Which Algorithms Suit Which Learning Environments? A Comparative Study of Recommender Systems in TEL

Simone Kopeinik(✉), Dominik Kowald, and Elisabeth Lex

Knowledge Technologies Institute, Graz University of Technology, Graz, Austria
{simone.kopeinik,elisabeth.lex}@tugraz.at, dkowald@know-center.at

Abstract. In recent years, a number of recommendation algorithms have been proposed to help learners find suitable learning resources online. Next to user-centered evaluations, offline-datasets have been used to investigate new recommendation algorithms or variations of collaborative filtering approaches. However, a more extensive study comparing a variety of recommendation strategies on multiple TEL datasets is missing. In this work, we contribute with a data-driven study of recommendation strategies in TEL to shed light on their suitability for TEL datasets. To that end, we evaluate six state-of-the-art recommendation algorithms for tag and resource recommendations on six empirical datasets: a dataset from European Schoolnets TravelWell, a dataset from the MACE portal, which features access to meta-data-enriched learning resources from the field of architecture, two datasets from the social bookmarking systems BibSonomy and CiteULike, a MOOC dataset from the KDD challenge 2015, and Aposdle, a small-scale workplace learning dataset. We highlight strengths and shortcomings of the discussed recommendation algorithms and their applicability to the TEL datasets. Our results demonstrate that the performance of the algorithms strongly depends on the properties and characteristics of the particular dataset. However, we also find a strong correlation between the average number of users per resource and the algorithm performance. A tag recommender evaluation experiment reveals that a hybrid combination of a cognitive-inspired and a popularity-based approach consistently performs best on all TEL datasets we utilized in our study.

Keywords: Offline study · Tag recommendation · Resource recommendation · Recommender systems · ACT-R · SUSTAIN · Technology enhanced learning · TEL

1 Introduction

Recommender systems have grown to become one of the most popular research fields in personalized e-learning. A tremendous amount of contributions has been presented and investigated over its last fifteen years of existence [1]. However, up to now there are no generally suggested or commonly applied recommender

K. Verbert et al. (Eds.): EC-TEL 2016, LNCS 9891, pp. 124–138, 2016.
DOI: 10.1007/978-3-319-45153-4_10

system implementations for TEL environments. In fact, the majority of holistic educational recommender systems remain within research labs [2]. This may be partly attributed to the fact that proposed recommendation approaches often require either runtime-intensive computations or unavailable, expensive information about learning domains, resources and learner preferences. Furthermore, in informal learning settings, information like ontologies, learning object meta-data and even user ratings are very limited [3]. The spectrum of commonly available tracked learner activities varies greatly, but typically includes implicit usage data like learner-ids, some general information on learning resources, timestamps and indications of a user's interest in learning resources (e.g. opening, downloading or bookmarking) [4]. While existing research investigates the application of implicit usage data-based algorithms (e.g., [5–7]) on selected datasets, a more extensive comparative study directly opposing state-of-the-art recommendation algorithms is still missing. We believe such a study would benefit the community since we hypothesize that recommendation algorithms show different performance results depending on learning context and dataset properties as also suggested in [5,8]. This motivates our main research question: ***RQ1:*** *How accurate do state-of-the-art resource recommendation algorithms, using only implicit usage data, perform on different TEL datasets?*

To this end, we collected six datasets from different TEL domains such as social bookmarking, social learning environments, Massive Open Online Courses (MOOCs) and workplace learning to evaluate accuracy and ranking of six state-of-the-art recommendation algorithms. Results show a strong correlation between the average number of users per resource and the performance of most investigated algorithms. Further, we believe that content-based algorithms that match user characteristics with resource properties, could present an alternative for informal environments with sparse user-resource matrices. However, a prominent factor that hampers the finding and recommending of learning resources is the lack of learning object meta-data, which is resource-intensive to generate. Bateman et al. [9] proposed the application of tagging mechanisms to shift this task to the crowd. Furthermore, tag recommendations can assist and motivate the user in providing such semantic meta-data. Also, tagging supports the learning process, as it is known to foster reflection and deep learning [10]. Yet, so far, tag recommender investigations have been widely unattended in the TEL research community [11]. To this strand, we want to contribute with our second research question: ***RQ2:*** *Which computationally inexpensive state-of-the-art tag recommendation algorithm performs best on TEL datasets?*

The evaluation of three recommendation algorithms, implemented as six variations based on usage data and hybrid combinations, identifies a cognitive-inspired recommendation algorithm combined with a popularity-based approach as most successful.

2 Related Work

In general, there already exists a large body of research on recommender systems in the context of TEL, see e.g., [1,3,12]. Surveys like for example [11] additionally

discuss the potential of collaborative tagging environments and tag recommender systems for TEL. From the wide range of existing contributions, we identify two lines of research that are most related to our work, (i) data driven studies of tag recommendations and (ii) learning resource recommendations in the field of TEL.

