Leveraging User-Interactions for Time-Aware Tag Recommendations

Daniel Zoller
University of Würzburg
Data Mining and Information Retrieval (DMIR) Group
zoller@informatik.uni-wuerzburg.de

Stephan Doerfel
Micromata GmbH
Kassel
s.doerfel@micromata.de

Christian Pölitz
University of Würzburg
Data Mining and Information Retrieval (DMIR) Group
poelitz@informatik.uni-wuerzburg.de

Andreas Hotho
University of Würzburg
Data Mining and Information Retrieval (DMIR) Group
L3S Research Center
hotho@informatik.uni-wuerzburg.de

ABSTRACT

For the popular task of tag recommendation, various (complex) approaches have been proposed. Recently however, research has focused on heuristics with low computational effort and particularly, a time-aware heuristic, called BLL, has been shown to compare well to various state-of-the-art methods. Here, we follow up on these results by presenting another time-aware approach leveraging user-interaction data in an easily interpretable, on-the-fly computable approach that can successfully be combined with BLL.

We investigate the influence of time as a parameter in that approach, and we demonstrate the effectiveness of the proposed method using two datasets from the popular public social tagging system BibSonomy.

KEYWORDS

tag recommendation; user behavior; time-aware

1 INTRODUCTION

Tagging allows users to easily organize and share resources online. Users add keywords, called "tags", to resources and store them together, thus enabling themselves and others to retrieve these resources using those tags as cues. Tagging is the basis of social bookmarking systems, such as Delicious,[1] BibSonomy,[2] or Flickr.[3] However, it has also long found its way as a secondary feature into many applications, like web shops, wikis, blogs, or libraries.

The task of recommending tags has proven to be a fruitful line of research. Many approaches (see Section 3) utilize complex or computationally costly models, often relying on the full data collected in the system. However, [10] already showed that simple and fast popularity-based heuristics can achieve comparably good results as more difficult and more expensive state-of-the-art methods.

Recently, [19] confirmed this tendency in a large-scale experiment on multiple datasets, using a time-aware heuristic that transfers the base-level learning (BLL) equation to social tagging. BLL scores a tag based on the time that has passed since it was last used, taking all of a user's previous tag assignments into account. Next to its strong performance, the method has another significant advantage over comparably effective methods: The recommendations can be computed online – no offline training is required. The latter can be a decisive factor for system operators, especially where tagging is only a secondary feature and expensive offline training seems out of proportion due to the amount of required resources (hardware, time, and expertise).

In this paper, we follow the incentive to build such lightweight but effective methods. To that end, we use BLL and augment it with a time-aware heuristic relying on a new source of data, that is usually available in all tagging systems and can be exploited easily and on the fly: the interactions of users with the system. We exploit the context of requested pages in a tagging system to generate tags as recommendation candidates. We demonstrate that for the selection of those requests, time is a critical factor. We evaluate the success of our approach on a dataset of the public real-world tagging system BibSonomy.

Contributions: The main contributions of this investigation are:

(1) We describe a heuristic to harvest tag recommendation candidates from user interactions with the system before posting a resource.
(2) We test two variants of a tag recommender, that utilize the extracted recommendation candidates, one using a fix number of the most recent interactions and one using specific time-windows over which interactions are considered.
(3) We evaluate the overall gain in performance for recommendations of the interaction-based heuristics, as well as a hybridized version in combination with BLL.

Our results demonstrate that the time-aware approach provides better suggestions than the variant using a fixed number of previous interactions. Furthermore, a hybrid combining the time-aware, interaction-based recommender with BLL outperforms plain BLL. Moreover (like plain BLL), our proposed recommender is independent of the tagged resources (making no use of their contents) and respects the requirement of easy, online computability. We expect our approach to be relevant for the tag recommender community, as well as for operators of web systems who want to support their users with tag recommendations without spending effort and resources on costly optimization procedures.

[1] https://del.icio.us (for storing web links)
[2] https://www.bibsonomy.org (for storing web links and publications)
[3] https://www.flickr.com (for storing images)

RecTemp Workshop @ ACM RecSys '17, Como, Italy

The remainder of this paper is structured as follows: First, we introduce our recommender (Section 2) that leverages user interaction in a tagging system and the incorporation of the aspect of time. Then, we discuss related work in Section 3. We describe the dataset and experimental setup in Section 4. Next, we present our evaluation results for time-based and user interaction-based recommenders in Section 5. Finally, we conclude the paper in Section 7, after discussing our findings in Section 6.

