MovieLens, Book-Crossing, or EachMovie) in order to evaluate recommendation algorithms. These datasets are used as benchmarks to develop new recommendation algorithms and to compare them to other algorithms in given settings. In this paper, we explore datasets that capture learner interactions with tools and resources. We use the datasets to evaluate and compare the performance of different recommendation algorithms for learning. We present an experimental comparison of the accuracy of several collaborative filtering algorithms applied to these TEL datasets and elaborate on implicit relevance data, such as downloads and tags, that can be used to improve the performance of recommendation algorithms.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; K.3.m [**computers and education**]: Miscellaneous

Keywords

Recommendation algorithms, Technology Enhanced Learning, datasets, evaluation metrics

1. INTRODUCTION

Recommender systems have been researched and deployed extensively over the last decade in various application areas, including e-commerce and e-health. Several recommendation algorithms, such as content-based filtering [28], collaborative filtering [9] and their hybridizations [3] are widely discussed in the literature and in several surveys of the state-of-the-art. Also in the Technology Enhanced Learning (TEL) domain, the deployment of recommender systems has attracted increased interest during the past years [19]. By identifying suitable learning resources from a potentially overwhelming variety of choices [32], recommender systems offer a promising approach to facilitate both learning and teaching tasks.

Whereas several recommender systems have been implemented for use in learning scenarios in recent years, only a few researchers have attempted to validate their recommendation algorithms based on data that have been captured in a real-life setting [19]. In many cases, small-scale experiments are conducted in which a few learners or teachers are asked to rate the relevancy of suggested resources in a controlled experiment. Whereas such experiments offer valuable insights into the usefulness and relevancy of recommender systems for learning, stronger conclusions about the validity and generalizability of scientific experiments could be drawn if researchers have the possibility of verification, repeatability, and comparisons of results based on large datasets that capture learner interactions in real settings [6]. Such a collection would enable researchers to create repeatable experiments to gain valid and comprehensive knowledge about how certain recommendation algorithms performed on a certain dataset and in certain learning settings.

To collect relevant TEL related datasets, the first dataTEL Challenge[1] was launched as part of the first workshop on Recommender Systems for TEL (RecSysTEL) [18], jointly organized by the 4th ACM Conference on Recommender Systems and the 5th European Conference on Technology Enhanced Learning (EC-TEL 2010) in September 2010. In this call, research groups were invited to submit existing datasets from TEL applications that can be used for research purposes, among others for research on recommender

[1] http://adenu.ia.uned.es/workshops/recsystel2010/datatel.htm

systems for TEL. In this paper, we briefly present the collected datasets and evaluate the performance of several collaborative filtering algorithms on the datasets. The paper has three primary research contributions:

1. First, we present an analysis of datasets that capture learner interactions with tools and resources in TEL settings. These datasets can be used for a wide variety of research on learning analytics.
2. Second, the paper presents an experimental comparison of the accuracy of several collaborative filtering algorithms applied to TEL datasets.
3. Third, we research the extent to which implicit feedback of learners, such as reading information, downloads and tags, can be used to augment explicit relevance evidence in order to improve the performance of recommender systems for TEL.

The paper is organized as follows: Section 2 presents an analysis of datasets that capture learner interactions and that can be used for learning analytics. Section 3 presents an overview of existing recommendation algorithms, and in particular collaborative filtering algorithms, that can be applied to these datasets to suggest relevant resources to learners or teachers. Section 4 presents an overview of evaluation metrics that are commonly used to evaluate recommendation algorithms. Then, we present our evaluation results of the application of these algorithms to TEL datasets. We evaluate algorithms based on both explicit rating data and implicit relevance data, such as tags and downloads, that are available in some datasets. Results and opportunities for future research in this area are discussed in Section 6. Conclusions are drawn in Section 7.

2. DATATEL CHALLENGE

In this section, we present the objectives and results of the first dataTEL challenge that was targeted to collect TEL datasets. These datasets capture user interactions with tools and resources in learning settings and can be used for various purposes in the learning analytics research area. In this paper, we focus on the application of these datasets to validate recommendation algorithms and to tackle challenges to support recommendation for learning.

2.1 Objectives

In the world of recommender systems, it is a common practice to use public available datasets from different application environments (e.g. MovieLens, Book-Crossing, or EachMovie) in order to evaluate recommendation algorithms. These datasets are used as benchmarks to develop new recommendation algorithms and to compare them to other algorithms in given settings [6].

In such datasets, a representation of implicit or explicit feedback from the users regarding the candidate items is stored, in order to allow the recommender system to produce a recommendation. This feedback can be in several forms. For example, in the case of collaborative filtering systems, it can be ratings or votes (i.e. if an item has been viewed or bookmarked). In the case of content-based recommenders, it can be product reviews or simple tags (keywords) that users provide for items. Additional information is also required, such as a unique way to identify who provides this feedback (user identifier) and upon which item (item identifier). The user-rating matrix used in collaborative filtering is a well-known example [9].

Although recommender systems are increasingly applied in TEL, it is still an application area that lacks such publicly available and interoperable datasets. Although there is a lot of research conducted on recommender systems in TEL, they lack datasets that would allow the experimental evaluation of the performance of different recommendation algorithms using comparable, interoperable, and reusable datasets. This leads to awkward experimentation and testing such as using datasets from movies in order to evaluate educational recommendation algorithms. This practice seems to lack the necessary validity for proving recommendation algorithms for TEL [17].

