Modeling Activation Processes in Human Memory to Predict the Use of Tags in Social Bookmarking Systems

Christoph Trattner
NTNU & Know-Center
Norway & Austria
ctrattner@know-center.at

Dominik Kowald
Know-Center
Austria
dkowald@know-center.at

Paul Seitlinger
Graz University of Technology
Austria
paul.seitlinger@tugraz.at

Simone Kopeinik
Graz University of Technology
Austria
simone.kopeinik@tugraz.at

Tobias Ley
Tallinn University
Estonia
tley@tlu.ee

ABSTRACT

In recent years, several successful tag recommendation mechanisms have been developed that, among others, built upon Collaborative Filtering, Tensor Factorization, graph-based algorithms and simple "most popular tags" approaches. From an economic perspective, the latter approach has been convincing as calculating frequencies is computationally efficient and has shown to be effective with respect to different recommender evaluation metrics. In order to extend these conventional "most popular tags" approaches we introduce a tag recommendation algorithm that mimics the way humans draw on items in their long-term memory. Based on a theory of human memory, the approach estimates a tag's reuse probability as a function of usage frequency and recency in the user's past (base-level activation) as well as of the current semantic context (associative component).

Using four real-world folksonomies gathered from bookmarks in BibSonomy, CiteULike, Delicious and Flickr, we show how refining frequency-based estimates, by considering recency and semantic context, outperforms conventional "most popular tags" approaches and another existing and very effective but less theory-driven, time-dependent recommendation mechanism. By combining our approach with a resource-specific frequency analysis, our algorithm outperforms other well-established algorithms, such as Collaborative Filtering, FolkRank and Pairwise Interaction Tensor Factorization with respect to recommender accuracy and runtime. We conclude that our approach provides an accurate and computationally efficient model of a user's temporal tagging behavior. Moreover, we demonstrate how effective principles of recommender systems can be designed and implemented if human memory processes are taken into account.

This is a pre-print of the paper "Modeling Activation Processes in Human Memory to Predict the Use of Tags in Social Bookmarking Systems" accepted for publication in The Journal of Web Science. The final published version might look slightly different in style and content.
.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Keywords

personalized tag recommendations; time-dependent recommender systems; BLL equation; activation equation; ACT-R; human memory model; tag recency; BibSonomy; CiteULike; Delicious; Flickr

1. INTRODUCTION

One of the goals of Web Science as a new discipline is to understand the dynamics of human behavior and social interactions that shape the Web into a vast information network of content and people. As the Web evolves into a platform through which people interact with each other, communicate and express themselves, models of human behavior can shed light on why the Web forms as it does, and contribute to improving its underlying mechanisms. In this paper, we exemplify this idea in the context of social tagging. In particular, we show that a well-established model of human cognition both provides a good account of how people use tags and allows implementing an accurate and efficient tag recommendation mechanism.

When users categorize and tag resources on the Web (e.g., photos), they draw on their semantic-lexical memories to retrieve corresponding memory units. For instance, they might add the tag "Paris" as the photo shows the place they recently visited. Understanding the cognitive processes involved can help to predict individual tagging behavior [53] and to model phenomena on the collective level, such as the emergence of stable tag distributions [13]. To make appropriate memory units quickly available, human memory is very adaptive and tunes the activation of its units to statistical regularities of the environment (e.g., [4]): The more useful a memory unit has been and the stronger it is related to the current context (i.e., environmental cues), the higher is its activation level and hence, probability of being retrieved.

We assume that these activation processes also determine a user's tagging behavior and that the usage probability of a tag can be derived from estimates of its activation in the user's memory. According to [2], the activation of a tag should depend on at least two variables: i) the general usefulness of a tag in a user's tagging history and ii) its associations to the current context, i.e., to elements of the resource to be tagged. This means that a memory unit (e.g., the tag "recommender") is more likely to be brought into

consciousness, if we use it often and if it fits the current topic (e.g., "webscience"). In the next subsection, we present a simple formalism, which allows for a psychologically meaningful calculation and combination of the two variables of usefulness and context. As we will show below, this formalism helps to identify gaps in the current recommender research, namely to reconsider recent attempts to introduce time-dependent dynamics into recommender systems [62, 64]. We also show how this simple mechanism improves predictions of individual tagging behavior and how it can be used to design and implement an accurate and efficient recommendation mechanism.

1.1 Formalizing the Activation of Memory Units

Consider a user retrieving a unit from his/her memory, such as a tag that he/she has used previously. To derive its usefulness in the current context, we determine the activation A_i of this unit of memory i. According to the following activation equation, which is part of the declarative module of the cognitive architecture ACT-R (e.g., [2]), the usefulness is given by:

$$A_i = B_i + \underbrace{\sum_j W_j \cdot S_{j,i}}_{AssociativeComponent(AC)} \quad (1)$$

The B_i component represents the base-level activation and quantifies the general usefulness of a unit i by considering how frequently and recently it has been used in the past. It is given by the base-level learning (BLL) equation:

$$B_i = ln(\sum_{j=1}^{n} t_j^{-d}) \quad (2)$$

, where n is the frequency of the unit's occurrences and t_j is the recency, i.e., the time (in seconds) since the jth occurrence. For example, if a user has applied the two tags "recognition" and "recommender" with equal frequency but "recommender" has dominated the user's recent bookmarks[1] the equation predicts a higher activation for "recommender". The exponent d accounts for the power law of forgetting that each unit's activation caused by the jth occurrence decreases in time according to a power function [2].

The second component of equation 1 represents the associative activation that tunes the base-level activation of the unit i to the current context. The context is given by any contextual element j important in the current situation (e.g., the tags "memory" and "recollection"). Through learned associations, the contextual elements are connected with tag i and can increase i's activation depending on the weight W_j and the strength of association $S_{j,i}$. To simplify matters, we use the tags associated with a given resource r (due to previous tag assignments of other users) as the contextual elements. We derived W_j from the number of times tag j has been assigned to r, and $S_{j,i}$ from the number of co-occurrences between the tags i and j. Section 4 contains a more detailed and formal description of all calculation steps.

1.2 Research Questions

The introduction of the activation equation to model retrieval of tags from memory leads us to a number of research questions. First, equation 2 models time-dependent decay, i.e., the effect of recency on a memory unit's activation, according to a power law. When looking at recent tag recommendation models which take into account the time-dependent dynamics, they formalize the recency of tag use by means of linear [22] or exponential decay functions [65]. Whereas a linear function can be rejected for theoretical reasons (e.g., [4]), and from 100 years of empirical research into human memory (e.g., [11]), it is not clear whether an exponential or power law provides a better account of time-dependent decay in the use of tags. In Section 3 we therefore investigate the question:

- *RQ1*: Is an exponential or power decay function more appropriate to account for the effect of recency on a tag's reuse probability?

Experiments have shown that a substantial amount of tag assignments can be explained by modeling the strength of memory traces of tags [53]. Hence, given equations 1 and 2 correspond with individual tagging behavior, we assume that their formalism can also be used to predict a user's future tag reuse. To examine this assumption, we followed a two-stage approach. First, since equation 2 formalizes a fundamental memory process in a very efficient, i.e., computationally effortless way, we wanted to explore its tag reuse prediction accuracy independent of the associative activation component. Hence, the second question is:

- *RQ2*: Does the base-level learning (BLL) equation provide a valid model of a user's tagging behavior in the past to predict future individual tag assignments?

Furthermore, given equation 2 allows for accurate tag reuse prediction, we investigate the accuracy of equation 1 and raise the question:

- *RQ3*: Does the additional consideration of the associative component evoked by the current context further improve the accuracy of the base-level learning (BLL) equation to predict the individual tag reuse?

Finally, in order to realize a complete tag recommender that goes beyond solely predicting individual tag reuse, we take the results of RQ3 and combine the activation equation with popular tags that have been applied to the target resource by other users. When also considering other users' tags, this allows us to introduce new tags to the target user, namely tags that have not been used by the target user before (e.g., [37, 33, 29]). To this end, we weight these tags based on their frequency in the resource's tag assignments, hereinafter referred to as MostPopular$_r$ (MP$_r$). This allows us to compare the performance of the combination of the activation equation and MP$_r$ with well-established approaches, such as Collaborative Filtering (CF), FolkRank (FR) and Pairwise Interaction Tensor Factorization (PITF) which leads to the fourth and final research question of this work:

- *RQ4*: Can the whole activation equation, that considers base-level and associative activation, be applied and extended to create an effective and computationally efficient tag recommendation mechanism compared to state-of-the-art baseline approaches?

To summarize, the four research questions consider different levels of complexity. While RQ1 only analyzes the past tagging behavior of a user (see Section 3), RQ2 and RQ3 predict the individual reuse of tags from the users' previous vocabulary (see Section 6.1.1), without the current context (RQ2) as well as with the current context (RQ3). Finally, RQ4 considers also the introduction of new tags by imitating popular tags from other users, and thereby

[1]In this paper we refer to a bookmark as a user's post of an URL and corresponding tag assignments to a social tagging system.

allows us to compare our approach with current state-of-the-art tag recommendation mechanisms (see Section 6.1.2).

The remainder of this paper is organized as follows: We begin discussing related work (Section 2) and describing our empirical analysis to tackle our first research question (Section 3). In Section 4 we explain our approach. Section 5 describes the datasets, the experimental setup and the baseline algorithms used for evaluation. Section 6 addresses research questions 2 - 4 and summarizes the settings and results of our extensive evaluation. Section 7 concludes the paper by discussing our findings when deriving tag recommender mechanisms from empirical, cognitive research. This is followed by a short outlook into our future work in Section 7.1.

2. RELATED WORK

Recent years have shown that tagging is an important feature of the Social Web, supporting users with a simple mechanism to collaboratively organize and find content [26]. Although tagging has demonstrated to significantly improve search [19, 9, 56] (and in particular tags provided by the individual), it is known that users are typically lazy in providing tags for instance for their bookmarked resources. It is therefore not surprising that recent research is investigating personalized tag recommenders to support individual user in their tag application process. To date, the two following approaches have been established: graph (collaborative) -based and content-based tag recommender systems.

For the latter strand of research, the most recognizable work is a study conducted by Heymann et al. [20]. The paper illustrates that page-text is a significantly better predictor for the user's social tags than anchor-texts or surrounding hosts. This was explored within the Stanford domain and for tags gathered from the bookmarking system Delicious. Furthermore, there is the work of Marek et al. [34, 35, 36] or Lin et al. [32] that show the same effect for page-title and page-content. Another relevant and recent research in this context has been contributed by Lorince and Todd [37], Floeck et al. [12] and Moltedo et al. [41] who show on a theoretical and empirical level that existing tags (such as for instance existing tag clouds in LastFM) have influenced the way people generate their own tags for a target resource.

Other related work (as pointed before) is the research on graph-based approaches ranking the user's individual tags for a target resource. The probably most notable research in this context is presented by Hotho et al. and Jaschke et al. [21, 23] who introduce an algorithm called FolkRank (FR) which uses the structure of folksonomies for searching and ranking. These rankings can also be used to recommend tags. Subsequent studies of Marinho et al. [39, 23] or Hamouda & Wanas [17] show how the classic Collaborative Filtering (CF) approach could be adopted for the recommendation of tags. Significant studies of Rendle et al. [49], Wetzker et al. [60], Krestel et al. [30] or Rawashdeh et al. [45] introduce a factorization model, a Latent Dirichlet Allocation (LDA) model or a Link-Prediction model, based on the Katz measure, respectively, to recommend tags to users.

Although the latter mentioned approaches perform reasonably well, they are computational expensive compared to simple "most popular tags" approaches. Furthermore, they ignore recent observations with regard to social tagging systems, such as the variation of the individual tagging behavior over time [63]. To that end, recent research has made the first promising steps towards more accurate graph-based models that also account for the variable of time [62, 64]. The approaches have shown to outperform some of the current state-of-the-art tag recommender algorithms.

