Personalized Topic-Based Tag Recommendation

Ralf Krestel[a,*], Peter Fankhauser[b]

[a]*L3S Research Center, Hannover, Germany*
[b]*DFKI, Saarbrücken, Germany*

Abstract

More and more content on the Web is generated by users. To organize this information and make it accessible via current search technology, tagging systems have gained tremendous popularity. Especially for multimedia content they allow to annotate resources with keywords (tags) which opens the door for classic text-based information retrieval. To support the user in choosing the right keywords, tag recommendation algorithms have emerged. In this setting, not only the content is decisive for recommending relevant tags but also the user's preferences.

In this paper we introduce an approach to personalized tag recommendation that combines a probabilistic model of tags from the resource with tags from the user. As models we investigate simple language models as well as Latent Dirichlet Allocation. Extensive experiments on a real world dataset crawled from a big tagging system show that personalization improves tag recommendation, and our approach significantly outperforms state-of-the-art approaches.

Keywords: Tag Recommendation, Personalization, Language Models, Topic Models

1. Introduction

The World Wide Web is growing at incredible speed. User generated content is uploaded by millions everyday. Web 2.0 or the Social Web are evolving rapidly. Sharing of user generated content is one of the predominant actions on the Web nowadays. To organize this content and to make it accessible to other users is the main purpose of sites like Flickr[1], LastFm[2], YouTube[3], or Delicious[4].

These sites allow users to annotate content with their own keywords (tags), opening up the possibility to retrieve content using traditional keyword search. Especially multimedia content like music, photos, or videos rely on manually added meta information. Adding keywords to content (tagging) is the only feasible way to organize multimedia data at that scale and to make it searchable. These keywords can be freely chosen by a user and are not restricted to any taxonomy. This results in some benefits like flexibility, quick adaption, and easy usability, but has also some drawbacks.

Tagging is considered a categorization process not a classification process, see Halpin et al. (2007). The underlying meaning has to be evaluated and inferred in the context of other tags and user information. Tags can even have no concrete meaning or are only interpretable by the user herself. In addition, tags can have various purposes. Some describe the annotated content, some refer to the user (e.g. "jazz", "myHolidays" or "to_read") as described in Bischoff et al. (2008). In practice allowing users to freely annotate means that tagging systems contain noise and are rather sparsely populated.

Studies by Golder and Huberman (2005) and Bollen and Halpin (2009) have shown that many users annotating a resource leads to a stable tag distribution for this resource, capturing its characteristics sufficiently. To support users in choosing tags, tag recommendation algorithms have emerged. For resources already annotated by lots of people this recommendation is rather straight forward. The tagging system can provide the most frequent tags assigned to the resource, or look at the tagging history of the user to make a more personalized recommendation. On this note, tag recommendation algorithms can be classified into user-centered and resource-centered ones, see Song et al. (2011).

In this paper we combine both perspectives to recommend personalized tags to users. To this end, we employ a mixture (Section 3.4) of simple language models (LM) (Section 3.2) and Latent Dirichlet Allocation (LDA) (Section 3.3) to estimate the probability of new tags based on the already assigned tags of a resource and a user, and introduce a principled approach for combining these estimates in Section 3.1. The potential advantage of employing LDA is the possibility to recommend tags not previously assigned to the resource or used by the user. This broadens the available vocabulary for tag recommendation. The potential advantage of combining the resource perspective with the user perspective is to filter general tags for a resource with the individual tagging preferences of a user.

In Section 4 we evaluate our approach on two real world datasets. We systematically analyze tag recommendation based on resources or users only, assess the possible merits of LDA

*Corresponding Author. Appelstr. 4, 30167 Hannover, Germany. Phone: +49-511-762-17704

Email address: krestel@l3s.de (Ralf Krestel)

[1]Flickr: http://www.flickr.com
[2]LastFm: http://www.last.fm
[3]YouTube: http://www.youtube.com
[4]Delicious: http://delicious.com

as opposed to language models, and compare our combined approach to FolkRank by Hotho et al. (2006) as a state-of-the-art personalized tag-recommender. Our evaluation shows that combining evidence from the resource and the user improves tag recommendation significantly, and that LDA helps, in particular for generalizing from individual tagging practices on resources. Moreover, our approach achieves significantly better accuracy than state-of-the-art approaches.