2 RECOMMENDERS

Tag recommenders provide recommendations during the process of posting a resource. When a user u is posting a resource r, the recommender will provide a list of tags that u might want to assign to r. For that task, we propose and describe a new time-aware, interaction-driven recommender which exploits the previous interactions of the active user (the one to provide recommendations for) with the tagging system in this section. Lastly, we describe a hybrid approach, combining our approach with BLL (Trattner et al. [19]).

2.1 Time-Aware Interaction-Driven Recommenders

In tagging systems, and more so in systems where tagging is a secondary feature, a user's interactions with the system are much more diverse than just adding and tagging new resources. Still, tag recommendation approaches commonly focus only on the result of these interactions (the tagged resources) to compute candidates to recommend. In this paper, we propose going beyond only the tagging activities and consider the context of all interactions that a user has with the system. To generate meaningful recommendations for a resource r, we focus especially on the most recent interactions, as they are likely to belong to the same context as r. The rationale behind the approach is, that by browsing and searching resources, users reveal their current interests, which are useful to select fitting tags. A user's interactions occur in the requests she sends to the system. Therefore, we call our recommender approach *Last Requests* (LR). To select requests from which candidate tags are computed we distinguish two options: First, a time-window based variant $\mathrm{LR_t}$: When user u stores a new resource r, recommendations are computed using all previous requests that u made during the time frame of length d immediately preceding the current posting of r. In the second variant, called $\mathrm{LR_n}$, only the last n interactions of u are considered. While the first variant is explicitly time-aware and thus more likely to capture interactions relevant in the context of the active post, the second variant is more broadly applicable, as users do not necessarily interact with the system prior to posting a new resource (e.g., when using a posting tool, like a bookmarklet or browser add-on). In such a case, older but perhaps still relevant information could be used.

In both variants, recommendation candidates are derived from each of the considered interactions. To yield a ranked list of recommendations, the resulting tags are ordered by their frequency among the interactions.[4] We describe methods for deriving recommendable tags from interactions in Section 4.3. Both recommenders can compute the candidates more efficiently than other recommenders that leverage all previously used tags of a user, because they only compute frequencies on a set of tags which is (usually much) smaller than the complete list of a user's previously used tags, let alone the complete historical posting data in the system.

[4] In the case of ties, we return the tags in lexicographic order.

2.2 Hybrid with BLL

Both proposed variants of the tag recommender cannot provide results for new users that post resources immediately after registration without interacting with the system. Also, it is not unusual that users do not interact with the system before they store a new resource. This fact also impedes the time-aware variant $\mathrm{LR_t}$. To compensate, we combine each of the interaction-based methods with BLL into a hybrid recommender. BLL has been shown to outperform other (time-dependent) approaches (see Section 3). It uses all tags previously used by the active user and computes a recency-based ranking on them: Let T_u be the set of tags previously used by user u and $timep(p)$ the timestamp when user u stored post p. Further, let $Y_{t,u}$ be the set of tag assignments for tag t of the user u (i.e., we add a tuple (u, r, t) every time a user u annotates a resource r with tag t to the corresponding set), and $time(y)$ be the timestamp of the tag assignment $y \in Y_{t,u}$. The BLL-score of each tag in T_u is calculated as $\ln(\sum_{y \in Y_{t,u}} (timep(p) - time(y))^{-d})$ and normalized by the softmax function over all scores. We set $d = 0.5$ for our evaluation – the setting which obtained the best results in [19].

Our interaction-based recommender approaches are combined with BLL in the following way: First, the interaction-based method is used to compute a ranked list L_1 of candidates (possibly of length 0). Similarly, BLL also yields a ranked list L_2 of candidate tags. From L_2, we remove all tags that also occur in L_1. Then the two lists are concatenated, such that the final ranked list of recommendations contains the suggestions from L_1 followed by those from L_2. We denote the two resulting hybrids by $\mathrm{LR_t}$ + BLL and $\mathrm{LR_n}$ + BLL.

3 RELATED WORK

Recommending tags can serve various purposes, such as increasing the chance of getting a resource annotated, reminding a user what a resource is about, and consolidating the vocabulary across users. Furthermore, as Sood et al. [18] pointed out, tag recommenders lower the effort of annotation by changing the process from a *generation* to a *recognition* task: rather than "inventing" tags, the user only needs to select some of the recommended tags.