In line with the latter strand of research, in this paper we present a novel graph-based tag recommender mechanism that uses the activation equation, which is based on the principles of a popular model of human cognition called ACT-R (e.g., [2, 3]). We show that the approach is not only very simple and straightforward but also reveal that the algorithm outperforms current state-of-the-art graph-based (e.g., [60, 21, 23]) and the leading time-based [64] tag recommender approaches.

3. MODELING RECENCY EFFECTS IN SOCIAL TAGGING SYSTEMS

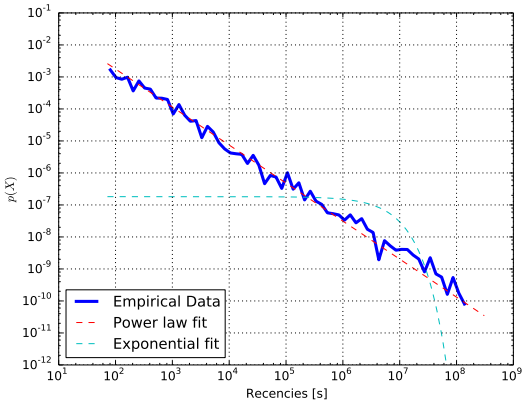
This section addresses our first research question as to whether the effect of recency decays according to an exponential or a power function. As described in Section 1, the same question has already been investigated in a different context (e.g., re-occurrence of words in New York Times headings) by Anderson and Schooler [4]. They found that the power function produces a better fit. Up to now, research on tag-based recommender systems has not applied a power function to model the temporal tagging patterns of users (only linear or exponential ones, see e.g., [22, 65, 64, 63, 7]). We therefore investigate, as to whether the results obtained by Anderson and Schooler [4] generalize to social tagging environments and thus, explore if users' tagging behavior justifies the application of the base-level learning (BLL) equation.

We approached this question investigating time-dependent user behavior in four representative dataset samples drawn from BibSonomy, CiteULike, Delicious and Flickr (for details see Section 5.1). We sorted a user's n bookmarks by time with the n^{th} bookmark being the most recently collected, and compared the tag assignments of her first n - 1 bookmarks with the n^{th} bookmark. Per user, we calculated the seconds elapsed since the last occurrence of each of the user's tags assigned to the n - 1 bookmarks. Additionally, we determined which of the user's tags had been reused in the n^{th} bookmark. To obtain a statistically reliable value, we pooled all users' tags with the same recency (seconds elapsed) and determined the proportion of tags reoccurring in the n^{th} bookmark as an estimate of the probability of future reuse.

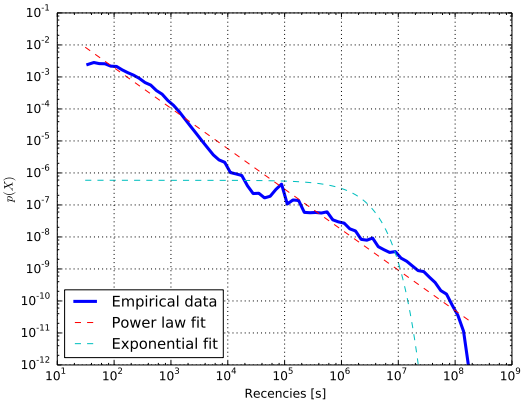
In Figure 1, we plotted the estimated probability $p(X)$ of tag reuse in the n^{th} bookmark against the number of seconds elapsed for each of the four datasets. The four plots in Figure 1 test the assumption of a power vs. exponential relationship by drawing the log-log-transformed re-occurrence probability against the seconds elapsed [4].

A glance at the plots in Figure 1 suggests that a power function might result in a better fit than the exponential function since it follows somewhat a straight line in a log-log-transformed plot (as suggested in [4]). To validate this hypothesis we made use of the python package *powerlaw* [1] which implements the method of Clauset et al. [8] to statistically quantify whether or not the observed empirical data can be better explained via a power law than an exponential function. As shown in Figure 1, in all four datasets the estimated power function (see also best values of x_{min} and α) provides a better fit for the data than an exponential function. To test for statical significance, we calculated the loglikelihood ratio R between the two observed functions and the empirical data as proposed in [1], where $R > 0$ and $p < .05$ means that the data is statistically more likely to follow a power distribution rather than an exponential one. As presented in Figure 1 this is the case in all four datasets. Note that the decay in Flickr is more pronounced than in BibSonomy, which might imply that scientific topics in BibSonomy (e.g., recommender research) do not change as fast as topics of photos of different leisure events (e.g., of the last weekend).

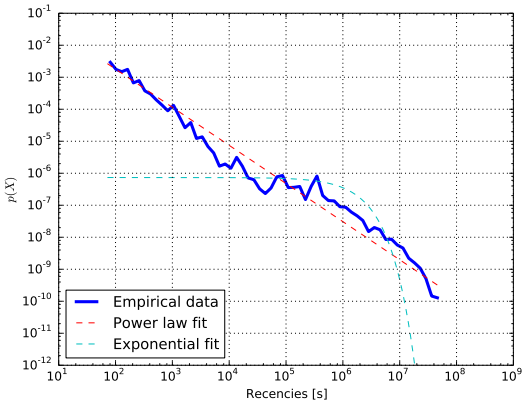
From this pattern of results we conclude that the findings revealed by Anderson and Schooler [4] generalize to social tagging



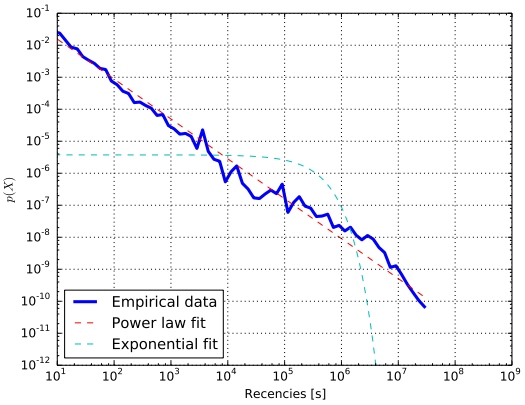
(a) BibSonomy (power vs. exponential fit)
Power law parameters: $x_{min} = 72$, $\alpha = 1.19$
Loglikelihood ratio: $R = 36.05$, $p < .001$



(b) CiteULike (power vs. exponential fit)
Power law parameters: $x_{min} = 31$, $\alpha = 1.26$
Loglikelihood ratio: $R = 246.08$, $p < .001$



(c) Delicious (power vs. exponential fit)
Power law parameters: $x_{min} = 32$, $\alpha = 1.19$
Loglikelihood ratio: $R = 60.92$, $p < .001$



(d) Flickr (power vs. exponential fit)
Power law parameters: $x_{min} = 2$, $\alpha = 1.25$
Loglikelihood ratio: $R = 117.55$, $p < .001$

Figure 1: Power law vs. exponential fit (first research question) of the time-dependent decay (measured in seconds) of individual tag reuse for BibSonomy, CiteULike, Delicious and Flickr. Parameters x_{min} and α of the fitted power function are also provided. Furthermore, values for R and p (p-value for R) are represented which is the loglikelihood ratio between the two candidate functions (power vs. exponential) fitted to the empirical data, where $R > 0$ and $p < .05$ means that the data is statistically more likely to follow a power distribution rather than an exponential one (which is the case in all four datasets - also visually).

environments: the effect of recency on the reuse probability of tags is more likely to follow a power law distribution than an exponential one. This speaks in favor of our approach's first component, the BLL equation, modeling a user's temporal tagging behavior via a power decay function.

The remainder of the present work deals with research questions 2 - 4 (see Section 1). Before we present the experimental setup (Section 5) and the results (Section 6) of the experiments addressing these questions, the next section describes our implementation of the two components of the activation equation, the BLL equation and the associative component, as a tag recommender.

4. A TAG RECOMMENDER BASED ON ACTIVATION IN MEMORY

The analysis in Section 3, revealed that the effect of recency on the reuse probability of tags follows a power law distribution. We therefore decided to implement the base-level learning (BLL) equation as a tag recommender and subsequently also extended the approach by the activation equation's second component, the associative component.

The first recommender is termed BLL as it implements the base-level activation equation (equation 2) in the form of a tag recommender using its two components of frequency and recency. Frequency-based models have been described in recommender systems research as "most popular tags" approaches [23]. There are different forms of these approaches, recommending either the most popular tags of the user, the resource or a mixture of both (see Section 5.3).

Recency-based recommender models (also referred to as time-dependent approaches) have been suggested in literature (e.g., [62, 64]) as an extension of "most popular tags" approaches. To date, this approaches modeled the time-dependent decay of tag reuse using a linear or exponential function (see Section 2) which is not in line with our findings. Hence, our second research question (whether base level activation can predict future tag use) translates

Symbol	Description
u	user
t	tag
r	resource
B	set of bookmarks / posts
B_{train}	train set
B_{test}	test set
B_t	set of bookmarks tagged by t
U	set of users
T	set of tags
R	set of resources
T_r	T of resource r
T_u	T of user u
Y	set of tag assignments
Y_u	Y of user u
Y_t	Y of tag t
Y_r	Y of resource r
Y_b	Y of bookmark b
$Y_{t,r}$	Y of tag t and resource r
$Y_{t,u}$	Y of tag t and user u
β	mixing parameter
d	decay parameter
c	context cue
$S(c,t)$	association strength between c and t
$A(t,u,r)$	activation of tag t for u and r
$B(t,u)$	base-level activation of tag t for u
$\widetilde{T}_k(u,r)$	set of top k recommended tags for u and r
$T(u,r)$	set of relevant tags used by u for r

Table 1: Overview of notations used in this paper.

into whether it is possible to improve a "most popular tags" recommender with a recency component based on a power decay function.

4.1 Formalization

In this section we present the formalization of our proposed method. The notations we shall use throughout the paper are defined in Table 1. To realize this recommender the following steps where performed: For each tag in a user's training set B_{train}, we have calculated the base-level activation $B(t,u)$ of a given tag t in a user u's set of tag assignments, Y_u. First, we determined a reference timestamp $timestamp_{u,ref}$ (in seconds) that is the timestamp of the most recent bookmark of user u. In our dataset sample, $timestamp_{u,ref}$ corresponds to the timestamp of the user's bookmark that has been selected for the test set (see Section 5.2).

If j = 1 ... n indexes all tag assignments in Y_u, the recency of a tag assignment is given by $timestamp_{u,ref} - timestamp_{t,u,j}$. $B(t,u)$ of tag t for user u is given by the BLL equation:

$$B(t,u) = \ln(\sum_{j=1}^{n} (timestamp_{u,ref} - timestamp_{t,u,j})^{-d}) \tag{3}$$

, where d is set to .5 based on [2]. We also tried other d values, such as 1.2 based on the best α values of our empirical analysis in Section 3, but this did not lead to better results in terms of recommender accuracy. Thus, we decided to keep the value from the literature.

In order to map the values onto a range of 0 to 1 we applied a softmax function as proposed in related work [40]:

$$\underset{T_u}{\text{soft}\max}(B(t,u)) = \frac{\exp(B(t,u))}{\sum_{t' \in T_u} \exp(B(t',u))} \tag{4}$$

, where t' is a tag in T_u, the set of tags used by user u in the past.