2.1 Learning Resource Recommendations

Verbert et al. [5] studied the influence of different similarity measures on user- and item-based collaborative filtering for the prediction of user ratings. Additionally, they compared user-based collaborative filtering on implicit learner data, among four different datasets and first used and analyzed the prominent TEL datasets, TravelWell and MACE. Fazeli et al. [6] showed that the integration of social interaction can improve collaborative filtering approaches in TEL environments. Niemann and Wolpers [7] investigated the usage context of learning objects as a similarity measure to predict and fill in missing user ratings and subsequently improve the database for other recommendation algorithms such as collaborative filtering. The approach is evaluated in a rating prediction setting. The suggested approach does not require any content information of learning objects and thus could also be applied to cold start users, but not cold start items. For further research examples, a broad overview on data-driven learning recommender studies is given in [13]. In contrast to previous work, we do not focus on a specific algorithm or dataset but we study the performance of a range of recommendation algorithms on various TEL datasets.

2.2 Tag Recommendations

Considerable experiments exploring learning resource annotation through tags are presented in [14], in which generally the suitability of tagging within the learning context was investigated. Results claim guidance to be an important factor for the success of tagging. Diaz et al. [15] investigated automated tagging of learning objects utilizing a computationally expensive variant of Latent Dirichlet Allocation [16] and evaluated the tagging predictions in a user study. In [17], an approach to automatically tag learning objects based on their usage context was introduced, which builds on [7]. It shows promising results towards the retrospective enhancement of learning object meta-data. However, their approach cannot be used in online settings as it is based on context information of resources that is extracted from user sessions. In this work, we concentrate on tag recommendation algorithms that are applicable also in online settings.

3 Evaluation

In this work, we evaluate six recommendation algorithms in terms of performance on six TEL datasets from different application areas such as social bookmarking systems (BibSonomy, CiteULike), MOOCs (KDD15), open social learning

(MACE, TravelWell) and workplace learning (Aposdle). We evaluate two recommender application cases, (i) the recommendation of learning resources to support finding relevant information and (ii) the recommendation of tags to support the annotation of learning resources.

3.1 Methodology

For evaluation, we split each dataset into a training and a test set, following a common evaluation protocol used in recommender systems research [18,19]. To predict the future based on the past, each user's activities are sorted in chronological order by the timestamp the activities were traced in the systems. For the tag recommender evaluation, we put the latest post of a user (i.e. all tags assigned by a user to a resource) into the test set and the remaining posts of this user into the training set (see [18]). When evaluating resource recommendations, this process slightly differs. We select 20 % of most recent activities of a user for testing and the remains for training (see [19]). Also, to ensure that there is enough training data available per user, we only consider users with at least five available activities. For the tag recommender test sets, we only consider users with at least two available posts. This procedure avoids a biased evaluation as no data is deleted from the original datasets.

3.2 Algorithms

For the purpose of this study, we selected well-established, computationally inexpensive tag and resource recommendation strategies (for a more substantial review on complexity please see [20]) as well as approaches that have been proposed and discussed in the context of TEL. All algorithms of this study as well as the evaluation methods are implemented in Java as part of our *TagRec* recommender benchmarking framework [21], which is freely available via GitHub[1].

Most Popular (MP). MP is a simple approach to rank items according to their frequency of occurrence [22]. The algorithm can be implemented on user-based, resource-based or group-based occurrences and is labeled respectively, as MP_U, MP_R and MP. $\text{MP}_{U,R}$ describes a linear combination of MP_U and MP_R.

Collaborative Filtering (CF). This approach calculates the neighborhood of users (CF_U) or resources (CF_R) to find items that are new to a user by either considering items that similar users engaged with or items that are similar to resources the target user engaged with in the past [23]. The neighborhood is defined by the k most similar users or resources, calculated by the cosine-similarity measure on the binary user-resource matrix. Tag recommendations require the triple: (user, resource, tag). Therefore, we implemented an adaptation of CF_U for tag recommendations [24]. Accordingly, the neighborhood of a user is determined through a user's tag assignments instead of resource engagements. As suggested by literature [25], we set k to 20 for all CF implementations.

[1] https://github.com/learning-layers/TagRec/.

Content-Based Filtering (CB). CB recommendation algorithms rate the usefulness of items by determining the similarity between an item's content with the target user profile [26]. In this study, we either use topics (if available) or otherwise tags to describe the item content. The similarity between the item vector and the user vector is calculated by the cosine-similarity measure.