To this end, the dataTEL Theme Team of the STELLAR Network of Excellence[2] launched the first dataTEL Challenge that invited research groups to submit existing datasets from TEL applications that can be used as input for TEL recommender systems. A special dataTEL Cafe event took place during the RecSysTEL 2010 workshop in Barcelona to discuss the submitted datasets and to facilitate dataset sharing in the TEL community.

2.2 Collected Datasets

Seven datasets have been collected as a result of the first dataTEL challenge. In this paper, we use datasets that include usage related data (such as ratings, tags, reads or downloads) as a basis to demonstrate and evaluate recommendation algorithms for learning. We present an overview of datasets that include such usage data, including information on the data elements that are available and basic statistics of the number of resources, users and activities that are stored.

Some of these datasets are already publicly available, whereas others are still under preparation and not yet publicly accessible. An up-to-date overview of datasets is available at the dataTEL website[3]. We expect an increasing amount of learning related datasets in the upcoming year.

2.2.1 Mendeley dataset

The first dataset was submitted by Mendeley [11] and includes usage data of papers that are available through the Mendeley scientific portal[4]. Mendeley is a research platform that helps users to organize research papers and collaborate with colleagues. In the context of learning, such a dataset provides useful data for recommender systems that suggest papers to learners or teachers, or suitable peer learners on the basis of common research or learning interests. Examples of paper recommenders that have been evaluated in TEL settings are InLinx (Intelligent Links) [1], Papyres [25] and pioneering work on the application of recommender systems in TEL conducted by Tang and McCalla [30]. Although

[2] http://www.teleurope.eu/pg/groups/9405/datatel/

[3] http://www.teleurope.eu/pg/pages/view/50630/

[4] http://www.mendeley.com/

research on paper recommenders has been elaborated more extensively in the Research2.0 domain that emerged in recent years, the dataset is currently one of the few available datasets that captures a very large number of user activities. This dataset can be used meaningfully for research on TEL recommender systems in contexts where papers are considered as learning resources. Three files are included in the Mendeley dataset that capture data since 2009:

- *Online article view log.* The online article view set include a random sampling of 200.000 users that are extracted from usage logs. Time at which each view occurred is provided.
- *Library readership.* The library readership set includes 41.220 user libraries that contain more than 20 articles. From the 13.313.548 library entries, 2.655.578 (19.95%) have been read by users.
- *Library stars.* The library stars set provides data on articles that have been starred by users. 186.976 (1.40%) of the 13.313.548 library entries have been starred.

Among others, this dataset is useful for research on (1) extraction of users interests, on the basis of articles that have been starred, read or added to libraries by users, and evolutions in these interests on the basis of time recordings, (2) identification of users who share common interests, on the basis of their usage behavior, and (3) identification of implicit quality/relevance indications of individual articles by analyzing their usage data.

2.2.2 APOSDLE-DS dataset

The APOSDLE-DS dataset originates from the APOSDLE project [16], which ran from March 2006 to February 2010. APOSDLE is an adaptive work-integrated learning system that aims to support learning within everyday work tasks. It recommends resources (documents, videos, links) and colleagues who can help a user with a task.

The dataset captures 1500 user activities of 6 users during an evaluation period of 3 months. The activities captured are *perform task*, *view resource*, *edit annotation*, *perform topic*, *selected learning goal*, *adapting experience level*, *adding resource to collection*, *browse data*, *being contacted*, *contacting person* and *creating new learning path.* The dataset also includes 163 descriptions of documents and document fragments on which these activities were performed.

From the collected data, the *adding resource to collection* action can provide direct information about the relevance of a resource. This action occurred 581 times within the evaluation period. *Creating a new learning path* is considered as an attempt to plan learning activities over a longer time period and can provide a solid basis for research on the recommendation of sequences of resources. Unfortunately, this action occurred only a few times (< 25). Also direct collaboration activities are rare: *being contacted* occurred 11 times and *contacting person* 69 times. Implicit data to cluster users who share similar interests or goals are available more extensively (*149 perform task*, *861 perform topic* and *414 select learning goal* activities). Whereas the current collection contains data of only a few users and may be too small for statistical analysis, the dataset provides a good example of relevant learning activities to be captured in learning settings.

2.2.3 ReMashed dataset

The ReMashed dataset was collected within the ReMashed environment [8] that focuses on community knowledge sharing. The main objective of ReMashed is to offer personalized recommendations from the emerging information space of a community. The ReMashed dataset is based on aggregating contributions of the users in the ReMashed portal. This portal aggregates Web 2.0 contributions from a range of remote services (delicious, Youtube, Flickr, Slideshare, blogs, and twitter) of the users. The data collection started in February 2009 and is still ongoing. It includes information about interests (learning goals), bookmarks, tags, ratings and contents. Until now, 140 users are registered. In total, 23.000 tags and 264 ratings are given to 96.000 items.

The ReMashed dataset includes only publicly available contributions from users. Although, the data is publicly available, the dataset is not prepared yet for public access as it requires anonymization and the commitment of the users.

2.2.4 Organic.Edunet dataset

The Organic.Edunet dataset was collected on a learning portal for organic agriculture educators [20]. The portal provides access to more than 10.500 learning resources from a federation of 11 institutional repositories. The portal mostly focuses on serving school teachers and university tutors and has attracted almost 12.000 unique visitors from more than 120 countries, out of which about 1.000 are registered users. This dataset contains data from the initial operational phase of the portal that took place in the context of the EC-funded Organic.Edunet project.