Since the emergence of social bookmarking, the topic of tag recommendations has raised considerable interest among researchers. As for all recommender domains, tag recommender algorithms can roughly be classified into three classes: *Content-based* algorithms use the content of resources, for instance to compute similarities between items and to present items that are similar to the ones the active user previously liked. *Collaborative* algorithms make use of the relations between the users and the items, for instance by identifying similar users and suggesting items similar users liked. The third class are algorithms that exploit both data sources, sometimes called *hybrid* recommenders. A major drawback of content-based approaches is that the usability of a resource's content depends on the type of resource. For example, when the tagged resources are textual, using words from those resources (e.g., from the title, like Lipczak et al. [11] did) can be successful. However, the same is much harder when the resources are images. Moreover, even for textual resources, the implementation of the recommendation (e.g., the word selection strategy) is resource dependent. Thus, since

our goal is to create a simple and highly versatile recommender, we focus only on the second class of recommender algorithms – those that exploit the folksonomy structure between users, tags, and resources, which exists in any tagging system.

An evaluation of collaborative algorithms, such as collaborative filtering, the FolkRank algorithm [9], and simpler, popularity-based methods was performed in [10] on various datasets. FolkRank outperformed the other methods. However, the hybrid heuristic based recommender, that combined users' frequently used tags with tags that were frequently used to annotate the resource, was second. Rendle and Schmidt-Thieme [16] produced recommendations with a statistical method based on factor models. They factorized the folksonomy structure to find latent interactions between users, resources and tags. Using a variant of the stochastic gradient descent algorithm, the authors optimized an adaptation of the Bayesian Personal Ranking criterion [15]. Seitlinger et al. [17] proposed an approach that simulates human category learning in a three-layer connectionist network. In the input layer, Latent Dirichlet Allocation is used to characterize the resource (and user). [14] introduced a slightly different folksonomy graph model in which edges are weighted and directed. On the resulting graph, PageRank is used to produce a ranking of tags. [12] proposed 'TagRank', a variant of topic-sensitive PageRank upon a tag-tag correlation graph which they integrate into a hybrid with collaborative filtering and popularity-based algorithms. The selection of the algorithms for the hybrid is guided by a greedy algorithm. A drawback of all the presented algorithms is their reliance on complex methodology that uses the full corpus of folksonomy data to learn a recommendation model. While they are suitable approaches to boost performance, they also require a lot of effort in terms of additional computation time, hardware, implementation (e.g., additional data structures, methods to update the trained models), and expertise. Due to the fact that a folksonomy changes over time, the learned models must be updated regularly to fit the current data.

In [21], the authors introduce GIRP, a temporal tag usage pattern model. It uses an exponential function that considers the first- and last-time usage of a tag. A short-term interests model is proposed in [20], recommending the most popular tags of users based on recent data, that is, data from a time-window of fixed length (one day or higher). It is found that a window of 30 days works best on the overall BibSonomy dataset. Recently, Trattner et al. [19] presented a comprehensive study of various tag recommender strategies, including their own development based on a model of human memory (BLL). In contrast to GIRP [21], BLL models the temporal tag usage using a power function rather than an exponential function (see Section 2.2). They compare BLL with other methods (including several of those mentioned above) and find that BLL performs better than the time-dependent algorithm GIRP and other methods based on matrix factorization. Only more computationally expensive models achieve a higher F-score on the evaluated datasets. Most of these models extend basic models by re-ranking the tag candidates by the semantic context of the resource.

In this work, we assume the perspective of a tagging system operator or, respectively, the operator of a system that includes tagging as a secondary feature. We aim at supporting the tagging process with as little cost as possible while still delivering good results. Following the strong results of BLL in [19], and given its low computational effort and the convenient fact that it requires no extra data structures nor precomputed values (see Section 2.2), we use this approach as our baseline. Our method is similar to that of Yin et al. [20], however, instead of using tags of the previous posts, we use tags extracted from previous user interactions with the tagging system. We will show that the time-frame for collecting such interactions is critical and that time-frames of less than a day are worth considering.

4 EXPERIMENTAL SETUP

In this section, we introduce the tagging system BibSonomy, of which we use data for our study. We further describe the datasets with all preprocessing steps, the experiments and their evaluation.

4.1 BibSonomy

BibSonomy [1] allows users to collect references to (scientific) publications and bookmarks to websites, and to annotate these with arbitrary keywords, so called tags. While entering the metadata of a resource in a form, the user can also enter tags, which she can later use for retrieval. The system assists the user with a tag autocompletion relying on her previously used tags. Next to the possibility of filtering resources by tag and/or user, users can find new interesting resources through a full text search. Additionally, users can form groups in which they can share posts and literature. On group pages, all group members' resources are displayed. Overview pages for websites and publications enable users to see who else bookmarked a specific resource and the tags they used to describe the resource. Detail pages for publications show the metadata that the user who saved the resource had entered for the publication. While browsing the system, a user can copy resources of other users into her own collection. Another feature allows users to group tags to concepts (e.g., the tags "time" and "tag" to the concept "recsys"). These concepts can be used for retrieving resources. BibSonomy is a popular target for spammers, that is, for users who store links to advertisements to promote their visibility. For that reason, users are classified by a learning algorithm and manually by the system's administrators. For our analysis, we only used data generated by users that were not marked as spammers.