Usage Context-Based Similarity (UCbSim). This algorithm was introduced by [27] and further discussed in the TEL context by [7,28]. The approach is inspired by paradigmatic relations known in lexicology, where the usage context of a word is defined by the sequence of words occurring before or after it in the context of a sentence. The equivalent to a sentence in online activities is defined as a user session, which describes the usage context. In line with literature [7], we calculate the significant co-occurrence of two items i and j by the mutual information (MI):

$$MI_{i,j} = log_2 \frac{O}{E} \tag{1}$$

where O is the number of observed co-occurrences and E the number of expected co-occurrences. The similarity ($sim_{i,j}$) between two objects is given by their cosine-similarity, where each object is described as a vector of its 25 highest ranked co-occurrences. For this study, we recommend resources that are most similar to the resources a user engaged with in her last session. Further, we conclude a session if no user interaction is observed for 180 min.

Base Level Learning Equation with Associative Component ($\mathbf{BLL}_{AC}$). This cognitive-inspired tag recommendation algorithm, mimics retrieval from human semantic memory. A detailed description and evaluation can be found in [29]. It is based on equations from the ACT-R architecture [30] that model the availability of elements in a person's declarative memory as activation levels A_i. Equation 2 comprises the base-level activation B_i and an associative component that represents semantic context. To model the semantic context we look at the tags other users have assigned to a given resource, with W_j representing the frequency of appearance of a tag_j and with S_{ji} representing the normalized co-occurrence of tag_i and tag_j, as an estimate of the tags' strength of association.

$$A_i = B_i + \sum_j W_j S_{ji} \tag{2}$$

With $B_i = ln(\sum_{j=1}^{n} t_j^{-d})$, we estimate how useful an item (tag) has been in an individual person's past, with n determining the frequency of tag use in the past, and t_j representing recency, i.e., the time since a tag has been used for the j^{th} time. The parameter d models the power law of forgetting and is in line with [30] set to .5. We select the most relevant tags according to the highest activation values. As BLL_{AC}+MP_R, we denote a linear combination of this approach with MP_R.

SUSTAIN. SUSTAIN [31] is a cognitive model aiming to mimic humans' category learning behavior. In line with [19], which suggested and analyzed the model

to boost collaborative filtering, we implemented the first two layers, which depict an unsupervised clustering mechanism that maps inputs (e.g., resource features) to outputs (e.g., activation values that decide to select or leave a resource).

In the initial training phase, each user's personal attentional tunings and cluster representations are created. The number of clusters per user evolves incrementally through the training process (i.e., a new cluster is only recruited if a new resource cannot be assimilated with the already existing clusters). As input features describing a resource, we select either topics (if available) or tags. The total number of possible input features determines the clusters' dimension. Further, the clustering algorithm has three tunable parameters, which we set in line with [31] as follows: attentional focus $r = 9.998$, learning rate $\eta = .096$ and threshold $\tau = .5$, where the threshold specifies the sensitivity to new cluster creation. The resulting user model is then applied to predict new resources from a candidate set that is given by the 100 highest ranked resources according to CF_U. For the prediction, we calculate and rank an activation value for each resource given by the highest activated cluster in the user model and select the most relevant items accordingly. As SUSTAIN+CF_U, we denote a linear normalized combination of SUSTAIN and CF_U.

3.3 Datasets

Table 1 summarizes the dataset properties such as posts, users, resources, tags, topics and their relations, as descriptive statistics. For the purpose of this study, we use *sparsity* to designate the percentage of resources that are not described by topics or tags. A more elaborate presentation of the datasets follows.

BibSonomy. The university of Kassel provides SQL dumps[2] of the open social bookmarking and resource sharing system *BibSonomy*, in which users can share and tag bookmarks and bibliographic references. Available are four log data files that report users' tag assignments, bookmark data, bibliographic entries and tag to tag relations. Since topics are not allocated [32], we used the tag assignment data, which was retrieved in 2015.

CiteULike. CiteULike is a social bookmarking system for managing and discovering scholarly articles. Since 2007, CiteULike datasets[3] are published on a regular basis. The dataset for this study was retrieved in 2013 (resource recommendation dataset) and 2015 (tag recommendation dataset). Three log data files report on users' posting of articles, bibliographic references, and group membership of users. Activation data of user posts, including tags, have been used for this study. Topics are not available.

[2] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/.
[3] http://www.citeulike.org/faq/data.adp.

KDD15. This dataset origins from the KDD Cup 2015[4], where the challenge was to predict dropouts in Massive Open Online Courses (MOOCs). The MOOC learning platform was founded in 2013 by Tsinghua University and hosts more than 360 Chinese and international courses. Data encompasses course dates and structures (courses are segmented into modules and categories), student enrollments and dropouts and student events. For the purpose of this study, we filtered the event types *problem*, *video* and *access* that indicate a student's learning resource engagement. There are no tags in this dataset but we classify categories as topics.