2. Related Work

In recent years interest in tag recommendation was sparked within the research community. The growing importance of tagging systems led to the development of sophisticated tag recommendation algorithms. The various approaches applied for the Data Discovery Challenge 2009 (Eisterlehner et al. (2009)) represent a good overview.

2.1. Collaborative Filtering

A popular approach to tag recommendation has been collaborative filtering (Herlocker et al. (2004)), taking into account similarities between users, resources, and tags.

Mishne (2006) introduces an approach to recommend tags for weblogs, based on similar weblogs tagged by the same user. Chirita et al. (2007) realize this idea for the personal desktop, recommending tags for web resources by retrieving and ranking tags from similar documents on the desktop.

Jäschke et al. (2007) compare two variants of collaborative filtering and FolkRank (Hotho et al. (2006)), a graph based algorithm for recommendations in folksonomies. For collaborative filtering, once the similarity between users on tags, and once the similarity between users on resources is used for recommendation. FolkRank uses random walk techniques on the user-resource-tag (URT) graph based on the idea that popular users, resources, and tags can reinforce each other. These algorithms take co-occurrence of tags into account only indirectly, via the URT graph. Our evaluation shows that our approach achieves significantly better accuracy than FolkRank, and even the simple and scalable combination of smoothed language models achieves competitive accuracy.

Xu et al. (2006) describe a way to recommend a few descriptive tags to users by rewarding co-occuring tags that have been assigned by the same user, penalizing co-occuring tags that have been assigned by different users, and boosting tags with high descriptiveness. An interactive approach in the context of a photo tagging site based on co-occurence is presented in Garg and Weber (2008). After the user enters a tag for a new resource, the algorithm recommends tags based on co-occurence of tags for resources which the user or others used together in the past. After each tag the user assigns or selects, the set is narrowed down to make the tags more specific. Sigurbjörnsson and van Zwol (2008) also look at co-occurence of tags to recommend tags based on a user defined set of tags. The co-occuring tags are then ranked and promoted based on e.g. descriptiveness.

Heymann et al. (2008) employ association rule mining on the tag sets of resources for collective tag recommendation. The mined association rules have the form $T_1 \rightarrow T_2$, where T_1 and T_2 are tag sets. On this basis tags in T_2 are recommended, when all tags in T_1 are available for the resource, and the confidence for the association rules is above a threshold. In Krestel et al. (2009) we have shown that tag recommendation based on LDA achieves significantly better accuracy than this approach, and recommends more specific tags, which are more useful for tag-based search.

Wetzker et al. (2010) introduce an approach for personalized tag recommendation based on tensor calculus. Their approach is similar to the approach based on language models presented in this paper, but differs with respect to normalization of tag weights and the way, the resource perspective is taken into account. By using a more principled probabilistic approach for combining the resource perspective with the user perspective, the approach in this paper can benefit more readily from better estimates of tag probabilities, based, e.g., on Latent Dirichlet Allocation.

2.2. Clustering

A general problem of tagging systems is their sparsity. This has lead to a number of approaches using clustering in order to map the sparse tagging space to fewer dimensions.

Symeonidis et al. (2008) employ dimensionality reduction to personalized tag recommendation. Whereas Jäschke et al. (2007) operate on the URT graph directly, Symeonidis et al. (2008) use generalized techniques of SVD (Singular Value Decomposition) for n-dimensional tensors. The 3-dimensional tensor corresponding to the URT graph is unfolded into 3 matrices, which are reduced by means of SVD individually, and combined again to arrive at a more dense URT tensor approximating the original graph. The algorithm then suggests tags to users, if their weight is above some threshold. Rendle and Schmidt-Thieme (2010) introduce two more efficient variants of this approach using canonical decomposition and pairwise interaction tensor factorization. These tensor based techniques can be readily compared to the approach presented in this paper. The user-tag and resourcetag perspectives combined in this paper correspond to 2 of the 3 matrices, and the LDA can be seen as an alternative dimensionality reduction technique. Indeed, on the bibsonomy dataset (see Section 4), both approaches achieve similar accuracy.