The dataset was collected from January 2010 until September 2010 and includes information about 345 tags, 250 ratings and 325 textual reviews that these users have provided. The particularity of this dataset is the fact that ratings are collected upon three different dimensions/criteria: the usefulness of a resource as a learning tool, the relevance to the organic thematic, and the quality of its metadata. This allows for the deployment of an elaborate multi-criteria recommendation service within the portal.

2.2.5 MACE dataset

The MACE dataset originates from the MACE project [34], which ran from September 2006 to September 2009. The MACE portal provides advanced graphical metadata-based access to learning resources in architecture that are stored in different repositories all over Europe. Therefore, MACE enables architecture students to search through and find learning resources that are appropriate for their context. From 2007 until now, 1.148 users registered at the portal. The portal offers access to about 150.000 learning resources, from which 12.000 have been accessed by registered users. These objects hold together about 47.000 tags, 12.000 classification terms and 19.000 competency values. Tags were assigned by logged in users and the classification and competency terms by domain experts.

Most user actions with the MACE portal were logged, including *search activities*, using facetted search, social tags, geographical locations, classifications and/or competencies, *access* of learning resources, *download* of resources, *social tagging*, including *add tag*, *add comment* and *add rating*, and *access of user pages*. The time of each user activity is recorded. The dataset provides useful and rich data for various research purposes. In addition to explicit rating feedback, access time, downloads, tags and comments can provide useful implicit indications that can be used to gain knowledge about user interests. The availability of a relatively large set of both explicit and implicit relevance data makes this dataset a potentially useful candidate for recommender research.

2.2.6 Travel well dataset

The Travel well dataset was collected on the Learning Resource Exchange portal [33] that makes open educational resources available from 20 content providers in Europe and elsewhere. Most registered users are primary and secondary teachers who come from a variety of European countries. The dataset contains data from the pilot phase which was conducted during the EC-funded MELT-project. These data were collected from August 2008 until February 2009 on 98 users. The dataset includes explicit interest indicators that can be used to infer the relevance of a resource for the user. Users can rate resources on a scale of 1 to 5 for usefulness and add tags to resources. In total, 16.353 user activities were recorded on 1.923 resources.

The particularity of the dataset is that it contains information of the home country, mother tongue and spoken languages of users. Additionally, it has metadata on the origin of the educational resource and its language. The dataset thus allows tracking the interests of users on 'travel well' resources, indicating that the user and resource come from different countries and that the language of the resource is different from that of the users mother tongue. Additionally, this dataset is useful for research on extraction of teacher interests and identification of teachers who share common interests, on the basis of their tags and ratings. The availability of a relatively large set of such explicit relevance indicators makes this dataset a potentially useful candidate for recommender research in TEL.

2.3 Summary

Table 1 summarizes the details of the collected datasets, including information on the number of users, items and activities that are captured and details on the data elements that are provided. The MACE, Organic.Edunet and Mendeley datasets are the largest datasets that collected user data of 1.148, 1.000 and 200.000 users. The Travel well and ReMashed datasets contain ratings and tags of 98 and 140 users, respectively. The current sample of APOSDLE captures data of relatively few users.

Of interest in this discussion are the data elements that are provided by the datasets. Explicit relevance feedback, such as ratings by users, are provided in the MACE, ReMashed, Organic.Edunet and Travel well datasets. These datasets provide ratings on a five point likert scale and are interesting datasets for evaluating recommender algorithms. Mendeley provides information on articles that are starred by a user (1 if the article has been starred and 0 otherwise), but the semantics of such stars in user libraries may be different for different users (i.e. a star can indicate relevance feedback, but may as well indicate that the user wants to read the article at a later stage). Therefore, the application of such data for recommendation is less straightforward.

In addition to ratings/stars, most datasets include additional user interactions, such as tags, downloads or the inclusion of a resource in a user library. In Section 5.2, we research the extent to which such activities can be used to improve the performance of recommendation algorithms. The APOSDLE dataset includes a wide variety of additional learner related activities, including tasks that are performed by a user, her learning goals and learning paths that she constructed. Whereas the dataset may be too sparse to draw conclusions at this point, the capturing of such activities has a big potential for building recommender systems for learning. The application of this dataset for recommendation for learning is further discussed in Section 6.

The Mendeley, APOSDLE and Travel well datasets are openly available. For the Organic.Edunet, MACE and ReMashed datasets, legal protection rules apply. Details and contact information to obtain the datasets are included in the dataset descriptions. In the remainder of this paper, we report on experimental results with these datasets.

3. RECOMMENDER SYSTEMS

Recommender systems apply data analysis techniques to help users find items that are likely of relevance. Recommender algorithms are often categorized into three areas: collaborative filtering, content-based filtering and hybrid filtering. Collaborative filtering is the most widely implemented and most mature technology [3]. Collaborative recommender systems recognize commonalities between users on the basis of their ratings or implicit relevance indications and generate new recommendations based on inter-user comparisons. Content-based filtering matches content resources to user characteristics [28]. These algorithms base their predictions on individual information and ignore contributions from other users. Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer drawbacks [3].