4.2 Datasets

Our experiments rely on two types of data gathered from the real-world tagging system BibSonomy: posts and user interactions. The latter type of data is rarely published – due to privacy concerns. However, BibSonomy makes such data available to researchers in the form of collected HTTP-request server logs.[5] Thus, at the moment, BibSonomy is the only source enabling the analyses presented here. The methodology, however, is transferable to other tagging systems.

Request Log Data: The request log data contains every web request any user made to the tagging system. We removed all non-human requests like redirects to other pages, or requests by bots or other applications (using the user agent information). Also, we only considered requests to HTML sites and excluded system pages (e.g., the login page). We used two different time frames for the evaluation: (i) from 2006-01-01 (the start date of BibSonomy) through 2011-12-31 (a dataset already used in behavioral analyses in previous work [5]) and (ii) a more recent share, ranging from 2014-07-01

[5] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

Table 1: Statistics about the content datasets, where $|U|$ is the number of users, $|P|$ the number of posts, $|R|$ the number of resources, $|T|$ the number of tags and $|Y|$ the number of tag assignments.

Dataset	$\|U\|$	$\|P\|$	$\|R\|$	$\|T\|$	$\|Y\|$
Bib_1	3,674	180,807	162,364	74,634	672,249
Bib_2	1,912	53,098	46,441	33,937	207,709

through 2016-06-30. For the remainder of this paper, we refer to the older dataset as Bib_1 and to the newer dataset as Bib_2. The older dataset contains 2,074,182 and the newer 246,472 requests.

Content Data: The second type of data is the data generated by the users of the system by annotating resources with tags. We split the content dataset of BibSonomy into the same two time frames as the interaction data. We applied several cleaning steps to each dataset. First, we removed system tags (like myown or imported). To filter imports, we deleted all posts of a user that shared a posting date with another post.[6] The remaining tags were normalized by decapitalizing and removing all non-alphanumeric characters from the tag string. We did not prune our datasets using a p-core, to avoid biases to the results (see Section 4.4). Furthermore, the goal of this research is to produce broadly applicable recommender strategies, but restricting the dataset to only its dense part would neglect new users, rare tags and rare resources, thus providing an incomplete impression on the overall performance. The statistics for the two cleaned content datasets can be found in Table 1.

4.3 Extraction of Tag Recommendation Candidates from User Interaction

While browsing in a tagging system, the user queries different page types. For extracting tag recommendation candidates from requests to BibSonomy, we are using the following methods for the different page types. After the extraction, we also normalized the extracted tags as described in the previous section.

Tag Pages: In BibSonomy, users can restrict the global collection of resources or the collections of other users, groups and search results by tags on a separate tag page for the corresponding entity. All pages also support to filter with more than one tag. We represented a request to one of the tag pages with the specified tag(s).

User Pages: For user pages, we extracted the user's tags that she used before the request was made for her own posts. We only considered user pages where the logged-in user requested a user page of another user. This is the same representation that [13] used for their analysis.

Resource Pages: Also, like [13], we represent a publication or website overview page with the tags that any user used to describe the requested resource. We restrict the tags extracted from a details page to the page owner's tags.

Concept Pages: Users can request concept pages for users, groups, or globally, showing resources tagged with the keywords the user, group or all users defined as subtags of the concept. We added all subtags of the concept to the set of considered tag candidates.

Search Pages: We represent a search request to BibSonomy with the terms of the search query after removing stop words using a multi-language stop word list.

[6]Furthermore, we removed user accounts that are used by libraries, like DBLP, from the two BibSonomy datasets.

External Referer: When users use external engines for searching and click on results linking to BibSonomy, the external source is logged as referer in the request logs. For candidate extraction, we tokenize the value of the 'q' url parameter (because it is the most commonly used parameter by search engines) of these requests and remove stop words.

Tags of a copied post: While browsing by tags or searching, the system presents the user with the resources that match her entered query. Next to every resource, the user can click on a copy button. This click is recorded in the request logs. We extract the tags of the copied resource to represent this type of request.

While some of the request types (e.g., concepts) are specific to BibSonomy, most of them are usually present in a tagging system, representing search options as well as the typical navigation paradigm in a folksonomy.