Table 1. Properties of the six datasets that were used in our study. $|P|$ depicts the number of posts, $|U|$ the number of users, $|R|$ the number of resources, $|T|$ the number of tags, $|Tp|$ the number of topics, $|AT_r|$ the average number of tags a user assigned to one resource, $|ATp_r|$ the average number of topics describing one resource, $|AR_u|$ the average number of resources a user interacted with, $|AU_r|$ the average number of users that interacted with a specific resource. The last two parameters SP_t and SP_{tp} describe the sparsity of tags and topics, respectively.

	$\lvert P\rvert$	$\lvert U\rvert$	$\lvert R\rvert$	$\lvert T\rvert$	$\lvert Tp\rvert$	$\lvert AT_r\rvert$	$\lvert ATp_r\rvert$	$\lvert AR_u\rvert$	$\lvert AU_r\rvert$	SP_t	SP_{tp}
BibSonomy	82539	2437	28000	30889	0	4.1	0	33.8	3	0	100
CiteULike	105333	7182	42320	46060	0	3.5	0	14.7	2.5	0	100
KDD15	262330	15236	5315	0	3160	0	1.8	17.2	49.4	100	1.1
TravelWell	2572	97	1890	4156	153	3.5	1.7	26.5	1.4	3.2	28.7
MACE	23017	627	12360	15249	0[a]	2.4	0	36.7	1.9	31.2	100
Aposdle	449	6	430	0	98	0	1.1	74.8	1	100	0

[a]Generally the dataset contains topics but unfortunately, at this point, we do not have them available.

MACE. In the MACE project an informal learning platform was created that links different repositories from all over Europe to provide access to meta-data-enriched learning resources from the architecture domain. The dataset encompasses user activities like the accessing and tagging of learning resources and additional learning resource descriptions such as topics and competences [33]. At this point, unfortunately, we do not posses access to competence and topic data. However, user's accessing of learning resources and tagging behavior were used in our study.

TravelWell. Originating from the Learning Resource Exchange platform[5], the dataset captures teachers search for and access of open educational resources from a variety of providers all over Europe. Thus, it covers multiple languages and subject domains. Activities in the dataset are supplied in two files with

[4] http://kddcup2015.com/information.html.
[5] http://lreforschools.eun.org.

either bookmarks or ratings which both include additional information about the learning resource [34]. Relevant information to our study encompasses user names, resource names, timestamps, tags and categories.

Aposdle. An adaptive work integrated learning system that origins from the Aposdle EU project. The target user group are workers from the innovation and knowledge management domain. The dataset origins from a workplace evaluation that also included a context-aware resource recommender. Three files with user activities, learning resource descriptions with topics but no tags and a domain ontology were published [35]. The very small dataset has only six users. For the purpose of our evaluation study, we considered the user actions *VIEW_RESOURCE* and *EDIT_ANNOTATION* as indications for learning resource engagements.

3.4 Metrics

For the performance evaluation of the selected recommendation algorithms (MP, CF, CB, UCbSim, BLL, Sustain) we use the further described metrics recall, precision and f-measure, which are commonly used in recommender system research [5,36]. Additionally, we look at nDCG, which was reported to be the most suitable metric for evaluations of item ranking [37].

When calculating recall and precision, we determine the relation of recommended items $\hat{I}_u$ for a user u, to items that are of a user's interest I_u. Items relevant to a user are determined by the test set. All metrics are averaged over the number of considered users in the test set.

Recall. Recall (R) indicates the proportion of the k recommended items that are relevant to a user (i.e., correctly recommended items), to all items relevant to a user.

$$R@k = \frac{|I_u \cap \hat{I}_u|}{|I_u|} \tag{3}$$

Precision. The precision (P) metric indicates the proportion of the k recommended items that are relevant to a user.

$$P@k = \frac{|I_u \cap \hat{I}_u|}{|\hat{I}_u|} \tag{4}$$

F-measure. The F-measure (F) calculates the harmonic mean of recall and precision. This is relevant as recall and precision normally do not develop symmetrically.

$$F@k = 2 \cdot \frac{(P@k \cdot R@k)}{(P@k + R@k)} \tag{5}$$

nDCG. Discounted Cumulative Gain (DCG) is a ranking quality metric that calculates usefulness scores (gains) of items based on their relevance and position in a list of k recommended items and is calculated by

$$DCG@k = \sum_{i=1}^{k} \left(\frac{2^{B(i)} - 1}{log_2(1+i)}\right) \tag{6}$$

where $B(i)$ is 1 if the i^{th} recommended item is relevant and 0 if not. To allow comparability of recommended lists with different item counts, the metric is normalized. nDCG is calculated as DCG divided by the ideal DCG value iDCG, which is the highest possible DCG value that can be achieved if all relevant items are recommended in the correct order, formulated as $nDCG@k = \frac{DCG@k}{iDCG@k}$.