To investigate our third research question (as to whether the BLL equation can be further improved by also considering the associative component evoked by the current context) we have implemented equation 1 in form of:

$$A(t,u,r) = \underbrace{\underbrace{\underset{T_u}{\text{soft}\max}(B(t,u))}_{BLL} + \sum_{c \in T_r} (|Y_{c,r}| \cdot S(c,t))}_{BLL_{AC}} \tag{5}$$

To calculate the variables of the associative component, i.e., to model a user's semantic context, we looked at the set of tags T_r assigned by other users to the given resource r. A user's semantic context certainly consists of a greater variety of aspects, such as content words in the title or in the page text. However, since not all of our datasets contain title information or page text and other studies have convincingly demonstrated the impact of a resource's prominent tags on a user's tagging behavior (e.g., [37, 33]), we decided to approximate the context by means of other users' tags.

When applying the formula to a recommender system, related literature [54, 58] suggests to use a measure of normalized tag co-occurrence to represent the strength of an association. Accordingly, we define the co-occurrence between two tags as the number of bookmarks in which both tags are included. To add meaning to the co-occurrence value, the overall frequency of the two tags is also taken into consideration. This is done by normalizing the co-occurrence value according to the Jaccard coefficient (6) following the approach described in [54]:

$$S(c,t) = \frac{|B_c \cap B_t|}{|B_c \cup B_t|} \tag{6}$$

In our implementation, $S(c,t)$ is calculated as an association value between a tag previously given by the target user (t) and a tag that has been assigned to a resource of interest (c). Based on a tag co-occurrence matrix that depicts the tag relations of an entire data set, information about how many times two tags co-occur ($B_c \cap B_t$) in bookmarks is retrieved and set into relation with the number of bookmarks in which at least one of the two tags appear ($B_c \cup B_t$). We set the attentional weight W_c of c to the number of times c occurred in the tag assignments of the target resource, i.e., $|Y_{c,r}|$.

Hence, the associative component in equation 5 works in a similar way as resource-based Collaborative Filtering in the tag recommender literature [57]. This means, that tags with a higher similarity to the target resource (measured by tag co-occurrence) get a higher associative activation value than tags with a smaller usefulness in the current context.

Finally, to examine our fourth research question (as to whether the activation equation can be implemented in form of an effective recommender mechanism) we extended equation 5 by also considering the most popular tags in the tag assignments of the resource Y_r (MP_r, i.e., $\arg\max^k_{t \in T_r}(|Y_{t,r}|)$) [21]. This simple extension was necessary for the prediction of new and plausible tags that a user has not assigned in her previous tagging history (e.g., [37, 33, 29]). Therefore, we have selected MP_r over other methods like CF because as shown in related work [12, 52, 14, 15], users in social tagging systems are more likely to directly imitate tags that have

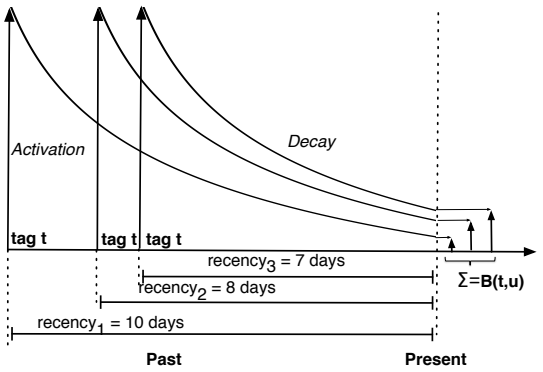


Figure 2: Example for applying the BLL equation (first component of the activation equation) to estimate the activation value of a tag t and to show the advantage over the conventional "most popular tags by user" (MP_u) approach.

already been assigned to a target resource. Finally, the top-k recommended tags for a given user u and resource r are calculated by the following equation:

$$\widetilde{T}_k(u,r) = \arg\max^k_{t \in T_u, T_r} (\underbrace{\beta \underbrace{\underset{T_u}{\text{soft}\max}(A(t,u,r)}_{BLL_{AC}} + (1-\beta) \underset{T_r}{\text{soft}\max}(|Y_{t,r}|)}_{BLL_{AC}+MP_r}) \quad (7)$$

, where β is used to weight the two components, i.e., the activation values $A(t,u,r)$ and the most popular tags of the target resource given by MP_r. Results presented in Section 6 were calculated using β = .5.

4.2 Illustration

In order to further clarify how we have applied the equations to characterize a user's individual tagging history, we provide two simple examples illustrated in Figures 2 and 3. That way, we also aim at demonstrating the advantage of our approach over conventional "most popular tags" approaches.

The example in Figure 2 shows how the BLL equation provides a more differentiated characterization of a user's tagging pattern than the "most popular tags by user" (MP_u) approach. In this example, a user u applied a tag t three times, i.e., n = 3. We assume that she applied the tag ten, eight and seven days ago. The three corresponding recency values are recency_1 = 10, recency_2 = 8 and recency_3 = 7. We have calculated the recency of a tag t's use by subtracting the timestamp of the j^{th} use of t from the timestamp of u's most recent bookmark. Each of the three uses of t activates the corresponding memory unit. In Figure 2, the upward directed arrows symbolize this hypothesized activation. Due to the power-law of forgetting, each activation decreases in time (represented by the sloping curves) and, according to [2], each of the three recency values is raised by the power d = -.5. Finally, the base-level activation of the memory unit for tag t is given by summing the remaining effects of the three tag uses, i.e., $ln(10^{-.5} + 8^{-.5} + 7^{-.5})$,

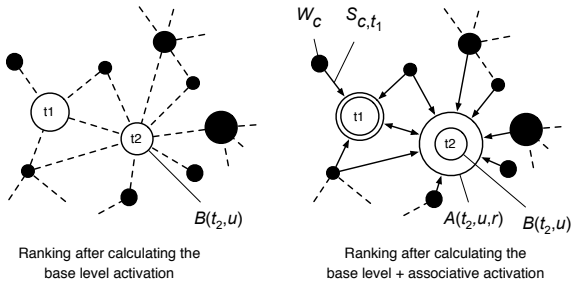


Figure 3: Example showing the impact of associative activation (second component of the activation equation).
***Note*: black filled nodes and unfilled nodes represent contextual and target tags, respectively; their sizes represent their attentional weights W_c (in case of contextual tags) and activation (in case of the target tags t_1 and t_2). The edge length represents the co-occurrence-based association strength $S_{c,t}$. Left panel: ranking based on base-level activation $B(t,u)$ not taking into account the contextual tags. Right panel: refined ranking after considering the associative activation evoked by contextual tags, resulting in the full activation $A(t,u,r)$.**

resulting in the base-level activation of .05. To the contrary, a conventional "most popular tags by user" (MP_u) approach, only takes into account the tag's usage frequency and thus, treats every tag assignment the same, independent of the time elapsed since its use. Given the user's entire set of tag assignments $Y_{t,u}$ encompasses 10 assignments, this approach would yield a value of .3 (3 / 10). This should demonstrate that the BLL equation allows for a more differentiated characterization of a user's tagging history than MPu.

In the example of Figure 3, we show the additional impact of the associative activation defined by the second component of the activation equation. The associative activation is evoked by the current context, i.e., the tags assigned by preceding users to the target resource (in the following called contextual tags). The left panel of Figure 3 shows two target tags, t_1 and t_2 exhibiting different base-level activation levels (represented by the circle size): t_1 reaches a higher base-level activation and thus, a higher ranking than t_2. This relationship changes when considering the influence of the contextual tags, as schematically visualized in the right panel of Figure 3. These contextual tags are represented by the black nodes. Depending on their weights W_j (represented by the size of the black-filled nodes) and strength of association $S_{j,i}$ (represented by the length of the edges), the contextual tags spread additional associative activation to the target tags t_1 and t_2, i.e., make them more easily available for retrieval and use. t_2 is stronger associated with the contextual tags and thus, receives stronger associative activation than t_1. Summarizing, we can see that t_2 is assigned a higher ranking than t_1 when considering both, the base-level and associative activation by means of the full activation equation.

5. EXPERIMENTAL SETUP

In this section we describe in detail the datasets, the evaluation method, the evaluation metrics and the baselines algorithms used for our experiments.

5.1 Datasets

For the purpose of our study and for reasons of reproducibility, our investigations focused on four well-known and freely-available folksonomy datasets. To test our approach on both, broad and nar-

Dataset	p	$\|B\|$	$\|U\|$	$\|R\|$	$\|T\|$	$\|Y\|$
BibSonomy	-	400,983	5,488	346,444	103,503	1,479,970
	3	41,764	788	8,711	5,757	161,509
CiteULike	-	3,879,371	83,225	2,955,132	800,052	16,703,839
	3	735,292	17,983	149,220	67,072	2,242,849
Delicious	-	1,416,151	15,980	931,993	180,084	4,107,107
	3	466,480	9,102	58,025	16,574	1,506,231
Flickr	-	864,679	9,590	864,679	127,599	3,552,540
	3	860,135	8,332	860,135	58,831	3,465,346

Table 2: Properties of the datasets, where $|B|$ is the number of bookmarks, $|U|$ the number of users, $|R|$ the number of resources, $|T|$ the number of tags and $|Y|$ the number of tag assignments. As shown in column "p", we applied both: a p-core pruning approach (represented by "3") as well as no p-core pruning (represented by "-").

row folksonomies [18] (in broad folksonomy many users are allowed to annotate a particular resource while in a narrow folksonomy only the user who has uploaded the resource is permitted to apply tags), datasets from BibSonomy, CiteUlike, Delicious (broad folksonomies) and Flickr (narrow folksonomy) were selected. These dataset have been also used in many of the related work in tag-based recommender systems and can be seen as the state-of-the-art benchmarking datasets (see also e.g., [16, 23, 10, 5]).

BibSonomy: The dataset of the social bookmark and publication sharing system BibSonomy[2] is freely available and can be downloaded for scientific purposes[3] (2013-07-01). For our evaluation we concentrated on the tags assigned to bookmarks which resulted in 400,983 bookmarks, 5,488 users, 346,444 resources, 103,503 tags and 1,479,970 tag assignments.

CiteULike: CiteULike[4] is a reference management system which gives free access to their data to researchers for non-commercial uses[5] (2013-03-10). The CiteULike dataset consists of 3,879,371 bookmarks, 83,225 users, 2,955,132 resources, 800,052 tags and 16,703,839 tag assignments.

Delicious: The dataset of the social bookmarking Web service Delicious[6] is freely available for scientific purposes and was crawled and provided by the University of Koblenz[7] (2010-01-07) within the Tagora EU project[8]. The dataset contains 47,208,747 bookmarks, 532,924 users, 17,262,480 resources, 2,481,698 tags and 140,126,586 tag assignments.

Flickr: Flickr[9] is an image hosting and sharing platform which also offers online community elements. As the Delicious dataset, the Flickr dataset is also provided by the University of Koblenz (see Delicious dataset) and contains 28,153,045 bookmarks, 319,686 users, 28,153,045 resources, 1,607,879 tags, and 112,900,000 tag assignments.

To reduce computational effort (see also Section 6.2), we applied the dataset pruning technique proposed by Gemmel et al. [16] to the very big Delicious and Flickr datasets. Thus, for these two datasets,

[2]http://www.bibsonomy.org/
[3]http://www.kde.cs.uni-kassel.de/bibsonomy/dumps
[4]http://www.citeulike.org/
[5]http://www.citeulike.org/faq/data.adp
[6]https://delicious.com/
[7]https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/
[8]http://www.tagora-project.eu/
[9]http://www.flickr.com/

Algorithm(s)	Parameter(s)	Value
CF	k	20
BLL, BLL$_{AC}$, BLL+MP$_r$, BLL$_{AC}$+MP$_r$	d	.5
APR, FR	d	.7
APR, FR	l	10
MP$_{u,r}$, GIRPTM, BLL+MP$_r$, BLL$_{AC}$+MP$_r$	β	.5
FM, PITF	k_U, k_R, k_T	256
FM, PITF	l	50
FM, PITF	α	0.01
FM, PITF	λ	.0

Table 3: Hyperparameters of the algorithms as used in the experiments.

we randomly selected 3%[10] of the user profiles (i.e., all the bookmarks of these users) in the folksonomies. However, as shown in Table 2, the pruned datasets of Delicious and Flickr still remain larger than the dataset of BibSonomy. Furthermore, according to Gemmel et al. [16], when following this pruning method, experiments on larger dataset samples provide near identical trends in the algorithmic results.