In this paper, we evaluate the performance of collaborative filtering (CF) on TEL datasets. Similar experiments on TEL settings have been reviewed in Manouselis et al. [19]. The basic idea of CF-based algorithms is to provide recommendations based on the opinions of other like-minded users. The opinions of users can be obtained explicitly from the users or by using implicit measures. Two approaches are distinguished for recommending relevant items to a user:

- *User-based collaborative filtering* computes similarities between users to find the most similar users and predicts a rating based on how similar users rated the item. In a first step, a user-based collaborative filtering algorithm searches users who share similar rating patterns with the active user. In a second step, ratings from these similar users are used to calculate a prediction for the active user.

Table 1: Overview datasets

	Mendeley	APOSDLE	ReMashed	Organic Edunet	Mace	Travel well
Collection period	1 year	3 months	2 years	9 months	3 years	6 months
Number of users	200.000	6	140	1.000	1.148	98
Number of items	1.857.912	163	96.000	10.500	12.000	1.923
Number of activities	4.848.725	1.500	23.264	920	461.982	16.353
Publicly available	+	+	-	-	-	+
reads	+	+	-	-	+	-
tags	-	(+)	+	+	+	+
ratings	(+)	-	+	+	+	+
download or add to collection	+	+	-	-	+	+
search	-	+	-	-	+	-
collaborations	-	+	-	-	-	-
learning goal/task	-	+	+	-	-	-
learning sequence	-	+	-	-	-	-
competencies/ experience level	-	+	-	-	+	-
time	+	-	-	-	+	+

- *Item-based collaborative filtering* applies the same idea, but uses similarity between items instead of users. The approach was popularized by Amazon.com - i.e. users who bought x also bought y. In a first step, an item-item matrix is built that determines relationships between pairs of items. In a second step, this matrix and the data on the active user are used to make a prediction. Once similar items are found, the prediction is then, for instance, computed by taking a weighted average of the target user ratings on similar items.

To enable empirical comparison of different approaches, we implemented different metrics to compute similarities between users and between items and different algorithms for computing predictions, including the standard weighted sum algorithm and simplified Slope One scheme [15]. The different approaches are presented briefly in this section. A more thorough review of various design options for collaborative filtering algorithms can be found in [17]. We report on experimental results in Section 5.

3.1 User-based Collaborative Filtering

User-based collaborative filtering assigns weights to users based on similarities of their ratings with that of the target user [5]. For calculating the similarity between a target user u and another user v, different similarity metrics can be used. We first present commonly used metrics. Then, we present the standard weighted sum algorithm for generating predictions based on these similarity computations.

3.1.1 Cosine similarity

In this case, two users are thought of as two vectors in the m-dimensional item-space. First, the set of items (I_{uv}) that both user u and user v have rated is selected. Then, similarity weights are calculated using the following formula

$$w_{uv} = \frac{\sum_{i\in I_{uv}} r_{vi} r_{ui}}{\sqrt{\sum_{i\in I_{uv}} r_{vi}^2 \sum_i r_{ui}^2}}$$

where r_{ui} is the rating of user u on item i and r_{vi} is the rating of user v on item i. Basically, the cosine similarity between user u and user v is the angle between the ratings vector of user u and the ratings vector of user v.

3.1.2 Pearson correlation.

In this case, similarity between two users u and v is measured by computing the pearson correlation between them using the following formula

$$w_{uv} = \frac{\sum_{i\in I_{uv}} (r_{vi} - r_v)(r_{ui} - r_u)}{\sqrt{\sum_{i\in I_{uv}} (r_{vi} - r_v)^2 \sum_i (r_{ui} - r_u)^2}}$$

where r_v and r_u denote the average ratings for users u and v, respectively. In essence, this similarity measure takes into account how much the ratings of other users for an item deviate from their average rating value.

3.1.3 Tanimoto-Jaccard

The Jaccard or Tanimoto Coefficient [31] measures the overlap degree between two sets by dividing the numbers of items observed by both users (intersection) and the number of different items from both sets of rated items (union). The similarity between two users u and v is defined as:

$$w_{uv} = \frac{|I_u \cap I_v|}{|I_u| + |I_v| - |I_u \cap I_v|}$$

where $|I_u|$ and $|I_v|$ represent the number of items that have been rated by user u and user v, respectively. This similarity metric considers only the number of items that have been rated in common and ignores rating values. The metric can be applied on binary datasets that do not contain rating values. In addition, studies have shown that the metric is advantageous in the case of extremely asymmetric distributed or sparse datasets [23].

3.1.4 Prediction Computation

After computing similarity weights, top-K users with maximum weights are selected as experts. Suppose u is a test user and i is a corresponding test item. Let τ_u be the set of experts who have rated i. The predicted rating $\widehat{r}_{ui}$ is computed as:

$$\widehat{r}_{ui} = r_u + \frac{\sum_{v \epsilon \tau_u} w_{uv}(r_{vi} - r_v)}{\sum_{v \epsilon \tau_u} w_{uv}}$$

Basically, the approach tries to capture how similar users rate the item in comparison to their average ratings. If τ_u is empty, i.e. no expert has rated the test item i, then the average rating of the user is outputted as the prediction.

3.2 Item-based Collaborative Filtering

Item-based collaborative filtering applies the same idea, but uses similarity between items instead of users. Once similar items are found, predictions are computed by taking a weighted average of the target user ratings on these similar items. We briefly describe the similarity computation and the prediction generation. The description is based on [29].