4 Results and Discussion

This section presents our results in terms of prediction accuracy (R, P, F) and ranking (nDCG). Six algorithms with a total of thirteen variations were applied on six TEL datasets from different learning settings. We consider metrics @5 as most relevant, as this seems to be a reasonable number of items to confront a learner with. Additionally, we report F@10 and nDCG@10. To best simulate real-life settings, we conducted the study on the unfiltered datasets.

4.1 Learning Resource Recommendations (*RQ1*)

In line with [5] who compared the performance of CF on different TEL datasets, we observe that the algorithms' performance values strongly depend on the dataset and its characteristics. Solely CF_U shows a stable behavior over all datasets. As expected, the performance of CF_U is related to the average number of resources a user interacted with. The SUSTAIN algorithm, which re-ranks the 100 best rated CF_U values, uses categories of a user's resources to construct learning clusters. Hence, the extent of the resource's descriptive features (we either use topics or tags, if topics are not available) is crucial to the success of the algorithm. Comparing our results of Table 2 with the dataset statistics of Table 1, we find that an average of at least three features per resource is needed to improve the performance of CF_U. Similarly, a poor performance of CF_R is reported for MACE, TravelWell and Aposdle, where the average number of users per resource is lower than two. MP as the simplest approach performs widely poor, except for MACE, where it almost competes with the more complex CF_U. This may relate to the number of learning domains covered by a learning environment. MACE is the only learning environment that is restricted to one subject, namely *architecture*.

The importance of a dense user resource matrix is underlined by our results. In fact, we find a strong correlation of .958 (t = 19.5502, df = 34, p-value < 2.2e−16) between the average number of users per resource ($|AU_r|$) (see Table 1)

Table 2. Results of our resource recommender evaluation. The accuracy estimates are organized per dataset and algorithm (*RQ1*). The datasets BibSonomy, CiteULike and MACE did not include topic information, thus for those three, we calculated CB_T and SUSTAIN on tags instead of topics. *Note:* the highest accuracy values per dataset are highlighted in bold.

Dataset	Metric	MP	CF_R	CB_T	CF_U	UCbSim	SUSTAIN	SUSTAIN + CF_U
BibSonomy	R@5	.0073	.0447	.0300	.0444	.0404	.0396	**.0530**
	P@5	.0154	.0336	.0197	.0410	.0336	.0336	**.0467**
	F@5	.0099	.0383	.0238	.0426	.0367	.0363	**.0496**
	F@10	.0102	.0380	.0226	.0420	.0351	.0374	**.0497**
	nDCG@5	.0088	.0416	.0270	.0440	.0371	.0392	**.0541**
	nDCG@10	.0103	.0490	.0313	.0509	.0440	.0469	**.0629**
CiteULike	R@5	.0051	**.0839**	.0472	.0567	.0716	.0734	.0786
	P@5	.0048	**.0592**	.0353	.0412	.0558	.0503	.0553
	F@5	.0050	**.0694**	.0404	.0477	.0627	.0597	.0650
	F@10	.0042	.0601	.0362	.0488	.0573	.0530	**.0618**
	nDCG@5	.0048	**.0792**	.0427	.0511	.0686	.0704	.0717
	nDCG@10	.0054	**.0901**	.0504	.0635	.0802	.0815	.0863
KDD15	R@5	.0067	**.4774**	.1885	.4325	.4663	.3992	.4289
	P@5	.0018	.2488	.1409	.2355	**.2570**	.2436	.2377
	F@5	.0029	.3074	.1612	.3050	**.3314**	.3025	.3059
	F@10	.0034	.2581	.1244	.2773	**.3195**	.2756	.2769
	nDCG@5	.0053	**.3897**	.1927	.3618	.3529	.3227	.3608
	nDCG@10	.0081	**.4740**	.2090	.4281	.4465	.3939	.4284
TravelWell	R@5	.0035	.0257	.0174	.0404	.0471	**.0483**	.0139
	P@5	.0127	.0212	.0382	**.0425**	.0297	.0382	.0382
	F@5	.0056	.0232	.0240	.0414	.0365	**.0427**	.0204
	F@10	.0078	.0194	.0304	.0456	.0459	**.0481**	.0429
	nDCG@5	.0072	.0220	.0275	.0305	**.0491**	.0446	.0220
	nDCG@10	.0092	.0239	.0353	.0461	**.0631**	.0544	.0405
MACE	R@5	.0253	.0080	.0016	**.0283**	.0151	.0093	.0222
	P@5	.0167	.0079	.0023	**.0251**	.0213	.0065	.0190
	F@5	.0201	.0079	.0019	**.0266**	.0177	.0076	.0205
	F@10	.0169	.0116	.0031	**.0286**	.0189	.0155	.0241
	nDCG@5	.0248	.0082	.0014	**.0264**	.0165	.0079	.0215
	nDCG@10	.0281	.0136	.0026	**.0357**	.0282	.0157	.0302
Aposdle	R@5	.0	.0	.0	**.0026**	.0	.0	.0
	P@5	.0	.0	.0	**.0333**	.0	.0	.0
	F@5	.0	.0	.0	**.0049**	.0	.0	.0
	F@10	**.0196**	.0	.0151	.0045	.0	.0045	.0045
	nDCG@5	.0	.0	.0	**.0042**	.0	.0	.0
	nDCG@10	**.0152**	.0	.0103	.0042	.0	.0036	.0033