Since automatically generated tags affect the performance of the tag recommender systems, we excluded all of those tags from the datasets (e.g., we excluded the *no-tag*, *bibtex-import*-tag, etc.). Furthermore, we decapitalized all tags as suggested by related work in the field (e.g., [49]). The overall dataset statistics can be found in Table 2. As shown in column "p", we applied both: a p-core pruning approach [6] (represented by "3") to capture the issues of data sparseness, as well as no p-core pruning (shown as "-") to capture the issue of cold-start users or items (see Doerfel et al. [10]), respectively. This p-core pruning is an iterative process where in each iteration all resources, tags and users are deleted that occur less than p times in a dataset. This algorithm terminates when no more tag assignments can be deleted which ensures that all resources, tags and users can be found at least p times in the remaining core [23, 33].

5.2 Methodology

To evaluate our tag recommender approach we used a leave-post-out method as proposed by popular and related work in this area (e.g., [23]). To that end, we created two datasets, one set for training and the other set for testing. To split up the dataset in two, we removed each user's latest bookmark in time from the original dataset and added it to the test set. Each bookmark in the test set consists of a collection of one or more tags to which we further refer as relevant tags. The now reduced version of the original dataset was used for training, the newly created one for testing. This procedure is a plausible simulation of a real-world environment as it retains the chronological order of a user's bookmarks and depicts a suggested offline-evaluation procedure for time-based recommender systems [7]. To quantify the performance of our approaches, a set of well-known, standard information retrieval performance metrics were used [23, 33]:

Recall (R) is calculated as the number of correctly recommended tags divided by the number of relevant tags, where $\widetilde{T}_k(u,r)$ denotes the k recommended tags and $T(u,r)$ the list of relevant tags of a user u for resource r that is determined by the bookmark in the test

[10]The reason for choosing this 3% limit was the fact that the PITF algorithm calculations (see also APR and FR in Section 5.3) took around 14 days on a 2.0 GHz six-core Intel Xeon E5-2620 processors with 128 GB of RAM (see Section 6.2), which we found to be a fair upper runtime limit for any of our calculations (which we performed in memory).

set B_{test} [59]:

$$R@k = \frac{1}{|B_{test}|} \sum_{u,r \in B_{test}} \frac{|\widetilde{T}_k(u,r) \cap T(u,r)|}{|T(u,r)|} \quad (8)$$

Precision (P) is calculated as the number of correctly recommended tags divided by the number of recommended tags k [59]:

$$P@k = \frac{1}{|B_{test}|} \sum_{u,r \in B_{test}} \frac{|\widetilde{T}_k(u,r) \cap T(u,r)|}{k} \quad (9)$$

F1-score (F1) combines precision and recall into one score [59]:

$$F1@k = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k} \quad (10)$$

Mean reciprocal rank (MRR) is the sum of the reciprocal ranks of all relevant tags in the list of recommended tags. This means that a higher MRR is achieved if relevant tags occur at the beginning of the recommended tag list [45]:

$$MRR = \frac{1}{|B_{test}|} \sum_{u,r \in B_{test}} \frac{1}{|T(u,r)|} \sum_{t \in T(u,r)} \frac{1}{rank(t)} \quad (11)$$

Mean average precision (MAP) is an extension of the precision metric that additionally looks at the ranking of recommended tags. MAP is described in the subsequent formula, where B_k is 1 if the recommended tag at position k is among the relevant tags and 0 otherwise. $P_{u,r}@k$ depicts Precision@k calculated for user u and resource r [45]:

$$MAP = \frac{1}{|B_{test}|} \sum_{u,r \in B_{test}} \frac{1}{|T(u,r)|} \sum_{k=1}^{|\widetilde{T}_k(u,r)|} B_k \cdot P_{u,r}@k \quad (12)$$

In particular, we report R@k, P@k, MRR and MAP for k = 10 and F1-Score ($F_1@k$) for k = 5 recommended tags[11].

5.3 Baseline Algorithms

We compared the results of our approach to several baseline tag recommender algorithms. The algorithms were selected based on their popularity in the community, performance and novelty (see also [38, 5]). Hyperparameters for the algorithms, as they were used for the experiments, are found in Table 3.

MostPopular (MP): This approach recommends for any user $u \in U$ and any resource $r \in R$ the same set of tags $\widetilde{T}(u,r)$. This set of tags is weighted by the frequency in all tag assignments Y [24]:

$$\widetilde{T}_k(u,r) = \arg\max_{t \in T}^{k}(|Y_t|) \quad (13)$$

MostPopular$_u$ (MP$_u$): The *most popular tags by user* approach suggests the most frequent tags in the tag assignments of the user Y_u [24].

MostPopular$_r$ (MP$_r$): The *most popular tags by resource* algorithm weights the tags based on their frequency in the tag assignments of the resource Y_r [24].

MostPopular$_{u,r}$ (MP$_{u,r}$): This algorithm is a mixture of the most popular tags by user and resource approaches:

$$\widetilde{T}_k(u,r) = \arg\max_{t \in T_u, T_r}^{k}(\beta|Y_{t,u}| + (1-\beta)|Y_{t,r}|) \quad (14)$$

The β parameter can be used to balance the influence of the user and the resource components [24] and was set to .5 as it is also done in our approaches.

Collaborative Filtering (CF): Marinho et al. [39] described how the classic Collaborative Filtering (CF) approach [50] can be adapted for tag recommendations. Since folksonomies have ternary relations (users, resources and tags), the classic CF approach can not be applied directly. Thus, the neighborhood N_u^k of a user u is formed based on the tag assignments in the user profile Y_u. Furthermore, in CF-based tag recommendations only the subset V_r of users that have tagged the active resource r are taken into account when calculating the user neighborhood. The set of n recommended tags can then be determined based on this neighborhood [39, 23]:

$$\widetilde{T}_k(u,r) = \arg\max_{t \in T}^{k}(\sum_{v \in N_u^k} sim(Y_u, Y_v) \cdot \delta(v,r,t)) \quad (15)$$

, where $\delta(v,r,t) := 1$ if $(v,r,t) \in Y_t$ and 0 else. The only variable parameter here is the number of users k in the neighborhood which has to be set in advance. We used a neighborhood size k of 20 as suggested in related work [16][12]. There are different ways to calculate the similarity $sim(Y_u, Y_v)$ between two users u and v. For our experiments we applied the Jaccard's similarity. We also tried the Okapi BM25 similarity measure (usually the best measure to calculate the similarity between users) [43, 44, 61] where we reached almost the same results as with Jaccard's, but with a significantly higher computational effort, especially in the case of the bigger datasets.

Adapted PageRank (APR): Hotho et al. [21] adapted the well-known PageRank algorithm [42] in order to rank the nodes within the graph structure of a folksonomy. This is based on the idea that a resource is important if it is tagged with important tags by important users. Thus, the folksonomy has to be converted into an undirected graph where the set of nodes s is the disjoint union of all users U, resources R and tags T: $s = U \cup R \cup T$. The co-occurences of users and resources, users and tags and resources and tags are treated as weighted edges in this graph and can also be represented as an adjacency matrix A. The update of the weightings is done using the following formula where $\vec{p}$ is a preference vector and d is a variable to set its impact [21]:

$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p} \quad (16)$$

For recommending tags, the preference vector $\vec{p}$ is used to give higher weights to the target user and resource of the recommendation task. While all other users and resources get a weight of 1, they get a weight of $1 + |U|$ and $1 + |R|$ [23]. Please have a look at the next paragraph about FolkRank to get more information about the used implementation and parameters.

FolkRank (FR): The FolkRank algorithm is an extension of the Adapted PageRank approach that was also proposed by Hotho et al. [21]. This extension gives a higher importance to the preference vector $\vec{p}$ using a differential approach, where $\vec{w}^{(0)}$ is the weighting vector calculated using the Adapted PageRank algorithm with $\vec{p}$ = 1 and $\vec{w}^{(1)}$ is the result with a $\vec{p}$-setting as described above:

$$\vec{w} = \vec{w}^{(1)} - \vec{w}^{(0)} \quad (17)$$

Our Adapted PageRank and FolkRank implementations are based on an open-source Java implementation provided by the University

[11] $F_1@5$ was also used as the main performance metric in the ECML PKDD Discovery Challenge 2009: http://www.kde.cs.uni-kassel.de/ws/dc09/.

[12] We also tested other values for k but observed that CF did not generated significant higher values of estimate when setting $k > 20$.

of Kassel[13]. In this implementation the parameter d is set to .7 and the maximum number of iterations l is set to 10 [23].

	p	Metric	MP_u	GIRP	BLL	BLL_{AC}
BibSonomy	-	F_1@5	.152	.157	.162	**.169***
		MRR	.114	.119	.125	**.133**
		MAP	.148	.155	.162	**.172**
	3	F_1@5	.215	.221	.228	**.292**$^{***\circ\circ\circ}$
		MRR	.202	.210	.230	**.286**$^{***\circ\circ\circ}$
		MAP	.238	.247	.272	**.345**$^{***\circ\circ\circ}$
CiteULike	-	F_1@5	.185	.194	.201***	**.211**$^{***\circ\circ\circ}$
		MRR	.165	.182	.193***	**.205**$^{***\circ\circ\circ}$
		MAP	.194	.213	.227***	**.242**$^{***\circ\circ\circ}$
	3	F_1@5	.272	.291	.300**	**.336**$^{***\circ\circ\circ}$
		MRR	.268	.294	.319***	**.365**$^{***\circ\circ\circ}$
		MAP	.305	.337	.366***	**.424**$^{***\circ\circ\circ}$
Delicious	-	F_1@5	.170	.184	.196***	**.231**$^{***\circ\circ\circ}$
		MRR	.155	.178	.197***	**.230**$^{***\circ\circ\circ}$
		MAP	.180	.207	.230***	**.274**$^{***\circ\circ\circ}$
	3	F_1@5	.193	.194	.206**	**.311**$^{***\circ\circ\circ}$
		MRR	.170	.177	.193**	**.296**$^{***\circ\circ\circ}$
		MAP	.198	.207	.227**	**.364**$^{***\circ\circ\circ}$
Flickr	-	F_1@5	.435	.509	.523*	**.523***
		MRR	.360	.445	.466***	**.466*****
		MAP	.468	.590	.619***	**.619*****
	3	F_1@5	.488	.577	.592*	**.592***
		MRR	.407	.511	.533***	**.533*****
		MAP	.527	.676	.707***	**.707*****

Table 4: F_1@5, MRR and MAP values for BibSonomy, CiteULike, Delicious and Flickr (with no core and p-core = 3) showing that the BLL equation provides a valid model of a user's tagging behavior to predict tags (second research question). Moreover, the results imply that using the activation equation (BLL_{AC}) to also take into account semantic cues (i.e, associations with resource tags) can further improve this model (third research questions). The symbols *, ** and *** indicate statistically significant differences based on a Wilcoxon Ranked Sum test between BLL, BLL_{AC} and GIRP at α level .05, .01 and .001, respectively; $^{\circ}$, $^{\circ\circ}$ and $^{\circ\circ\circ}$ indicate statistically significant differences between BLL_{AC} and BLL at the same α levels.

Factorization Machines (FM): Rendle [46] introduced Factorization Machines which combine the advantages of Support Vector Machines (SVM) with factorization models to build a general prediction model that is also capable of tag recommendations. More information about the used framework and parameters can be found in the next paragraph describing the PITF approach.