4.4 Evaluation

For tuning the parameters t and n of the two recommender heuristics, LR_t and LR_n, we split each of the two datasets into a validation and a test set: each part contains 50 % of the posts. We use the validation set to determine the best parameters t and n. Then we use the test to evaluate a hybrid with BLL (cf. Section 2.2).

The choice of the evaluation setup often has a strong influence on the experiments. For example, Cremonesi et al. [3] observed that different sampling strategies yield different outcomes, while Doerfel et al. [4] showed that different restrictions on the dataset can also lead to different (contradictory) results. Both suggest that the scenario should be selected such that it resembles reality as much as possible and that choices should be based on the use case rather than on issues like sparsity. Therefore, we adapt the rating-based *temporal leave-one-out* method introduced in [2] to our scenario. In our experiments, we consider each post as a test post. Moreover, we ensure that the algorithms only use data from before the creation of the test posts. More specifically, for each post p, we do the following: We select all posts (and requests) that have been created before p and use it to compute recommendations for p based on the user and resource of p. Then we compare the recommended tags to the actual tags of p and evaluate the number of correctly predicted tags. This scenario is the most realistic offline evaluation scenario as it considers each occasion for recommendations and uses only data resembling exactly the state of the system at the time the test post was actually created.

Metric: We use the standard information performance metric F-score for measuring the quality of the recommendations [7]. Online systems usually present users with only a limited number of tag recommendations while saving a new resource (often five tags). For that reason, we report F@5, that is, the F-score computed for the set of the first five suggested tags from the ranked list of recommendations (cf. Section 2.2). The parameters t and n are selected based on the best F@5 score found in experiments on the validation sets.

Significance: To test the obtained results for statistically significant differences, we use the Wilcoxon signed-rank test. Since we consider a large number of posts for the evaluation, we compute two versions – considering either the posts or the users as the population. For the latter case, we averaged the obtained F@5 scores over all posts per user in the test set, and we conduct the significance test based on these averages. To indicate statistical significance, we use the symbol * after a reported F-score when the user-based test

indicates significance, and we use $^+$ for the post-based test, in both cases using the α-level of 0.01.

5 RESULTS

In this section, we present the results of our experiments. In the first set of experiments, we use the validation datasets to tune our two approaches LR_t and LR_n, using various settings for t and n, thus including different sets of user interactions. In all these experiments, we consider only those posts where the situation was suitable for the respective approach, that is, where interactions with the system had been recorded. Thus, the number of considered posts for which an approach is tested varies from recommender type and setting. On the same subsets of posts, we also evaluate BLL and the hybrid with BLL for comparison. Eventually, in Section 5.2, we use the test datasets to evaluate the impact of our approach on the overall performance, that is, we evaluate the recommender strategies using all posts without any restriction.

5.1 Time-Aware Interaction-Driven Recommenders

First, we report the results for the recommenders LR_t and LR_n (see Section 2.1). We report results averaged over all posts where the respective heuristic was applicable (i.e., where there were observable interactions), and we compare to BLL on the same set of posts.

In our first evaluation, we vary the parameter n – the number of included previous interactions – of the LR_n recommender from 1 to 10 on the validation set and report the results of the best parameter in Table 2a. We find that on Bib_1 and Bib_2 similar configurations ($n = 6$ for Bib_1 and $n = 7$ for Bib_2) worked best for LR_n. Further, we can observe that LR_n alone is clearly inferior to BLL in terms of F@5. Combining results by concatenating the lists of recommended tags (LR_n + BLL), as described in Section 2.2, can improve the F@5 score, but still cannot reach that of plain BLL. Our hypothesis is that the last requests are too far in the past and thus have no relevance for the current post. Thus, the time of the considered interactions is a critical factor.

Therefore, in Table 2b, we switch the mode of selecting requests to time-windows, using LR_t. We vary the considered time-window for including requests from one minute to 30 days.[7] Although the number of posts for which the heuristic is applicable grows with the selected time-window length, we observe decreasing scores for LR_t on both datasets – more evidence for the above hypothesis. We find that using a time-window t of one minute yields the highest performance on (the respective validation sets of) both datasets, Bib_1 and Bib_2. Other than before with LR_n, the time-aware interaction-based heuristic LR_t yields F@5 scores comparable to BLL on both Bib_1 and Bib_2. When combining the two recommenders ($LR_{t=T}$ + BLL), the scores improve significantly, by ten percentage points over plain BLL on both sets. The improvements obtained for Bib_1 are significant according to the Wilcoxon signed-rank test, considering both the users and the posts as entities. On Bib_2, the significance of the hybrid's improvement is confirmed when the posts are used as entities in the test.