and the performance (F@5) of all considered algorithms but MP. This is especially visible when comparing KDD15 ($|AU_r| = 49.4$) and Aposdle ($|AU_r| = 1$). KDD15 is our only MOOC dataset. It differs predominantly through its density but also through the structural nature of the learning environment, where each course is hierarchically organized in modules, categories and learning resources.

Table 3. Results of our tag recommender evaluation. We see that the cognitive-inspired $\text{BLL}_{AC} + \text{MP}_R$ clearly outperforms its competitors (*RQ2*). *Note*: the highest accuracy values per dataset are highlighted in bold.

Dataset	Metric	MP_U	MP_R	$\text{MP}_{U,R}$	CF_U	BLL_{AC}	$\text{BLL}_{AC} + \text{MP}_R$
BibSonomy	R@5	.3486	.0862	.3839	.3530	.3809	**.4071**
	P@5	.1991	.0572	.2221	.2066	.2207	**.2359**
	F@5	.2535	.0688	.2814	.2606	.2795	**.2987**
	F@10	.1879	.0523	.2131	.1875	.2028	**.2237**
	nDCG@5	.3449	.0841	.3741	.3492	.3851	**.4022**
	nDCG@10	.3712	.0918	.4070	.3693	.4095	**.4343**
CiteULike	R@5	.3665	.0631	.3933	.3639	.4114	**.4325**
	P@5	.1687	.0323	.1829	.1698	.1897	**.2003**
	F@5	.2310	.0427	.2497	.2315	.2597	**.2738**
	F@10	.1672	.0294	.1825	.1560	.1797	**.1928**
	nDCG@5	.3414	.0600	.3632	.3457	.4016	**.4140**
	nDCG@10	.3674	.0631	.3926	.3596	.4221	**.4385**
TravelWell	R@5	.2207	.0714	.2442	.1740	.2491	**.2828**
	P@5	.1000	.0366	.1333	.0800	.1300	**.1400**
	F@5	.1376	.0484	.1724	.1096	.1708	**.1872**
	F@10	.1125	.0388	.1356	.0744	.1287	**.1426**
	nDCG@5	.2110	.0717	.2253	.1622	.2525	**.2615**
	nDCG@10	.2411	.0800	.2686	.1730	.2783	**.2900**
MACE	R@5	.1306	.0510	.1463	.1522	.1775	**.1901**
	P@5	.0576	.0173	.0618	.0631	.0812	**.0812**
	F@5	.0799	.0259	.0869	.0893	.1114	**.1138**
	F@10	.0662	.0170	.0692	.0615	.0829	**.0848**
	nDCG@5	.1146	.0463	.1296	.1502	.1670	**.1734**
	nDCG@10	.1333	.0483	.1477	.1568	.1835	**.1902**

Contradicting [13], which suggested to use MOOCs datasets to evaluate TEL recommendations, our findings indicate that recommender performance results calculated on MOOCs are not representative for other, typically sparse, TEL environments. This is especially true for small-scale environments such as Aposdle, where the evaluation positively shows that algorithms based on implicit usage data do not satisfy the use case. For Aposdle, which has only six users, none of the considered algorithms showed acceptable results. While approaches based on individual user data (CB_T, SUSTAIN) may work in similar settings, we suppose this is hindered by the unfortunate association of topics, which do not describe the content of a resource but rather the application type (e.g., template) and the poor allocation of topics to resources which is on average 1.16. We believe that learning environments that serve only a very small number of

users, such as often the case in work place or formal learning settings, should draw on recommendation approaches that build upon a thorough description of learner and learning resources as incorporated in ontology-based recommender systems.

4.2 Tag Recommendations (*RQ2*)

The tag recommender evaluation was limited to the four datasets of our study that feature tags. Contrary to the results of the resource recommender study, we can observe a clear winner, which performs best on all datasets and metrics as depicted in Table 3. $\text{BLL}_{AC} + \text{MP}_R$ combines frequency and recency of a user's tagging history, which is enhanced by context information and therewith also recommends tags that are new to a user. Because runtime and complexity are considered very important factors in most TEL environments [8], we also emphasize the results of $\text{MP}_{U,R}$ that outperforms the comparably cost-intensive CF_U in three of four settings, and hence forms a good alternative for runtime-sensitive settings. An extensive evaluation of runtime and memory for tag recommendation algorithms can be found in [18].