Pairwise Interaction Tensor Factorization (PITF): This approach proposed by Rendle and Schmidt-Thieme [49] is an extension of factorization models based on the Tucker Decomposition (TD) model that explicitly models the pairwise interactions between users, resources and tags. The FM and PITF results presented in this paper were calculated using the open-source C++ tag recommender framework provided by the University of Konstanz[14]. We set the dimensions of factorization k_U, k_R and k_T to 256, the learning rate α to .01, the regularization constant λ to .0 and the number of iterations l to 50 as suggested in [49][15].

Temporal Tag Usage Patterns (GIRP): This time-dependent tag-recommender algorithm was presented by Zhang et al. [64] and is based on the frequency and the temporal usage of a user's tag assignments. In contrast to our BLL and BLL_{AC} approaches, GIRP models the temporal tag usage with an exponential function rather than a power function (see Section 3).

GIRP with Tag Relevance to Resource (GIRPTM): This is an extension of the GIRP algorithm by the resource component (MP_r), which is also done in our BLL+MP_r and BLL_{AC}+MP_r approaches [64].

6. RESULTS AND DISCUSSION

In this section we present the results of our experiments in respect to recommender accuracy and runtime.

6.1 Recommender Accuracy

The presentation of the evaluation results is organized in line with our research questions 2 - 4, as introduced in Section 1. With respect to the recommender accuracy, we will turn our attention first, to the BLL equation and its validity to model individual tagging behavior (RQ2), second, to the impact of context information when added to the BLL equation (BLL_{AC}) (RQ3) and third, to a comparison of our context enriched BLL implementation ($BLL_{AC}+MP_r$) with state-of-the-art baseline approaches (RQ4). We report these points in two subsections where the first one looks solely at the individual tag reuse prediction (RQ2 and RQ3) and the second one at the prediction of tag reuse in combination with tag imitation (RQ4).

6.1.1 Predicting Tag Reuse

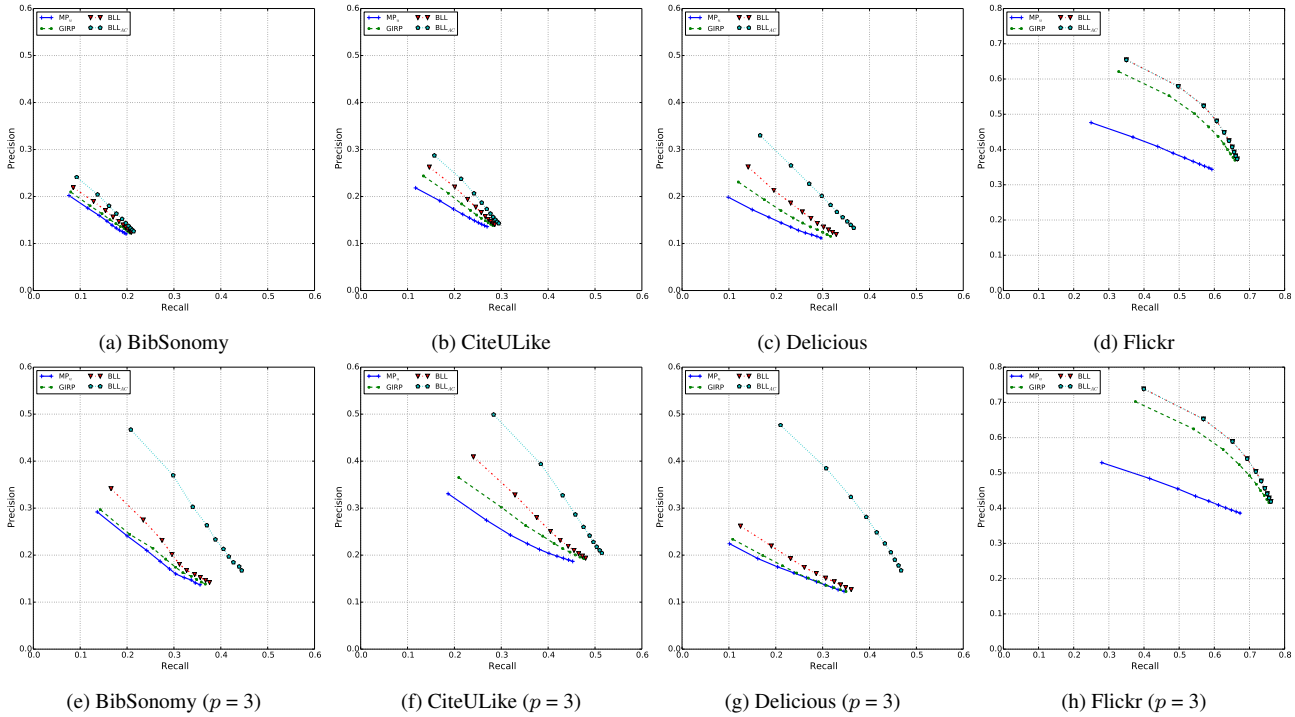
The BLL equation models the user's tagging behavior with respect to frequency and recency. While the frequency of tag use is a fairly common parameter for tag recommendations, the factor of time, that models the effects of a user's long term memory (as described through recency), is expected to bring additional value to tag recommendation approaches. That is why we investigate our second research question by determining the effect of the recency component on tag assignments.

When comparing BLL with MP_u and GIRP, the results reported in Table 4 and Figure 4 clearly show that the time-dependent algorithms BLL and GIRP both outperform the frequency-based MP_u approach. Looking further at the two time-dependent algorithms, BLL reaches higher levels of accuracy than the less theory-driven GIRP algorithm in both settings (with no core and p-core = 3). Even more apparent is the impact of the recency component in the narrow folksonomy (Flickr). Unlike the broad folksonomies (BibSonomy, CiteULike and Delicious), the Flickr dataset has no tags of other users available for the target resource. Therefore, a user needs to assign tags without having the inspiration of previously given tags. We assume that the user, to this end, needs to draw on her long term memory that the BLL equation aims to mimic. In

[13]http://www.kde.cs.uni-kassel.de/code

[14]http://www.informatik.uni-konstanz.de/rendle/software/tag-recommender/

[15]We also conducted experiments with factors of 64, 128 and 512 and with more and less than 50 iterations. Across all datasets the setting of 256 factors and 50 iterations showed almost always the best results. Factors less than 256 decreased the results significantly while factors higher than 256 did not result in any higher estimates while varying the number of iterations. The same is true for α and λ.



(a) BibSonomy (b) CiteULike (c) Delicious (d) Flickr

(e) BibSonomy (p = 3) (f) CiteULike (p = 3) (g) Delicious (p = 3) (h) Flickr (p = 3)

Figure 4: Recall/Precision plots for BibSonomy, CiteULike, Delicious and Flickr (no core and p-core = 3) showing the performance of BLL and BLL$_{AC}$ along with MP$_u$ and GIRP for 1 - 10 recommended tags (k).

summary, these results provide strong evidence that the BLL equation provides a valid model of a user's tagging behavior to predict tags (second research question). These results are further proved to be statistically significant based on a Wilcoxon Rank Sum test that is also shown in Table 4.

By expanding BLL to BLL$_{AC}$, we implement the activation equation as explained in Section 4 in order to address our third research question. The activation equation enriches the base-level activation (i.e., frequency and recency of tag use) by adding contextual activation through tags previously assigned to the target resource. Looking at the results of this experiment, as illustrated in Table 4 and Figure 4, a number of interesting aspects appear. For one thing, the results demonstrate that BLL$_{AC}$ reveals only a small improvement over BLL, when applied on the unfiltered datasets (no p-core) of the broad folksonomies (BibSonomy, CiteULike and Delicious). However, this changes when looking at the results for the p-core pruned datasets (p = 3). Caused by the higher number of tags assigned to each resource, the contextual activation gains impact. This leads to considerably increased values for all of the used metrics ($F1$@5, MRR, MAP). One might wonder why the results of BLL and BLL$_{AC}$ are the same in the case of the narrow folksonomy (Flickr). This is, in fact, an expected outcome. The Flickr dataset represents a narrow folksonomy and thus, resources are tagged by only one user (i.e, the one that has uploaded it), the model of the resource component does not generate additional value. Since the fine-tuning or re-ranking of the user tags based on context cues increases the recommender accuracy in the broad folksonomies, we can also answer the third research question positively.

6.1.2 *Predicting Tag Reuse and Tag Imitation*

To address our fourth and last research question, we combine our BLL$_{AC}$ approach with MP$_r$, which leads to BLL$_{AC}$+MP$_r$. Hereby, BLL$_{AC}$ models the context-aware user component while MP$_r$ further models the resource component to complementary take into account new tags that have not been used by the target user in the past. The results presented in Table 5 show that this approach outperforms a set of state-of-the-art baseline algorithms as well as BLL+MP$_r$ (without contextual activation of the user tags). Moreover, the three time-dependent algorithms (GIRPTM, BLL+MP$_r$ and BLL$_{AC}$+MP$_r$) produce higher estimates (F_1@5, MRR and MAP) across all datasets as well as in both settings (with no core and p-core = 3). Moreover, an important observation is that our BLL$_{AC}$+MP$_r$ approach also significantly outperforms GIRPTM, the currently leading, graph-based time-depended tag recommendation algorithm. Particularly good results are shown for ranking-dependent metrics such as MRR and MAP. This observation clearly illustrates the advantages of our approach that is build upon long-standing models of human memory theory, over the less-theory driven GIRPTM algorithm that also utilizes time information of social tags.

Another aspect worth discussing is the contrast of the results illustrated in Table 4, where BLL$_{AC}$ reaches substantially higher levels of accuracy than BLL, to the results outlined in Table 5, where BLL$_{AC}$+MP$_r$ only indicate marginal improvements over BLL+MP$_r$. In our opinion, this effect appears because the resource tag information depicted in MP$_r$ is congruent with data used for the contextual activation in BLL$_{AC}$. This finding suggests that the use of different resource metadata, such as title or body-text, may be valuable when specifying the context in BLL$_{AC}$ (see also Section 7). Similar observations can be made when looking at the Recall/Precision curves in Figure 5 that show the recommender performance of the approaches for 1 - 10 recommended tags (k).