3.2.1 Item similarity computation

The computation of similarities between items proceeds in a similar way than computing similarities between users in user-based CF. The basic idea in similarity computation between two items i and j is to first isolate the users who have rated both items and then to apply a similarity computation technique to determine the similarity w_{ij}. We illustrate the approach using the cosine similarity metric. Alternative similarity measures such as pearson correlation (see previous section) are also commonly applied to calculate similarity between items.

To compute the cosine similarity, we first isolate the co-rated cases (i.e., cases where the users rated both i and j). Let the set of users who both rated i and j be denoted by U, then the cosine similarity is given by

$$w_{ij} = \cos(\vec{i}, \vec{j}) = \frac{\sum_{u \in U} r_{ui} r_{uj}}{\sqrt{\sum_{u \in U} r_{ui}^2} \sqrt{\sum_{u \in U} r_{uj}^2}}$$

where r_{ui} is the rating of user u on item i and r_{uj} is the rating of user u on item j. Thus, this formulation views two items and their ratings as vectors, and defines the similarity between them as the angle between these vectors.

3.2.2 Prediction computation

In the case of item-based predictions, a weighted sum technique computes the prediction of an item i for a user u by computing the sum of the ratings given by the user on items similar to i. Each rating is weighted by the corresponding similarity w_{ij} between items i and j. Formally, we can denote the prediction of item i for user u as

$$\widehat{r}_{ui} = \frac{\sum_{allsimilaritemsj} w_{ij}(r_{ui})}{\sum_{allsimilaritemsj} w_{ij}}$$

Basically, this approach tries to capture how the active user rates the similar items. The weighted sum is scaled by the sum of the similarity weights to make sure the prediction is within the predefined range.

3.2.3 Slope One scheme

The Slope One scheme [15] is an alternative scheme to compute item-based CF predictions that simplifies the implementation of standard item-based collaborative filtering algorithms. The scheme is based on a simple "popularity differential". Let the set of users who both rated i and j be denoted by U. Given a training set c, and any two items j and i with ratings r_{uj} and r_{ui} respectively by some user u in U, then the average deviation of item i with respect to item j is considered as:

$$dev_{j,i} = \sum_{u \epsilon U} \frac{r_{uj} - r_{ui}}{card(U)}$$

The slope one scheme then simplifies the prediction formula to

$$\widehat{r}_{ui} = r_u + \frac{1}{card(R_j)} \sum_{i \epsilon R_j} dev_{j,i}$$

Details are presented in [15]. The advantage is that this implementation of Slope One does not depend on how the user rated individual items, but only on the user average rating and on which items the user has rated. Experimental results are presented in Section 5.

4. EVALUATION METRICS

In this paper, we focus on the measurement of accuracy and coverage of recommendation algorithms, which can be measured by offline analysis of data:

- *Accuracy* measures how well the system generates a list of recommendations. Measures typically used are *precision*, *recall* and *F1*. *Precision* indicates how many recommendations were useful to the user, whereas *recall* measures how many desired items appeared among the recommendations. *F1* is the harmonic mean of precision and recall - that is, $(2{*}precision{*}recall)/(precision+recall)$.
- *Predictive accuracy* evaluates the accuracy of a system by comparing the numerical recommendation scores

against the actual user ratings for the user-item pairs in the test dataset. Mean Absolute Error (MAE) between ratings and predictions is a widely used metric. MAE is a measure of the deviation of recommendations from their true user-specified values. The MAE is computed by first summing absolute errors of the N corresponding ratings-prediction pairs and then computing the average. The lower the MAE, the more accurately the recommendation engine predicts user ratings. Root Mean Squared Error (RMSE) and Correlation are also used as statistical accuracy metric.

- *Coverage* is a measure of the percentage of items and users for which a recommendation system can provide predictions. A prediction is impossible to be computed in case that no or very few people rated an item or in case that the active user has zero correlations with other users.

A more comprehensive review of evaluation metrics for collaborative filtering algorithms can be found in Herlocker et al. [10].

5. EXPERIMENTAL RESULTS

In this section, we present our experimental results of applying collaborative filtering techniques to TEL datasets. We used the Apache Mahout[5] framework for comparing the performance of different collaborative filtering algorithms on datasets. Apache Mahout is an open source framework that provides implementations of standard item-based and user-based collaborative filtering algorithms and implementations of different metrics to compute similarities between users and between items, including pearson, cosine and tanimoto measures.

First, we present results of collaborative filtering algorithms and the influence of different similarity metrics on datasets that contain ratings, including the MACE and Travel well datasets. We also compare these results with accuracy results of algorithms on the MovieLens dataset [5], that is often used by the recommender system community to evaluate algorithms. Then, we present results of collaborative filtering algorithms applied to binary data without ratings, such as data of Mendeley. In this set of experiments, we used implicit relevance indications such as tags and downloads as a basis to generate recommendations.

5.1 Collaborative filtering based on ratings

In a first set of experiments, we applied collaborative filtering algorithms to datasets that contain rating data. First, we compare the influence of different similarity metrics on collaborative filtering. For this first set of experiments, we selected all users from the MACE and the Travel well collection who provided at least 5 ratings. User ratings were randomly split into two sets - observed items (80%) and held-out items (20%). Ratings for the held-out items were to be predicted. We used the Mean Absolute Error (MAE) as the evaluation metric for predictive accuracy in this experiment.

Results are presented in Figure 1. These results indicate that item-based CF based on tanimoto similarity outperforms item-based CF based on pearson and cosine similarity measures for both the MACE and Travel well datasets. In contrast, the use of cosine and pearson measures on the MovieLens dataset improves predictive accuracy of item-based collaborative filtering. These results are consistent with previous experiments that demonstrate that the use of the tanimoto similarity measure on datasets that are very sparse, such as the MACE and Travel well datasets, is beneficial [23].