Figure 1 shows the trend of the F@5 scores calculated on BLL compared to the hybrid LR_t + BLL for the time-windows ranging from one minute to five minutes. We can observe that BLL remains roughly constant when t increases. On the other hand, the combination of LR_t and BLL decreases from 0.42 to 0.35 when we increase the time-window t from one minute to five minutes, but outperforms BLL for every considered time-window. For the dataset Bib_2, we find similar results, except that LR_t decreases faster (figure omitted due to space limitations).

[7] 30 days is the setting for which Yin et al. [20] report the best results on the overall dataset for their approach based in previous posts.

Table 2: F-scores for LR_n and LR_t with their corresponding best configurations N and T. For comparison both BLL and the hybrid recommender are reported for the subset. We also report the number of considered users $|U|$ and posts $|P|$. Symbols * and + indicate a statistically significant difference (see Section 4.4 for details).

(a) Results for LR_n. We report the significant difference of BLL compared to both other methods.

	Bib_1 (N=6)			Bib_2 (N=7)		
	$\frac{\lvert P\rvert}{\lvert U\rvert}$	F@5		$\frac{\lvert P\rvert}{\lvert U\rvert}$	F@5	
BLL	62,541 / 1,998	**0.27**	$^{*+/*+}$	17,788 / 509	**0.32**	$^{*+/*+}$
$LR_{n=N}$		0.13			0.10	
$LR_{n=N}$ + BLL		0.16			0.17	

(b) Results for LR_t. We report the significant difference of LR_t and LR_t + BLL compared to BLL.

	Bib_1 (T=1m)			Bib_2 (T=1m)		
	$\frac{\lvert P\rvert}{\lvert U\rvert}$	F@5		$\frac{\lvert P\rvert}{\lvert U\rvert}$	F@5	
BLL	6,974 / 927	0.31		910 / 163	0.35	
$LR_{t=T}$		0.32	$^{*+}$		0.33	
$LR_{t=T}$ + BLL		**0.42**	$^{*+}$		**0.45**	$^{+}$

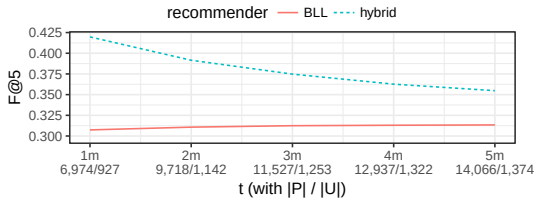

Figure 1: F-scores for BLL and the hybrid LR_t + BLL on the test set of Bib_1. The time-window t ranges from one to five minutes. For each t the number of posts and users in the considered subset are reported.

5.2 Overall Performance

In the previous section, we saw that the time-window based LR_t recommender achieved better results than BLL on those subsets of the data where the respective heuristic was applicable. In this section, we evaluate the hybrid of LR_t with BLL on the complete test sets to get an impression of its overall impact as an improvement over plain BLL. In the following, we combine LR_t with those parameters that produced the best results on the validation sets. Results are given in Table 3. The combination with the request-based LR_t recommender improves the recommendation result by about three per cent on the Bib_1 test set and one per cent on the Bib_2 test set. A Wilcoxon signed-rank test, conducted on the average F-score of each user, indicates that the difference is significant for Bib_1.

Table 3: F@5-scores of hybrid recommender and BLL on the full BibSonomy test sets. Symbols * and + indicate a statistically significant difference (see Section 4.4 for details).

	Bib_1	Bib_2
BLL	0.265	0.315
$LR_{t=1m}$ + BLL	**0.274**$^{*+}$	**0.318**$^{+}$

Testing with all posts as entities, significance is confirmed for both Bib_1 and Bib_2. Thus, we can conclude that exploiting user interactions in very short time-windows immediately before posting a resource can boost the performance of the already well-performing recommender algorithm BLL.

6 DISCUSSION AND LIMITATIONS

In our studies, we found that the time-window based recommender LR_t provides better recommendations than the recommender LR_n, which uses the last requests. We also saw that in those situations where it is applicable its results are comparable to the more complex algorithm BLL. Combining the time-window variant with BLL into a hybrid significantly improves the performance. In the following we discuss several aspects and limitations of our study.

Timing: Overall, we could demonstrate the applicability of our time-aware interaction-driven heuristic. We saw particularly that short time spans immediately before the posting of a resource yield good recommendations. It seems that users store posts in bursts, each representing different aspects of their (shifting) interests. Thus, relying on very recent tags is a reasonable approach.