In summary, our results clearly imply that the activation equation by Anderson et al. [2] can be used to implement a highly effective recommender approach. Overall, the simulations demonstrate

	p	Measure	MP	MP$_r$	MP$_{u,r}$	CF	APR	FR	FM	PITF	GIRPTM	BLL+MP$_r$	BLL$_{AC}$+MP$_r$
BibSonomy	-	F_1@5	.013	.074	.192	.166	.175	.171	.122	.139	.197	.201	**.202**
		MRR	.008	.054	.148	.133	.149	.148	.097	.120	.152	.158	**.159**
		MAP	.009	.070	.194	.173	.193	.194	.120	.150	.200	.207	**.209**
	3	F_1@5	.047	.313	.335	.325	.260	.337	.345	.356	.350	.353	**.358**
		MRR	.035	.283	.327	.289	.279	.333	.329	.341	.334	.349	**.350**
		MAP	.038	.345	.403	.356	.329	.414	.408	.421	.416	.435	**.439**
CiteULike	-	F_1@5	.002	.131	.253	.218	.195	.194	.111	.122	.263	.270***	**.271*****
		MRR	.001	.104	.229	.201	.233	.233	.110	.141	.246	.258***	**.259*****
		MAP	.001	.134	.280	.247	.284	.284	.125	.158	.301	.315***	**.317*****
	3	F_1@5	.013	.270	.316	.332	.313	.318	.254	.258	.336	.346***	**.351*****
		MRR	.012	.243	.353	.295	.361	.366	.282	.290	.380	.409***	**.415*****
		MAP	.012	.294	.420	.363	.429	.436	.326	.334	.455	.489***	**.497*****
Delicious	-	F_1@5	.033	.140	.236	.228	.211	.229	.157	.185	.253	.270***	**.274*****
		MRR	.025	.113	.214	.214	.206	.221	.141	.178	.236	.262***	**.267*****
		MAP	.026	.146	.257	.262	.246	.270	.168	.211	.286	.320***	**.327*****
	3	F_1@5	.058	.399	.355	.397	.290	.396	.394	.404	.370	.405***	**.417**$^{\circ\circ}$
		MRR	.041	.341	.330	.341	.284	.365	.361	.372	.329	.377***	**.392**$^{\circ\circ\circ}$
		MAP	.047	.443	.406	.441	.336	.466	.463	.478	.419	.483***	**.504**$^{\circ\circ\circ}$
Flickr	-	F_1@5	.023	-	.435	.417	.328	.334	.297	.316	.509	.523*	**.523***
		MRR	.023	-	.360	.436	.352	.355	.300	.333	.445	.466***	**.466*****
		MAP	.023	-	.468	.581	.453	.459	.384	.426	.590	.619***	**.619*****
	3	F_1@5	.026	-	.488	.493	.368	.378	.361	.369	.577	.592*	**.592***
		MRR	.026	-	.407	.498	.398	.404	.375	.390	.511	.533***	**.533*****
		MAP	.026	-	.527	.663	.513	.523	.481	.502	.676	.707***	**.707*****

Table 5: F_1@5, MRR and MAP values for BibSonomy, CiteULike, Delicious and Flickr (with no core and p-core = 3) showing that our BLL$_{AC}$+MP$_r$ approach outperforms state-of-the-art baseline algorithms (fourth research question). The symbols *, ** and *** indicate statistically significant differences based on a Wilcoxon Ranked Sum test between BLL+MP$_r$, BLL$_{AC}$+MP$_r$ and GIRPTM at α level .05, .01 and .001, respectively; $^{\circ}$, $^{\circ\circ}$ and $^{\circ\circ\circ}$ indicate statistically significant differences between BLL$_{AC}$+MP$_r$ and BLL+MP$_r$ at the same α levels.

that our tag recommender approach exceeds the performance of well-established and effective recommenders, such as MP$_{u,r}$, CF, APR, FM and even the currently leading time-dependent approach GIRPTM [64] (fourth research question). Finally, it is indispensable to highlight that BLL$_{AC}$+MP$_r$, despite its simplicity, appears to be even more successful than the sophisticated FR and PITF algorithms. Again, these results are further proved to be statistically significant based on a Wilcoxon Rank Sum test that is also shown in Table 5.

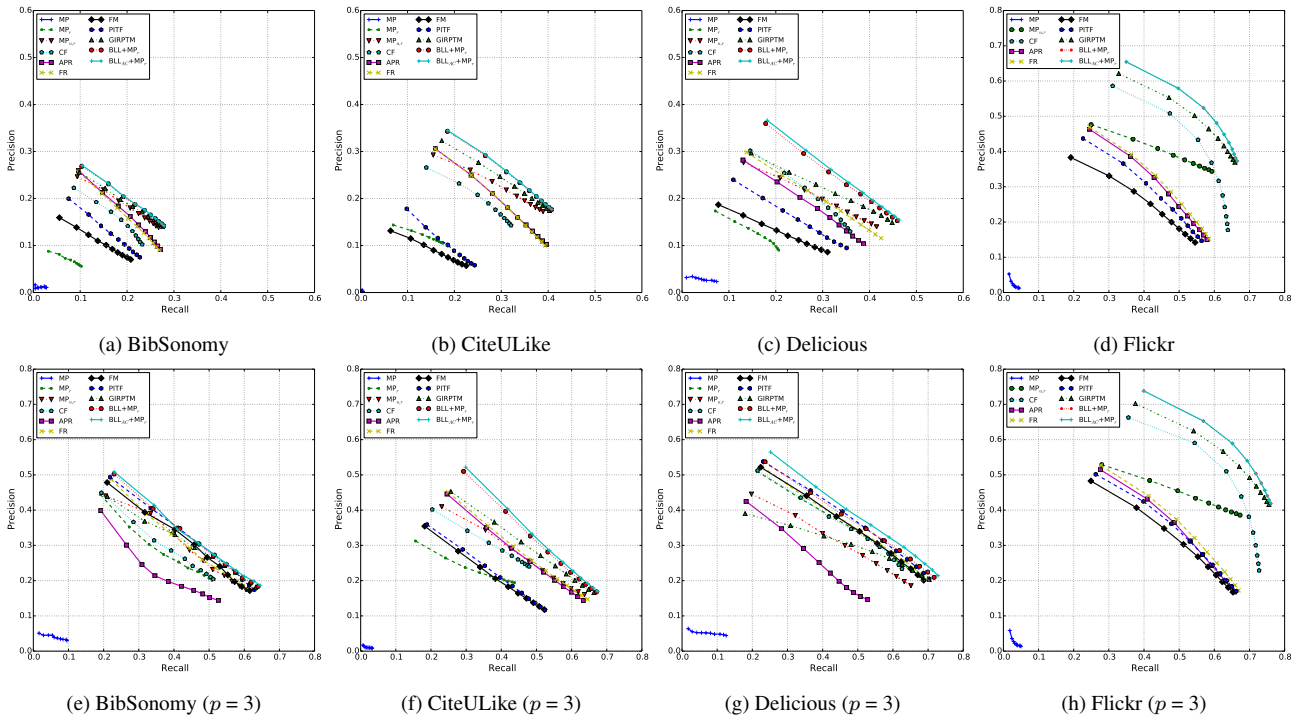
6.1.3 *Validation of the results in the ECML PKDD Discovery Challenge 2009 Dataset*

In order to increase the reproducibility of our results and to ensure that our results can be compared over different papers, we conducted another experiment on the well-known ECML PKDD discovery challenge 2009 dataset[16]. The dataset is an rather "old" snapshot (from 2009) of BibSonomy at p-core level 2 consisting of 64,406 bookmarks, 1,185 users, 22,389 resources, 13,276 tags and 253,615 tag assignments, but is used in many of the related work. Additionally, the dataset provides already a given train/test split, which further ensures the comparability of results. The winning algorithm based on the $F1@5$ evaluation metric in this tag recommender challenge was an optimized ensemble of factorization machines algorithms and was proposed by Rendle et al. [48]. In Table 6, we summarize the results presented in [48] together with the results of the novel time-dependent approaches of our work.

The $F1@5$ estimates indicate that the dataset and the splitting method is of advantage for resource-based approaches since MP$_r$ clearly outperforms MP$_u$. Interestingly, GIRP [64], reaches an even lower $F1@5$ score than MP$_u$ which also indicates that the information of time seems not to be important in this setting. However, BLL reaches a higher $F1@5$ score than MP$_u$ which again shows the advantage of its power decay function. Another indication of the importance of the current context in form of resource tags, is given by the very good results of our BLL$_{AC}$ approach which are similar to the results of APR. Although, BLL$_{AC}$ still recommends only tags already used by the given user, it adjusts the ranking using already assigned resource tags (i.e., the current context).

Our complete algorithm, BLL$_{AC}$+MP$_r$, reaches a $F1@5$ score of .308 and thus, again outperforms other sophisticated methods such as GIRPTM, CF, FR, FM and PITF. With regard to the final ECML PKDD discovery challenge 2009 ranking, this would result in the 8th position without any optimizations to the dataset or the length of the recommended tag list. Additionally, our algorithm is much more efficient in terms of computational complexity than the better performing approaches (especially the ones based on Factorization Machines, see also Table 7) and can be executed for this dataset on a single machine in a few seconds. Summed up, the results of this experiment show that our approach is capable

[16] http://www.kde.cs.uni-kassel.de/ws/dc09/



(a) BibSonomy (b) CiteULike (c) Delicious (d) Flickr

(e) BibSonomy (p = 3) (f) CiteULike (p = 3) (g) Delicious (p = 3) (h) Flickr (p = 3)

Figure 5: Recall/Precision plots for BibSonomy, CiteULike, Delicious and Flickr (with no core and p-core = 3) showing the performance of BLL+MP$_r$ and BLL$_{AC}$+MP$_r$ along with state-of-the-art baseline mechanisms for 1 - 10 recommended tags (k).

of providing high estimates of recommender accuracy in different settings without the need of dataset optimization or complex calculation steps.

6.2 Runtime

In addition to recommender accuracy, we investigated the runtime of our approaches both, in terms of computational complexity and monitored runtime. Table 7 shows the complexity of all algorithms in ascending order. We can see that the popularity-based algorithms MP$_u$, MP$_r$ and MP$_{u,r}$, that count frequencies by simply iterating over the tag assignments of the user (Y_u) and/or the resource (Y_r), provide linear runtime. For the time-based algorithms GIRP, GIRPTM, BLL and BLL+MP$_r$ we can observe similar behavior. An additional term is introduced, when calculating BLL$_{AC}$ and BLL$_{AC}$+MP$_r$. This term describes the initialization of the co-occurrence matrix that holds the semantic context. The matrix is built by iterating over each bookmark b in the set of bookmarks B of a folksonomy and checking the tag assignments of b (i.e., Y_b). Even though this calculation step increases the computational complexity of the approach, this step only needs to be performed once, which may be done offline (especially for big datasets) and subsequently, it may not effect the online runtime in a live system.

Moreover, we can see that BLL$_{AC}$ and BLL$_{AC}$+MP$_r$ show better performance than the other state-of-the-art methods such as CF, APR, FR, FM and PITF. As our theoretically motivated model allows us to rely on relatively little but meaningful operations considering only user tag frequency, recency and semantic context in terms of resource tags, our algorithm outperforms the former. CF on the other hand, processes not only the tag assignments Y_u of the target user, but additionally the tag assignments of each user v in the set of users (i.e., neighbors) that have tagged the target resource (V_r). In cases where there are no other users available that have tagged the target resource (i.e., cold-start resources), V_r becomes the set of all users which then could lead to much higher computational costs as expected (see our other runtime experiment in Figure 6 described in the next paragraph). With regard to APR/FR (depending on the number of nodes $|U|$, $|R|$ and $|T|$) and FM/PITF (depending on the dimensions of factorization k_U, k_R and k_T), even multiple iterations l are computed (see also Section 5.3), which leads to higher runtime complexities.

To furthermore proof the theoretical assumptions made in our complexity analysis, a real runtime experiment was carried out. In particular, we conducted an experiment on an IBM System x3550 server with two 2.0 GHz six-core Intel Xeon E5-2620 processors and 128 GB of RAM using Ubuntu 12.04.2 and Java 1.7 to determine the overall runtime performance[17] of the algorithms presented above. All algorithms were executed as single core single treat instance to ensure that the measured run-time is not affected by the implementation. The results of this evaluation (in milliseconds) can be found in Figure 6. As expected, the experiment proofs further evidence that the popularity-based approaches, such as MP$_u$, MP$_r$ and MP$_{u,r}$, the time-dependent approaches GIRP and GIRPTM and also our theory-based approaches perform significantly better than the more sophisticated graph-based approaches such as APR, FR, FM and PITF.

7. CONCLUSION

With this paper, we showed that it is worthwhile to analyze in more depth the human-computer interaction that is involved in the

[17]We report the overall runtime since it would not be fair to compare the live prediction time of a model-based approach, that precalculates the recommendations during the training phase as this is for instance the case with FM or PITF, against the live prediction time of a memory-based approach (e.g., MP$_u$, MP$_r$, MP$_{u,r}$, BLL, etc.).