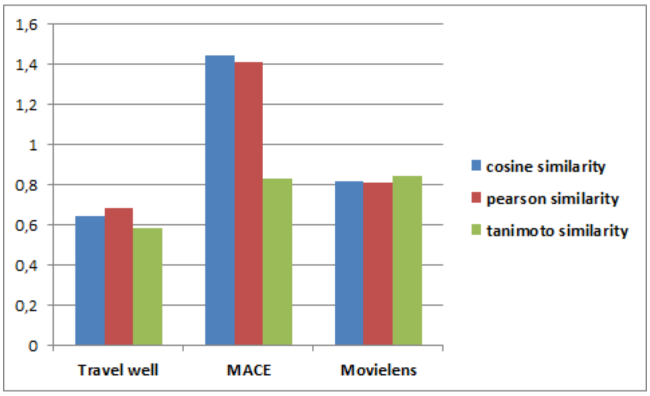

Figure 1: MAE of item-based collaborative filtering based on different similarity metrics

In a second experiment, we compared results of item-based, user-based and slope-one collaborative filtering schemes. For each dataset, we used the best performing similarity measure. Results are presented in Figure 2 and indicate that also the best choice of algorithm is dataset dependent. In the case of MACE, standard item-based collaborative filtering outperforms user-based and slope-one collaborative filtering. For Travel well data, user-based collaborative filtering outperforms the other schemes. The simplified Slope One scheme gives the most accurate results for the MovieLens dataset - which is consistent with findings reported in [14].

Whereas predictive accuracy results of the best performing algorithms on MACE and Travel well data are comparable to reported results of collaborative filtering schemes applied to the MovieLens dataset, the major bottleneck of applying these collaborative filtering schemes to the collected TEL data is the limited coverage of the approach. In MACE, only 113 of 1.148 users provided explicit relevance feedback in the form of ratings. In addition, only 1.706 of 12.000 accessed resources were rated. In the Travel well dataset, more users have provided ratings (56 out of 98), but the number of resources that have been rated by multiple users is very small. In order to address these sparsity issues, we elaborate on the use of implicit relevance indicators and the use of binary data for collaborative filtering in the next section.

5.2 Collaborative filtering based on implicit relevance data

Implicit feedback techniques appear to be attractive candidates to improve recommender performance in the TEL domain, where explicit feedback ratings are often sparse. Behaviors most extensively investigated as sources for implicit

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

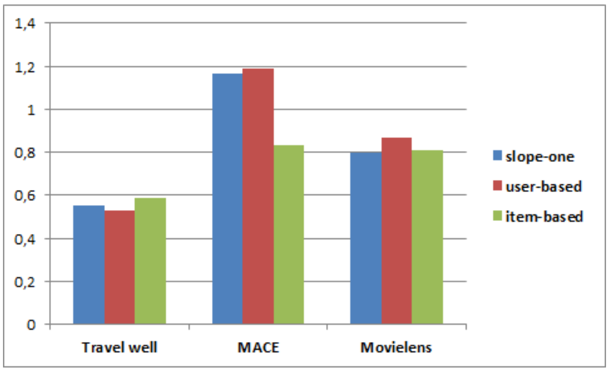

Figure 2: MAE of user-based, item-based and slope-one collaborative filtering

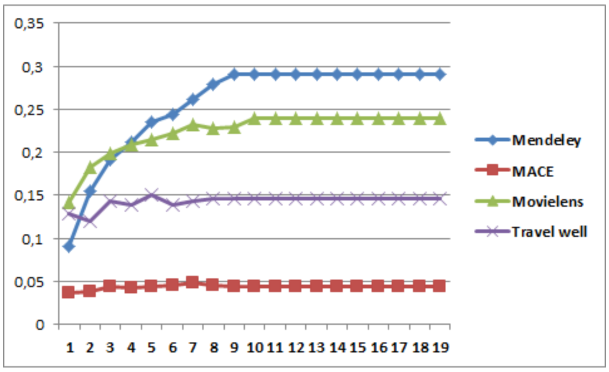

Figure 3: F1 of user-based collaborative filtering with increasing number of neighbors

feedback in other areas have been reading, saving and printing [12]. Morita and Shinoda [24] show that there is a strong tendency for users to spend a greater length of time reading those articles rated as interesting, as opposed to those rated as not interesting. This finding has been replicated by others in similar environments [13]. Other behaviors that have been explored include printing, saving, tagging and bookmarking [27].

We explore the use of implicit relevance data in the Travel well, MACE and Mendeley datasets. In addition to explicit rating data, the Travel well dataset includes 11.943 tags that are provided by 76 users on 1.791 resources. In the MACE dataset, 48.004 tags are provided by 283 users on 6.673 resources. In addition, MACE includes: (1) information about the access of resources (*resultViewed event*), including the date and time when the user viewed the resource, (2) search terms that were used by the user, (3) information about downloaded resources (*save event*) and (4) comments that were added by the user (*addComment event*). The Mendeley dataset provides data about library readership and library stars.