Applicability: The number of posts and users for which the time-window based recommender LR_t can provide tag recommendations is only a relatively small subset of the data. One reason for this phenomenon may be the fact that BibSonomy offers browser extensions for posting resources, thus a useful means of posting to BibSonomy without visiting the system first.

Generalizability of the Results: Since we could evaluate our recommendation approach only on two BibSonomy datasets (due to the unavailability of suitable data from other systems), it remains an open research question to see how it would perform in other tagging systems. Heckner et al. [8] found that users of tagging systems with different resources tend to tag for different reasons, for example, Flickr (images) is used mostly for sharing, Delicious (websites) mainly for retrieving resources from one's own collection. Since the resource types in BibSonomy and Delicious are similar (references to documents; websites in both systems, publications only in BibSonomy), we hypothesize that they are often used for similar purposes. Also, most page types (e.g., a user page) exist in both systems. Thus, we would expect results of recommender algorithm experiments conducted on Delicious to be qualitatively similar to those found for BibSonomy. Another influence on the usage of tagging systems is the user interface which varies from system to system. For example, BibSonomy always links the entities of the folksonomy, while other systems may exclude some links or place links on different positions within a page.

Transferability of the Approach: While the above mentioned issues are limitations to the generalizability of the results in our study, our methods of extracting tag recommendations from interactions can easily be adapted to other tagging systems.

Computational Cost: In contrast to many other methods (cf. Section 3), our heuristic does not require large user profiles, as it draws the recommendations only from the interactions in the tagging system directly preceding a new post. Moreover, the heuristic is easy to implement into arbitrary tagging systems, and requires only data collected from browsing activities. While, in our experiments, we made use of the request logs to exploit interactions, in a production environment tags can be derived directly from the interactions and can be stored into a temporary cache. Thus, the recommendations can be computed directly without accessing additional data sources. Relying solely on counting occurrences and restrictions on small subsets of interactions, recommendations can be computed online without previous training. Consequently, they require only little effort, making them ideal candidates for systems where tagging is included as a secondary feature and for quickly prototyping a new tagging system.

Explainability: Finally, it is worth pointing out that, in our approach, the choice of the recommended tags can be easily explained. Explanations are suitable for increasing users' acceptance of the recommendations, particularly as the explanations reveals that no profiling of the user is necessary as only the few recent activities are exploited.

7 CONCLUSION

In this work, we have proposed a time-aware, interaction-driven tag recommendation heuristic, for the first time leveraging user interaction in a tagging system beyond the publicly visible posting. We have evaluated our approaches on data of the real-world tagging system BibSonomy. We have shown that time is a critical factor and that particularly interactions immediately before a new post are a good source for recommendable tags.

Finally, combining the time-aware approach with BLL leads to a hybrid recommender that outperforms both individual components. The result is an effective recommender system that is easy to implement and independent of the tagged resources and that requires no offline training. The approach is thus not only suitable for dedicated tagging systems, but also for broader web systems in which tagging is merely a secondary feature.

Future work: In this study, we have evaluated LR_t and LR_n with fixed parameters (time-window length or number of considered interactions) for all users. A chance for improving the results further would be an analysis of personalized parameters for each user or different user groups. It would also be conceivable to use more refined methods to detect the user's current context when posting a new resource. For example, by looking at the requests, it might be possible to distinguish situations where users do research regarding one specific topic from situations where users just "stumble" through the system changing their focus based on what they find on each page. Another topic for future work is the tag candidate extraction. For this study, we extracted only the directly requested tags or query terms from a request. In tagging systems, one could leverage the semantic context of such tags (arising from the co-occurrence with other tags on the same resources). Finally, instead of exploiting retrieval interaction on the level of individual requests, one could attempt to identify sessions. These might yield a more comprehensive understanding of the current user context than considering individual requests independently.

REFERENCES

[1] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. 2010. The social bookmark and publication management system BibSonomy. *The VLDB Journal* 19, 6 (2010), 849–875. DOI: https://doi.org/10.1007/s00778-010-0208-4

[2] Robin Burke. 2010. Evaluating the Dynamic Properties of Recommendation Algorithms. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 225–228. DOI: https://doi.org/10.1145/1864708.1864753

[3] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 39–46. DOI: https://doi.org/10.1145/1864708.1864721

[4] Stephan Doerfel, Robert Jäschke, and Gerd Stumme. 2016. The Role of Cores in Recommender Benchmarking for Social Bookmarking Systems. *ACM Transactions on Intelligent Systems and Technology* 7, 3 (February 2016), 40:1–40:33. DOI: https://doi.org/10.1145/2700485

[5] Stephan Doerfel, Daniel Zoller, Philipp Singer, Thomas Niebler, Andreas Hotho, and Markus Strohmaier. 2016. What Users Actually do in a Social Tagging System: A Study of User Behavior in BibSonomy. *ACM Transactions on the Web* 10, 2 (2016), 14:1–14:32. DOI: https://doi.org/10.1145/2896821

[6] Folke Eisterlehner, Andreas Hotho, and Robert Jäschke (Eds.). 2009. *ECML PKDD Discovery Challenge 2009 (DC09)*. CEUR-WS.org, Vol. 497.