Algorithm	$F1@5$
MP_u	.098
GIRP	.087
BLL	.104
MP_r	.288
$MP_{u,r}$	.290
CF	.295
APR	.231
FR	.290
GIRPTM	.248
FM	.296
PITF	.302
BLL_{AC}	.238
BLL_{AC}+MP_r	.308
Challenge winner [48]	.355

Table 6: $F1@5$ estimates for selected algorithms on the ECML PKDD Discovery Challenge 2009 dataset showing that our BLL_{AC}+MP_r is only outperformed by the winning algorithm (optimized ensemble of Factorization Machines [48]).

Algorithm	Complexity	Authors
MP	$\mathcal{O}(\lvert Y\rvert)$	Jäschke et al. [24]
MP_u	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert)$	Jäschke et al. [24]
GIRP	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert)$	Zhang et al. [64]
BLL	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert)$	Our approach
MP_r	$\mathcal{O}(\lvert R\rvert \cdot \lvert Y_r\rvert)$	Jäschke et al. [24]
$MP_{u,r}$	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert + \lvert R\rvert \cdot \lvert Y_r\rvert)$	Jäschke et al. [24]
GIRPTM	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert + \lvert R\rvert \cdot \lvert Y_r\rvert)$	Zhang et al. [64]
BLL+MP_r	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert + \lvert R\rvert \cdot \lvert Y_r\rvert)$	Our approach
BLL_{AC}	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert + \lvert B\rvert \cdot \lvert Y_b\rvert)$	Our approach
BLL_{AC}+MP_r	$\mathcal{O}(\lvert U\rvert \cdot \lvert Y_u\rvert + \lvert B\rvert \cdot \lvert Y_b\rvert + \lvert R\rvert \cdot \lvert Y_r\rvert)$	Our approach
CF	$\mathcal{O}(\lvert U\rvert \cdot \lvert V_r\rvert \cdot \lvert Y_v\rvert)$	Marinho et al. [39]
APR	$\mathcal{O}(\lvert U\rvert \cdot l \cdot (\lvert Y\rvert + \lvert U\rvert + \lvert R\rvert + \lvert T\rvert))$	Hotho et al. [21]
FR	$\mathcal{O}(\lvert U\rvert \cdot l \cdot (\lvert Y\rvert + \lvert U\rvert + \lvert R\rvert + \lvert T\rvert))$	Hotho et al. [21]
FM	$\mathcal{O}(l \cdot \lvert B\rvert \cdot (k_T \cdot \lvert T\rvert^2 + k_U \cdot k_R \cdot k_T))$	Rendle et al. [47]
PITF	$\mathcal{O}(l \cdot \lvert B\rvert \cdot (k_T \cdot \lvert T\rvert^2 + k_U \cdot k_R \cdot k_T))$	Rendle et al. [47]

Table 7: Computational complexity of BLL, BLL_{AC}, BLL+MP_r and BLL_{AC}+MP_r compared to state-of-the-art baseline algorithms in ascending order showing that our approaches provide a better runtime complexity than CF, APR, FR, FM and PITF.

generation and exploitation of the data which we would like to use to make recommendations. This involved designing an algorithm that is optimally tuned to the statistical structure of the data. In this particular case, we used a theory of human long-term memory to devise a model that predicts the reuse probability of a tag in social tagging, much in the same way as the human memory system makes use of memory traces for current tasks.

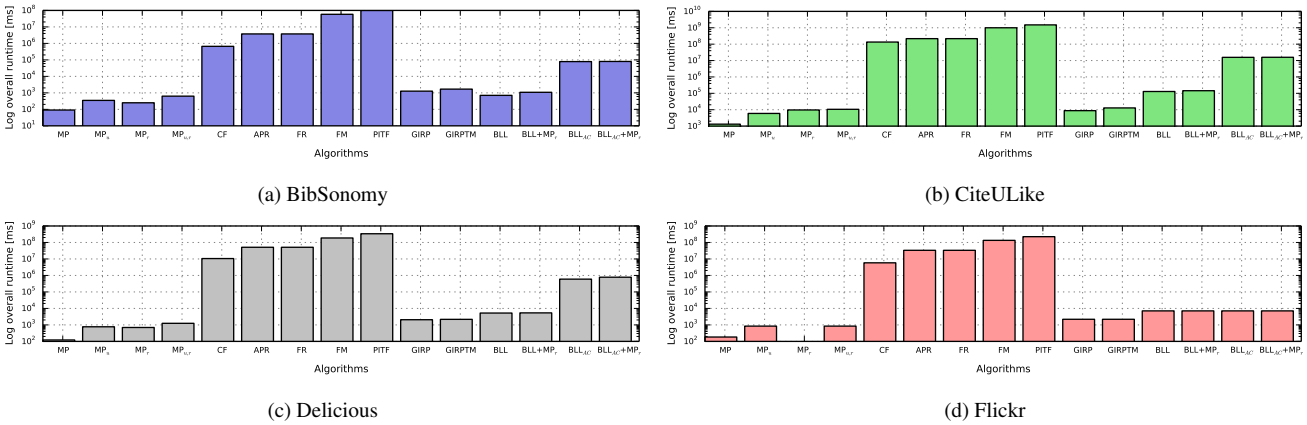
The first research question of this work dealt with the question whether an exponential or a power decay function is more appropriate to account for the effect of recency on a tag's reuse probability. In order to examine this question we performed an empirical analysis on four social tagging datasets (BibSonomy, CiteULike, Delicious and Flickr). The analysis showed that the effect of recency on the reuse probability of tags follows a power law distribution. This encourages the application of the BLL equation by Anderson et al. [2] as it models a user's temporal tagging pattern in form of a power forgetting function.

In order to tackle our research questions 2 - 4, we followed a three-step recommender evaluation strategy. We started by comparing the performance of BLL with MP_u to determine the effect of considering the recency of each tag use. Results of an additional comparison may differentiate our cognitive-psychological model from the less theory-driven GIRP approach introduced by Zhang et al. [64]. Our findings, tackling the second research question, clearly demonstrate that regardless of the evaluation metric and across all datasets, BLL reaches higher levels of accuracy than MP_u and even outperforms GIRP. Thus, processing the recency of tag use is effective to account for additional variance of users' tagging behavior and therefore, a reasonable extension of simple "most popular tags" approaches. Furthermore, the significant advantage over GIRP indicates that drawing on memory psychology guides the application of a reliable and valid model built upon long-standing, empirical research. The equations that Zhang et al. [64] used to implement their approach were developed from scratch rather than derived from existing research described above. As a consequence, [64] models the recency of tag use by means of an exponential function, which is clearly at odds with the power law of forgetting described in related work (e.g., [4]).

In a next step, we have extended BLL to BLL_{AC} using current context information based on the activation equation of Anderson et al. [2]. Where BLL gives the prior probability of tag reuse that is learned over time, the associative component tunes this prior probability to the current context by exploiting the current semantic cues from the environment (i.e., the previously assigned tags of the target resource). This is in line with how ACT-R models the retrieval from long-term memory. Our results show that this step significantly improves the "pure" BLL equation, especially in case of the p-core pruned datasets, where more context information (i.e, tag assignments of the target resource) are available to calculate the associative component (third research question).

In the last step, we combined BLL_{AC} with the frequency estimates of the most popular tags that have been applied by other users to the target resource in the past (i.e., MP_r) in order to be able to also recommend new tags, i.e., tags that have not been used by the target user before. Despite their simplicity and computational efficiency, our results show that this combination (BLL_{AC}+MP_r) significantly outperforms well-established mechanisms, such as CF, FR, PITF and GIRPTM, in terms of recommender accuracy and runtime. We assume this is the case because, in following some fundamental principles of human memory, our approaches are better adapted to the statistical structure of the environment (fourth research question). Moreover, the results of this experiment also show that there is only a small difference between BLL_{AC}+MP_r and BLL+MP_r (without contextual activation of the user tags), which suggests the use of additional context information, such as content-based features (e.g., the resource's title or body-text). This would also be in line with the studies of Marek et al. [34, 35, 36], who showed that the resource title has a big impact on tags in collaborative tagging systems and so could be a better alternative to represent context cues than the popular tags of the resource used in the current work.

Finally, a glance on the results shows an interdependency between the examined dataset and the performance of our approaches. While the distance to other strongly performing mechanisms is not large in case of broad folksonomies (BibSonomy, CiteULike and Delicious), this distance grows substantially larger in a narrow folksonomy (Flickr), where no tags of other users are available for the target user's resources. From this interdependency we conclude that applying a model of human memory is particularly effective if



(a) BibSonomy

(b) CiteULike

(c) Delicious

(d) Flickr

Figure 6: Overall runtime in milliseconds [ms] of BLL, BLL$_{AC}$, BLL+MP$_r$ and BLL$_{AC}$+MP$_r$ compared to state-of-the-art baseline algorithms for BibSonomy, CiteULike, Delicious and Flickr showing the full time to process the whole dataset samples (training and testing).

tag assignments are mainly driven by individual habits unaffected by the behavior of other users, such as it is done in Flickr.

7.1 Future Work

In future work, we will continue examining memory processes that are involved in categorizing and tagging Web resources. For instance, in a recent study [51], we introduced a mechanism by which memory processes involved in tagging can be modeled on two levels of knowledge representation: on a semantic level (representing categories or LDA topics) and on a verbal level (representing tags). Next, we will aim at combining this integrative mechanism with the activation equation to examine a potential correlation between the impact of recency (time-based forgetting) and the level of knowledge representation. We believe that conclusions drawn from cognitive science will help to develop an effective and psychologically plausible tag recommendation mechanism. We also plan to integrate our approach into an actual tagging system application. This will provide us with a real-life setting to test user acceptance. Furthermore, we are interested in extending our approach into the domain of content-based tag recommender systems, i.e., exploring additional context features such as title or body-text. Also, we want to test the activation equation in the context of collaborative item recommender systems, using tag and time information as input. We consider this promising, as preliminary experiments suggest [31], that the activation equation bears also a great potential to rank items efficiently. Finally, we are interested in investigating the impact of different tagging styles on tag recommender systems as suggested by [27] and the study of individual learning curves in the tag recommendation process as suggested by [25] for item recommendations.

8. REPRODUCIBILITY

Please note, that the source-code of all approaches introduced in this paper are implemented in our open-source tag-recommender framework *TagRec* [28, 55], which can be downloaded online for free from our GitHub repository[18]. Furthermore, we provide open-access to all data samples via e-mail request to ensure reproducibility of the methods described in our work.

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

9. ACKNOWLEDGMENTS

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" fellowship program by the first author. The work is supported by the Know-Center, the EU funded projects Learning Layers (Grant Agreement 318209) and weSPOT (Grant Agreement 318499) and the Austrian Science Fund (FWF): P 25593-G22. The Learning Layers and weSPOT projects are supported by the European Commission within the 7th Framework Program. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - 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. COMET is managed by the Austrian Research Promotion Agency (FFG).

APPENDIX

A. VALIDATION OF DATA SAMPLING

Table 8 reports on the validity of our data sampling strategy as presented in Section 5.1. Due to the computational complexity of some algorithms (e.g., PITF), we focused on randomly selecting 3% of the user profiles for the significant larger datasets Flickr and Delicious. In order to check whether this sampling did introduce any unwanted bias, we drew 5 additional random samples and repeated the estimation of of parameters for all algorithms. As highlighted, per algorithm, the table reveals only slight variance across the five samples for each of the four datasets (two for Delicious and two for Flickr). Test for statistical significance required a non-parametric method due to a significant violation of the normal distribution assumption. Therefore, to compare the five samples with respect to the $F1@5$ metric, we performed a Kruskal-Wallis test by ranks for every dataset, each yielding a non-significant effect for the sample:

- Delicious (no core): H(4) = 0.86, p = .99.
- Delicious (p = 3): H(4)=0.15, p = .99.
- Flickr (no core): H(4) = 2.94, p = .57.
- Flickr (p = 3): H(4) = 1.44, p = .84.