In a second set of experiments, we used these data as implicit relevance indications. In this set of experiments, we predict a fixed number of top-N recommendations and not the ratings. In this case, implicit relevance data are used to rank items to the user in order of decreasing relevance. Suitable evaluation metrics are Precision, Recall and F1. Similar to Sarwar et al. [29], our evaluations consider any item in the recommendation set that matches any item in the test set as a hit. The number of top-N items to be predicted was set to 10. The tanimoto similarity measure was used to compute similarities between users.

Performance results of user-based collaborative filtering on the F1 measure are presented in Figure 3. As can be seen in this figure, the size of the neighborhood affects the quality of the top-10 recommendations. In general, the quality increases as we increase the number of neighbors. However, after a certain point, the improvement gains diminish. Results indicate that implicit relevance indications can be used in a successful way. For Mendeley, we used library readership and starred articles as implicit relevance indications. Based on these data, a standard user-based collaborative filtering algorithm that predicts the top 10 most relevant items for a user has an F1 score of almost 30% - which is comparable to the application of user-based collaborative filtering on the MovieLens dataset (±25%). Reasonable results were also obtained for the Travel well dataset. Similar to the low accuracy results of user-based collaborative filtering on MACE data that were presented in the previous section, accuracy results remain low ($< 5\%$) when additional data about tags and downloads is incorporated. These results are consistent with previous studies of user-based collaborative filtering on extremely sparse datasets. To tackle this issue, part of our ongoing work is based on improving the performance based on alternative similarity measures [26]. We elaborate on useful extensions and future research directions for recommendation for learning in the next section.

6. DISCUSSION

The goal of this kind of dataset driven research on recommender systems is to gain deeper insights into both relevant similarity measures between users and between items and relevant data that can be taken into account to support recommendation for learning. Results of our study show that the tanimoto similarity measure gives most accurate results on the current TEL datasets that are very sparse. The best choice of algorithm (i.e. user-based, item-based or slope-one) is dataset dependant. These results are consistent with previous findings that have been reported in [22]. The results indicate that the successful operation of collaborative filtering in the context of real-life learning applications requires careful testing before their actual deployment.

It is important to note that the presented experiments serve only as a first step towards the understanding and appropriate specialization for recommendation for learning. This study has to be further complemented with experiments that will study the needs and expectations of the users, their information seeking tasks, and how recommended resources may be used in the context of their learning activities [21]. In this study, only very generic collaborative filtering algorithms have been tested. In the learning domain, researchers have proposed the use of additional learner or teacher attributes in recommendation processes [2]. Examples include knowledge or experience levels indicators, learning interests, learning goals, learning and cognitive styles, affects and background information. In addition to interests and preferences that are available in most datasets, the learning goal or competencies of a learner are often incorporated as a basis for generating learning recommendations [4]. Data on competencies or experience levels is available in the MACE and APOSDLE datasets. In addition, APOSDLE provides data on the learning goal of the learner when she is performing a task. Such data is useful to improve similarity measures between users and to find users who share similar goals, both as a basis to improve recommendation of relevant learning resources and to support recommendation of peer learners.

We aim to experimentally test the performance of variation against several attributes of learners or teachers that are proposed in the literature. In order to create evidence driven knowledge about the effect of recommender systems on learners and personalized learning, more experiments like the presented one are needed. The continuation of additional small-scale experiments with a limited amount of learners that rate the relevance of suggested resources only adds little contributions to an evidence driven knowledge base on recommender systems in TEL. The key research question remains how generic algorithms need to be modified in order to support learners or teachers. To give an example, from a pure learning perspective, the most valuable resources for a learner could be the recommendation of different opinions or facts that challenge the learners to disagree, agree and redefine their point of view. In order to enable such experiments, the capturing of learner or teacher data is a key requirement. Our ongoing research is focused on the development of a standardized data model that enables the uniform representation of both explicit and implicit relevance data of learners and teachers [7]. This data model will be standardized in collaboration with the CEN WS-LT Working Group on Social Data[6].

7. CONCLUSION

In this study, we presented datasets that capture learner interactions with tools and resources and that can be used for learning analytics research. We successfully applied several variations of user-based and item-based collaborative filtering algorithms to these datasets. Challenges to be tackled include sparsity of data and require further research on both implicit relevance indicators as well as similarity measures to find relevant items and/or users. To tackle these challenges, the further collection of sufficiently large datasets that capture learner interactions in different real-life learning settings is a key requirement.

8. ACKNOWLEDGMENTS

This work is supported by the dataTEL Theme Team and the STELLAR Network of Excellence. Katrien Verbert is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). The work of Nikos Manouselis has been funded with support by the EU project VOA3R - 250525 of the CIP PSP Programme (http://voa3r.eu). This research used data operated by European Schoolnet (EUN). EUN does not guarantee the accuracy of any data and cannot be held liable for any errors or omissions. The research leading to these results has also received funding from the European Community Seventh Framework Programme (FP7/2007-2013) under grant agreement no 231396 (ROLE).

[6] https://sites.google.com/site/censocialdata/home

9. REFERENCES

[1] C. Bighini, A. Carbonaro, and G. Casadei. Inlinx for document classification, sharing and recommendation. In *IEEE International Conference on Advanced Learning Technologies*, pages 91–95. IEEE Computer Society, 2003.

[2] P. Brusilovsky and E. Millán. User Models for Adaptive Hypermedia and Adaptive Educational Systems. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, chapter 1, pages 3–53. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[3] R. Burke. Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, chapter 12, pages 377–408. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[4] C.-M. Chen, H.-M. Lee, and Y.-H. Chen. Personalized e-learning system using item response theory. *Computers & Education*, 44(3):237–255, April 2005.