[7] Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *Journal of Machine Learning Research* 10 (2009), 2935–2962.

[8] M. Heckner, M. Heilemann, and C. Wolff. 2009. Personal Information Management vs. Resource Sharing: Towards a Model of Information Behaviour in Social Tagging Systems. In *Proceedings of the thrid AAAI Conference on Weblogs and Social Media*. San Jose, CA, USA.

[9] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. 2006. Information Retrieval in Folksonomies: Search and Ranking. In *The Semantic Web: Research and Applications: Third European Semantic Web Conference, ESWC 2006 Budva, Montenegro, June 11-14, 2006 Proceedings (LNCS)*, York Sure and John Domingue (Eds.), Vol. 4011. Springer Berlin Heidelberg, Berlin/Heidelberg, 411–426. DOI: https://doi.org/10.1007/11762256_31

[10] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. 2008. Tag Recommendations in Social Bookmarking Systems. *AI Communications* 21, 4 (2008), 231–247. DOI: https://doi.org/10.3233/AIC-2008-0438

[11] Marek Lipczak, Yeming Hu, Yael Kollet, and Evangelos Milios. 2009. Tag Sources for Recommendation in Collaborative Tagging Systems, See [6], 157–172.

[12] Feichao Ma, Wenqing Wang, and Zhihong Deng. 2013. TagRank: A new tag recommendation algorithm and recommender enhancement with data fusion techniques. In *Social Media Retrieval and Mining*, Shuigeng Zhou and Zhiang Wu (Eds.). Communications in Computer and Information Science, Vol. 387. Springer, Berlin/Heidelberg, 80–91. DOI: https://doi.org/10.1007/978-3-642-41629-3_7

[13] Thomas Niebler, Martin Becker, Daniel Zoller, Stephan Doerfel, and Andreas Hotho. 2016. FolkTrails: Interpreting Navigation Behavior in a Social Tagging System. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/2983323.2983686

[14] Maryam Ramezani. 2011. Improving graph-based approaches for personalized tag recommendation. *Journal of Emerging Technologies in Web Intelligence* 3, 2 (2011), 168–176. DOI: https://doi.org/10.4304/jetwi.3.2.168-176

[15] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Schmidt-Thieme Lars. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, Virginia, United States, 452–461.

[16] Steffen Rendle and Lars Schmidt-Thieme. 2009. Factor Models for Tag Recommendation in BibSonomy, See [6], 235–242.

[17] Paul Seitlinger, Dominik Kowald, Christoph Trattner, and Tobias Ley. 2013. Recommending tags with a model of human categorization. In *Proceedings of the 22nd International Conference on Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 2381–2386. DOI: https://doi.org/10.1145/2505515.2505625

[18] Sanjay Sood, Sara Owsley, Kristian Hammond, and Larry Birnbaum. 2007. TagAssist: Automatic Tag Suggestion for Blog Posts. In *Proceedings of the 1st International Conference on Weblogs and Social Media*. Boulder, Colorado, USA.

[19] Christoph Trattner, Dominik Kowald, Paul Seitlinger, Tobias Ley, and Simone Kopeinik. 2016. Modeling Activation Processes in Human Memory to Predict the Use of Tags in Social Bookmarking Systems. *Journal of Web Science* 2, 1 (2016), 1–16. DOI: https://doi.org/10.1561/106.00000004

[20] Dawei Yin, Liangjie Hong, Zhenzhen Xue, and Brian D. Davison. 2011. Temporal Dynamics of User Interests in Tagging Systems. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. AAAI.

[21] Lei Zhang, Jian Tang, and Ming Zhang. 2012. Integrating Temporal Usage Pattern into Personalized Tag Prediction. In *Proceedings of the 14th Asia-Pacific International Conference on Web Technologies and Applications (APWeb'12)*. Springer-Verlag, Berlin, Heidelberg, 354–365. DOI: https://doi.org/10.1007/978-3-642-29253-8_30