5 Conclusion

This paper presents a data-driven study that measures the performance of six known recommendation algorithms and variations thereof on altogether six TEL datasets from different application domains. Learning settings cover social bookmarking, open social learning, MOOCs and workplace learning. First, we investigate the suitability of three state-of-the-art recommendation algorithms (MP, CF, CB) and two approaches suggested for the educational context (UCbSim, SUSTAIN). The algorithms are implemented on implicit usage data. Our results show that satisfactory performance values can only be reached for KDD15, the MOOCs dataset. This suggests that standard resource recommendation algorithms, originating from the data-rich commercial domain are not well suited to the needs of sparse-data learning environments (*RQ1*). In a second study, we evaluate computationally inexpensive tag recommendation algorithms that may be applied to support learners' tagging behavior. To this end, we computed the performance of MP, CF and a cognitive-inspired algorithm, BLL_{AC}, on four datasets. Results show that a hybrid recommendation approach combining BLL_{AC} and MP_R clearly outperforms the remaining methods (*RQ2*).

Limitations and Future Work. This evaluation only covers performance measurements of resource and tag recommendation algorithms. Other relevant indicators, as described in [13], such as user satisfaction, task support, learning performance and learning motivation are not addressed in this research. Also, we would like to mention the restriction of data-driven studies to items that are part of a user's history (i.e., if a user did not engage with a specific learning resource in the usage data, the evaluation considers this resource as wrongly

recommended). However, this might not be the case. Thus, for future work, we plan to validate our results in an online recommender study. We believe that this would allow us to measure the real user acceptance of the recommendations.

Acknowledgments. We would like to gratefully acknowledge Katja Niemann who provided us with the MACE and TravelWell datasets, as well as the organizers of KDD Cup 2015 and XuetangX for making the KDD dataset available. This work is funded by the Know-Center, the EU-IP Learning Layers (Grant Agreement: 318209) and the EU-IP AFEL (Grant Agreement: 687916). The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria.

References

1. Drachsler, H., Verbert, K., Santos, O.C., Manouselis, N.: Panorama of recommender systems to support learning. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 421–451. Springer, Heidelberg (2015)
2. Khribi, M.K., Jemni, M., Nasraoui, O.: Recommendation systems for personalized technology-enhanced learning. In: Kinshuk, Huang, R. (eds.) Ubiquitous Learning Environments and Technologies, pp. 159–180. Springer, Heidelberg (2015)
3. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R.: Recommender systems in technology enhanced learning. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 387–415. Springer, Heidelberg (2011)
4. Verbert, K., Manouselis, N., Drachsler, H., Duval, E.: Dataset-driven research to support learning and knowledge analytics. Educ. Technol. Soc. **15**(3), 133–148 (2012)
5. Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., Duval, E.: Dataset-driven research for improving recommender systems for learning. In: Proceedings of LAK 2011, pp. 44–53. ACM (2011)
6. Fazeli, S., Loni, B., Drachsler, H., Sloep, P.: Which recommender system can best fit social learning platforms? In: Rensing, C., de Freitas, S., Ley, T., Muñoz-Merino, P.J. (eds.) EC-TEL 2014. LNCS, vol. 8719, pp. 84–97. Springer, Heidelberg (2014)
7. Niemann, K., Wolpers, M.: Usage context-boosted filtering for recommender systems in TEL. In: Hernández-Leo, D., Ley, T., Klamma, R., Harrer, A. (eds.) EC-TEL 2013. LNCS, vol. 8095, pp. 246–259. Springer, Heidelberg (2013)
8. Manouselis, N., Vuorikari, R., Van Assche, F.: Collaborative recommendation of e-learning resources: an experimental investigation. J. Comput. Assist. Learn. **26**(4), 227–242 (2010)
9. Bateman, S., Brooks, C., Mccalla, G., Brusilovsky, P.: Applying collaborative tagging to e-learning. In: Proceedings WWW 2007 (2007)
10. Kuhn, A., McNally, B., Schmoll, S., Cahill, C., Lo, W.-T., Quintana, C., Delen, I.: How students find, evaluate and utilize peer-collected annotated multimedia data in science inquiry with Zydeco. In: Proceedings of SIGCHI 2012, pp. 3061–3070. ACM (2012)
11. Klašnja-Milićević, A., Ivanović, M., Nanopoulos, A.: Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. Artif. Intell. Rev. **44**(4), 571–604 (2015)