	Delicious		Flickr	
Algorithm	No core	p = 3	No core	p = 3
MP_u	.175 / .165 / .169 / .168 / **.170**	.187 / .186 / .190 / .189 / **.193**	.432 / .443 / .439 / .438 / **.435**	.487 / .502 / .495 / .491 / **.488**
MP_r	.144 / .140 / .139 / .139 / **.140**	.402 / .400 / .398 / .402 / **.399**	-	-
$MP_{u,r}$	.238 / .230 / .233 / .233 / **.236**	.351 / .353 / .354 / .353 / **.355**	.432 / .443 / .439 / .438 / **.435**	.487 / .502 / .495 / .491 / **.488**
FR	.226 / .227 / .228 / .226 / **.229**	.393 / .394 / .393 / .397 / **.396**	.334 / .338 / .340 / .340 / **.334**	.376 / .383 / .384 / .383 / **.378**
GIRPTM	.258 / .250 / .253 / .252 / **.253**	.366 / .367 / .366 / .371 / **.370**	.506 / .517 / .512 / .508 / **.509**	.573 / .589 / .581 / .574 / **.577**
BLL_{AC}+MP_r	.278 / .269 / .273 / .272 / **.274**	.412 / .414 / .414 / .417 / **.417**	.519 / .532 / .526 / .520 / **.523**	.586 / .604 / .596 / .586 / **.592**

Table 8: $F1@5$ estimates for MP_u, MP_r, $MP_{u,r}$, FR, GIRPTM and BLL_{AC}+MP_r on 5 samples (i.e., 3% of randomly chosen users, see Section 5.1) of Delicious and Flickr (no core and p = 3). The results show very similar estimates among all five samples, which validates our chosen sampling strategy. The bold values (i.e., the fifth sample) are the reported ones in the rest of the paper.

B. REFERENCES

[1] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS One*, 9(1):e85777, 2014.

[2] J. R. Anderson, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1050, 2004.

[3] J. R. Anderson, J. M. Fincham, and S. Douglass. Practice and retention: a unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:1120, 1999.

[4] J. R. Anderson and L. J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

[5] L. Balby Marinho, A. Hotho, R. Jäschke, A. Nanopoulos, S. Rendle, L. Schmidt-Thieme, G. Stumme, and P. Symeonidis. *Recommender Systems for Social Tagging Systems*. SpringerBriefs in Electrical and Computer Engineering. Springer, Feb. 2012.

[6] V. Batagelj and M. Zaveršnik. Generalized cores. *arXiv preprint cs/0202039*, 2002.

[7] P. G. Campos, F. Díez, and I. Cantador. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *UMUAI*, pages 1–53, 2013.

[8] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[9] K. Dellschaft and S. Staab. Measuring the influence of tag recommenders on the indexing quality in tagging systems. In *Proc. of HT '12*, pages 73–82, New York, NY, USA, 2012. ACM.

[10] S. Doerfel and R. Jäschke. An analysis of tag-recommender evaluation procedures. In *Proc. of RecSys '13*, pages 343–346, New York, NY, USA, 2013. ACM.

[11] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Number 3. Teachers college, Columbia university, 1913.

[12] F. Floeck, J. Putzke, S. Steinfels, K. Fischbach, and D. Schoder. Imitation and quality of tags in social bookmarking systems–collective intelligence leading to folksonomies. In *On collective intelligence*, pages 75–91. Springer, 2011.

[13] W.-T. Fu. The microstructures of social tagging: a rational model. In *Proc. of CSCW '08*, pages 229–238. ACM, 2008.

[14] W.-T. Fu, T. Kannampallil, R. Kang, and J. He. Semantic imitation in social tagging. *ACM Trans. Comput.-Hum. Interact.*, 17(3):12:1–12:37, July 2010.

[15] W.-T. Fu, T. G. Kannampallil, and R. Kang. A semantic imitation model of social tag choices. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 66–73. IEEE, 2009.

[16] J. Gemmell, T. Schimoler, M. Ramezani, L. Christiansen, and B. Mobasher. Improving folkrank with item-based collaborative filtering. *Recommender Systems & the Social Web*, 2009.

[17] S. Hamouda and N. Wanas. Put-tag: personalized user-centric tag recommendation for social bookmarking systems. *Social network analysis and mining*, 1(4):377–385, 2011.

[18] D. Helic, C. Körner, M. Granitzer, M. Strohmaier, and C. Trattner. Navigational efficiency of broad vs. narrow folksonomies. In *Proc. of HT '12*, pages 63–72, New York, NY, USA, 2012. ACM.

[19] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. of WSDM '08*, pages 195–206, New York, NY, USA, 2008. ACM.

[20] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proc. of SigIR '08*, pages 531–538. ACM, 2008.

[21] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The semantic web: research and applications*, pages 411–426. Springer, 2006.

[22] C.-L. Huang, P.-H. Yeh, C.-W. Lin, and D.-C. Wu. Utilizing user tag-based interests in recommender systems for social resource sharing websites. *Knowledge-Based Systems*, 2014.

[23] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Proc. of PKDD'07*, pages 506–514. Springer, 2007.

[24] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in social bookmarking systems. *Ai Communications*, 21(4):231–247, 2008.

[25] Y. Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[26] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proc. of WWW '10*, pages 521–530, New York, NY, USA, 2010. ACM.

[27] C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier. Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proc. of HT '10*, pages 157–166, New York, NY, USA, 2010. ACM.

[28] D. Kowald, E. Lacic, and C. Trattner. Tagrec: Towards a standardized tag recommender benchmarking framework. In *Proc. of HT'14*, New York, NY, USA, 2014. ACM.

[29] D. Kowald, P. Seitlinger, C. Trattner, and T. Ley. Long time no see: The probability of reusing tags as a function of

frequency and recency. In *Proc. of WWW '14*, New York, NY, USA, 2014. ACM.

[30] R. Krestel and P. Fankhauser. Language models and topic models for personalizing tag recommendation. In *Proc. of WIIAT'10*, volume 1, pages 82–89. IEEE, 2010.

[31] E. Lacic, D. Kowald, P. Seitlinger, C. Trattner, and D. Parra. Recommending items in social tagging systems using tag and time information. *arXiv preprint arXiv:1406.7727*, 2014.

[32] Y.-l. Lin, C. Trattner, P. Brusilovsky, and D. He. The impact of image descriptions on user tagging behavior: A study of the nature and functionality of crowdsourced tags. *Journal of the Association for Information Science and Technology*, pages n/a–n/a, 2014.

[33] M. Lipczak. *Hybrid Tag Recommendation in Collaborative Tagging Systems*. PhD thesis, Dalhousie University, 2012.

[34] M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. *ECML PKDD discovery challenge*, pages 157–172, 2009.

[35] M. Lipczak and E. Milios. The impact of resource title on tags in collaborative tagging systems. In *Proc. of HT '10*, pages 179–188, New York, NY, USA, 2010. ACM.

[36] M. Lipczak and E. Milios. Efficient tag recommendation for real-life data. *ACM Trans. Intell. Syst. Technol.*, 3(1):2:1–2:21, Oct. 2011.

[37] J. Lorince and P. M. Todd. Can simple social copying heuristics explain tag popularity in a collaborative tagging system? In *Proc. of WebSci '13*, pages 215–224, New York, NY, USA, 2013. ACM.

[38] L. Marinho, A. Nanopoulos, L. Schmidt-Thieme, R. Jäschke, A. Hotho, G. Stumme, and P. Symeonidis. Social tagging recommender systems. In *Recommender Systems Handbook*, pages 615–644. Springer US, 2011.

[39] L. B. Marinho and L. Schmidt-Thieme. Collaborative tag recommendations. In *Data Analysis, Machine Learning and Applications*, pages 533–540. Springer, 2008.

[40] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proc. of RecSys '13*, New York, NY, USA, 2013. ACM.

[41] C. Moltedo, H. Astudillo, and M. Mendoza. Tagging tagged images: on the impact of existing annotations on image tagging. In *Proc. of ACMMM '2012 workshop on Crowdsourcing for multimedia*, pages 3–8. ACM, 2012.

[42] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[43] D. Parra and P. Brusilovsky. Collaborative filtering for social tagging systems: an experiment with citeulike. In *Proc. of RecSys '09*, pages 237–240. ACM, 2009.

[44] D. Parra-Santander and P. Brusilovsky. Improving collaborative filtering in social tagging systems for the recommendation of scientific articles. In *Proc. of WIIAT'10*, volume 1, pages 136–142. IEEE, 2010.

[45] M. Rawashdeh, H.-N. Kim, J. M. Alja'am, and A. El Saddik. Folksonomy link prediction based on a tripartite graph for tag recommendation. *Journal of Intelligent Information Systems*, pages 1–19, 2012.

[46] S. Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[47] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proc. of SigKDD '09*, pages 727–736. ACM, 2009.

[48] S. Rendle and L. Schmidt-Thieme. Factor models for tag recommendation in bibsonomy. In *ECML/PKDD 2008 Discovery Challenge Workshop*, pages 235–243, 2009.

[49] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proc. of WSDM '10*, pages 81–90, New York, NY, USA, 2010. ACM.

[50] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[51] P. Seitlinger, D. Kowald, C. Trattner, and T. Ley. Recommending tags with a model of human categorization. In *Proc. of CIKM '13*, New York, NY, USA, 2013. ACM.

[52] P. Seitlinger and T. Ley. Implicit imitation in social tagging: familiarity and semantic reconstruction. In *Proc. of CHI '12*, pages 1631–1640, New York, NY, USA, 2012. ACM.

[53] P. Seitlinger, T. Ley, and D. Albert. Verbatim and semantic imitation in indexing resources on the web: A fuzzy-trace account of social tagging. *Applied Cognitive Psychology*, 2014.

[54] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of WWW '08*, pages 327–336. ACM, 2008.

[55] C. Trattner, D. Kowald, and E. Lacic. Tagrec: Towards a toolkit for reproducible evaluation and development of tag-based recommender algorithms. *SIGWEB Newsl.*, (Winter):3:1–3:10, Feb. 2015.

[56] C. Trattner, Y.-l. Lin, D. Parra, Z. Yue, W. Real, and P. Brusilovsky. Evaluating tag-based information access in image collections. In *Proc. of HT '12*, pages 113–122, New York, NY, USA, 2012. ACM.

[57] K. H. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proc. of SAC'08*, pages 1995–1999. ACM, 2008.

[58] L. Van Maanen and J. N. Marewski. Recommender systems for literature selection: A competition between decision making and memory models. In *Proc. of CogSci '09*, pages 2914–2919, 2009.

[59] C. J. Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

[60] R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak. I tag, you tag: translating tags for advanced user models. In *Proc. of WSDM '10*, pages 71–80. ACM, 2010.

[61] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proc. of SigIR '08*, pages 155–162. ACM, 2008.

[62] D. Yin, L. Hong, and B. D. Davison. Exploiting session-like behaviors in tag prediction. In *Proc. of WWW '11 companion*, pages 167–168. ACM, 2011.

[63] D. Yin, L. Hong, Z. Xue, and B. D. Davison. Temporal dynamics of user interests in tagging systems. In *Twenty-Fifth AAAI conference on artificial intelligence*, 2011.

[64] L. Zhang, J. Tang, and M. Zhang. Integrating temporal usage pattern into personalized tag prediction. In *Web Technologies and Applications*, pages 354–365. Springer, 2012.

[65] N. Zheng and Q. Li. A recommender system based on tag and time information for social tagging systems. *Expert Syst. Appl.*, 2011.