[5] M. S. Desarkar, S. Sarkar, and P. Mitra. Aggregating preference graphs for collaborative rating prediction. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 21–28, New York, NY, USA, 2010. ACM.

[6] H. Drachsler, T. Bogers, R. Vuorikari, K. Verbert, E. Duval, N. Manouselis, G. Beham, S. Lindstaedt, H. Stern, M. Friedrich, and M. Wolpers. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2):2849–2858, 2010.

[7] H. Drachsler, H. G. K. Hummel, and R. Koper. Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model. *Int. J. Learn. Technol.*, 3(4):404–423, 2008.

[8] H. Drachsler, D. Pecceu, T. Arts, E. Hutten, P. van Rosmalen, H. Hummel, and R. Koper. Remashed - an usability study of a recommender system for mash-ups for learning. *International Journal of Emerging Technologies in Learning (iJet)*, Special Issue: ICL2009 on MashUps for Learning(5):7–11, January 2010.

[9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22:5–53, January 2004.

[11] K. Jack, J. Hammerton, D. Harvey, J. J. Hoyt, J. Reichelt, and V. Henning. Mendeleys reply to the datatel challenge. *Procedia Computer Science*, 1(2):1–3, 2010.

[12] D. Kelly and N. J. Belkin. Reading time scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 408–409, New York, NY, USA, 2001. ACM.

[13] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[14] D. Lemire, H. Boley, S. McGrath, and M. Ball. Collaborative filtering and inference rules for context-aware learning object recommendation. *International Journal of Interactive Technology and Smart Education*, 2(3), August 2005.

[15] D. Lemire and A. Maclachlan. Slope one predictors for online rating-based collaborative filtering. *Society for Industrial Mathematics*, 05:471–480, 2005.

[16] S. Lindstaedt, B. Kump, G. Beham, V. Pammer, T. Ley, A. Dotan, and R. De Hoog. Providing varying degrees of guidance for work-integrated learning. In *Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL: from innovation to learning and practice*, EC-TEL'10, pages 213–228, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] N. Manouselis and C. Costopoulou. Preliminary study of the expected performance of maut collaborative filtering algorithms. In M. D. Lytras, J. M. Carroll, E. Damiani, R. D. Tennyson, D. Avison, G. Vossen, and P. Ordonez De Pablos, editors, *The Open Knowlege Society. A Computer Science and Information Systems Manifesto*, volume 19 of *Communications in Computer and Information Science*, pages 527–536. Springer Berlin Heidelberg, 2008.

[18] N. Manouselis, H. Drachsler, K. Verbert, and O. Santos, editors. *Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010)*, volume 1 of *Procedia CS*. Elsevier, 2 edition, 2010.

[19] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, and R. Koper. Recommender systems in Technology Enhanced Learning. In R. L. S. B. Kantor P., Ricci F., editor, *Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners*, pages 387–415. Springer, 2010.

[20] N. Manouselis, K. Kastrantas, S. Alonso, J. Caceres, H. Ebner, and M. Palmér. Architecture of the organic.edunet web portal. *International Journal of Web Portals*, 1(1):71–91, 2009.

[21] N. Manouselis, R. Vuorikari, and F. V. Assche. Simulated Analysis of MAUT Collaborative Filtering for Learning Object Recommendation. In *Workshop proceedings of the EC-TEL conference: SIRTEL07 (EC-TEL07)*, pages 27–35, 2007.

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

[23] A. Mild and T. Reutterer. An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. *Journal of Retailing and Consumer Services*, 10(3):123–133, 2003.

[24] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[25] A. Naak, H. Hage, and E. Aïmeur. A multi-criteria collaborative filtering approach for research paper recommendation in papyres. In G. Babin, P. G. Kropf, and M. Weiss, editors, *MCETECH*, volume 26 of *Lecture Notes in Business Information Processing*, pages 25–39. Springer, 2009.

[26] K. Niemann, M. Scheffel, M. Friedrich, U. Kirschenmann, H.-C. Schmitz, and M. Wolpers. Usage-based object similarity. *Journal of Universal Computer Science*, 16(16):2272–2290, 2010.

[27] D. W. Oard and J. Kim. Implicit Feedback for Recommender Systems. In *AAAI Workshop on Recommender Systems, Madison, WI*, pages 81–83, July 1998.

[28] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*, chapter 10, pages 325–341. 2007.

[29] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM.

[30] T. Y. Tang and G. Mccalla. Smart recommendation for an evolving e-learning system. In *Workshop on Technologies for Electronic Documents for Supporting Learning, International Conference on Artificial Intelligence in Education (AIED)*, 2003.

[31] T. T. Tanimoto. An elementary mathematical theory of classification and prediction. *Internal Report IBM Corp*, 1958.

[32] S. Ternier, K. Verbert, G. Parra, B. Vandeputte, J. Klerkx, E. Duval, V. Ordonez, and X. Ochoa. The ariadne infrastructure for managing and storing metadata. *IEEE Internet Computing*, 13(4):18–25, 2009.

[33] R. Vuorikari and R. Koper. Ecology of social search for learning resources. *Campus-Wide Information Systems*, 26(4):272–286, 2009.

[34] M. Wolpers, M. Memmel, and A. Giretti. Metadata in architecture education - first evaluation results of the mace system. *Learning in the synergy of multiple disciplines*, 5794:112–126, 2009.