12. Manouselis, N., Drachsler, H., Verbert, K., Duval, E.: Recommender Systems for Learning. Springer, New York (2012)
13. Erdt, M., Fernandez, A., Rensing, C.: Evaluating recommender systems for technology enhanced learning: a quantitative survey. IEEE Trans. Learn. Technol. **8**(4), 326–344 (2015)
14. Lohmann, S., Thalmann, S., Harrer, A., Maier, R.: Learner-generated annotation of learning resources-lessons from experiments on tagging. J. Univ. Comput. Sci. **304**, 312 (2007)
15. Diaz-Aviles, E., Fisichella, M., Kawase, R., Nejdl, W., Stewart, A.: Unsupervised auto-tagging for learning object enrichment. In: Kloos, C.D., Gillet, D., Crespo García, R.M., Wild, F., Wolpers, M. (eds.) EC-TEL 2011. LNCS, vol. 6964, pp. 83–96. Springer, Heidelberg (2011)
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
17. Niemann, K.: Automatic tagging of learning objects based on their usage in web portals. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, E. (eds.) Design for Teaching and Learning in a Networked World, vol. 9307, pp. 240–253. Springer, Heidelberg (2015)
18. Kowald, D., Lex, E.: Evaluating tag recommender algorithms in real-world folksonomies: a comparative study. In: Proceedings of RecSys 2015, pp. 265–268. ACM (2015)
19. Seitlinger, P., Kowald, D., Kopeinik, S., Hasani-Mavriqi, I., Ley, T., Lex, E.: Attention please! a hybrid resource recommender mimicking attention-interpretation dynamics. In: Proceedings of International World Wide Web Conferences Steering Committee, WWW 2015, pp. 339–345 (2015)
20. Trattner, C., Kowald, D., Seitlinger, P., Kopeinik, S., Ley, T.: Modeling activation processes in human memory to predict the use of tags in social bookmarking systems. J. Web Sci. **2**(1), 1–16 (2016)
21. Kowald, D., Lacic, E., Trattner, C.: Tagrec: towards a standardized tagrecommender benchmarking framework. In: Proceedings of HT 2014. ACM, New York (2014)
22. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in Folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
23. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 291–324. Springer, Heidelberg (2007)
24. Marinho, L.B., Schmidt-Thieme, L.: Collaborative tag recommendations. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, pp. 533–540. Springer, Heidelberg (2008)
25. Gemmell, J., Schimoler, T., Ramezani, M., Christiansen, L., Mobasher, B.: Improving folkrank with item-based collaborative filtering. In: Recommender Systems & the Social Web (2009)
26. Basilico, J., Hofmann, T.: Unifying collaborative and content-based filtering. In: Proceedings of ICML 2004, p. 9. ACM (2004)
27. Friedrich, M., Niemann, K., Scheffel, M., Schmitz, H.-C., Wolpers, M.: Object recommendation based on usage context. Educ. Technol. Soc. **10**(3), 106–121 (2007)
28. Niemann, K., Wolpers, M.: Creating usage context-based object similarities to boost recommender systems in technology enhanced learning. IEEE Trans. Learn. Technol. **8**(3), 274–285 (2015)

29. Kowald, D., Kopeinik, S., Seitlinger, P., Ley, T., Albert, D., Trattner, C.: Refining frequency-based tag reuse predictions by means of time and semantic context. In: Atzmueller, M., Chin, A., Scholz, C., Trattner, C. (eds.) MUSE/MSM 2013, LNAI 8940. LNCS, vol. 8940, pp. 55–74. Springer, Heidelberg (2015)
30. Anderson, J.R., Schooler, L.J.: Reflections of the environment in memory. Psychol. Sci. **2**(6), 396–408 (1991)
31. Love, B.C., Medin, D.L., Gureckis, T.M.: Sustain: a network model of category learning. Psychol. Rev. **111**(2), 309 (2004)
32. Benchmark folksonomy data from bibsonomy, Knowledge and Data Engineering Group. University of Kassel, 2013/2015. http://www.kde.cs.uni-kassel.de/bibsonomy/dumps
33. Stefaner, M., Dalla Vecchia, E., Condotta, M., Wolpers, M., Specht, M., Apelt, S., Duval, E.: MACE – enriching architectural learning objects for experience multiplication. In: Duval, E., Klamma, R., Wolpers, M. (eds.) EC-TEL 2007. LNCS, vol. 4753, pp. 322–336. Springer, Heidelberg (2007)
34. Vuorikari, R., Massart, D.: Datatel challenge: European schoolnet's travel well dataset. In: Proceedings of RecSysTEL 2010 (2010)
35. Beham, G., Stern, H., Lindstaedt, S.: Aposdle-ds a dataset from the Aposdle work integrated learning system. In: Proceedings of RecSysTEL 2010 (2010)
36. Marinho, L.B., Hotho, A., Jäschke, R., Nanopoulos, A., Rendle, S., Schmidt-Thieme, L., Stumme, G., Symeonidis, P.: Recommender Systems for Social Tagging Systems. Springer, New York (2012)
37. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. Inf. Process. Manage. **43**(2), 531–548 (2007)