When content of resources is available, tag recommendation can also be approached as a classification problem, predicting tags from content. A recent approach in this direction is presented in Song et al. (2008). They cluster the document-term-tag matrix after an approximate dimensionality reduction, and obtain a ranked membership of tags to clusters. Tags for new resources are recommended by classifying the resources into clusters, and ranking the cluster tags accordingly.

2.3. LDA for Tag Recommendation

Latent Dirichlet Allocation, a variant of clustering in particular suitable for bag of words data, has recently gained some attention for tag recommendation. Si and Sun (2009) and Krestel et al. (2009); Krestel and Fankhauser (2009) introduce an

approach to *collective* tag recommendation using LDA. Xiance et al. employ LDA for eliciting topics from the words in documents (blogposts) and from the associated tags, where words and tags form disjoint vocabularies. On this basis they recommend new tags for new documents using their content only. Krestel et al. on the other hand use LDA to infer topics from the available tags of resources and then recommend additional tags from these latent topics. In this paper we extend these approaches for *personalized* tag recommendation by also taking the personal tagging practices of users into account. Moreover, we show that using a mixture of language models and latent topic models significantly improves the accuracy of tag recommendation.

Bundschus et al. (2009) introduce a combination of LDA based on the content and tags of resources and the users having bookmarked a resource. The underlying generative process elicits *user specific* latent topics from the resource content and seperately from the tags of the resource. The content-based topics and tag-based topics are in a one-to-one correspondence by the user-id. On this basis personalized tag recommendation is realized by first eliciting user specific topics from the resource content, and then using the corresponding tag-based topics for suggesting tags. Our approach does not require content, which may not be available, e.g., for multimedia data, but works exlusively on the tags.

Harvey et al. (2010) introduce a similar approach to personalized tag recommendation as proposed in this paper on the basis of LDA. Rather than decomposing the joint probability of a *tag* given the tag assignments for a resource and a user via an application of Bayes' rule (see Equation 5), they decompose the joint probability of *latent topics* given the tag assignments. On this basis, they introduce an extended Gibbs sampler which draws topics simultaneously from the user and the resource. This fully generative approach, however, requires some initial tags from the user to a given resource, in order to recommend additional tags. In contrast, our approach can also handle the arguably more realistic setting of suggesting tags for a new bookmark without any initial tags from the user.

3. Personalized Tag Recommendation mixing Language Models with Topic Models

In this section we present our approach to combine tag recommendation for users with tag recommendation for resources. We show that this combination helps to overcome the weaknesses of the individual approaches applied in isolation. On the one hand we take the user's interest and tagging preferences into account, and on the other hand we identify suitable tags for a particular resource. For both, user-centered and resource-centered, we investigate the use of two methods, Latent Dirichlet Allocation and language models for tag recommendation.

3.1. Goals and Approach

Tagging systems allow users to annotate resources with keywords. Tag recommendation aims at assistint the user with this task. As soon as the user decides to tag a new resource, the system suggests appropriate tags to alleviate the burden of coming up with new keywords and typing them for the user. The tag recommendation algorithm used in a system therefore also influences the tag distribution of resources since many users pick a recommended tag rather then conceiving new keywords. Hence, the recommendation algorithm is an important part of tagging systems. We now give a more formal definition of this task.

Given a set of resources R, tags T, and users U, the ternary relation $X \subseteq R \times T \times U$ represents the user specific assignment of tags to resources. A bookmark $b(r, u)$ for a resource $r \in R$ and a user $u \in U$ comprises all tags assigned by u to r: $b(r, u) = \pi_t \sigma_{r,u} X$[5]. The goal of personalized tag recommendation is to assist users bookmarking a new resource by reducing the cognitive load by suggesting tags for their bookmark $b(r, u)$. This can be based on other tag assignments to this resource and similar resources, or based on the user and similar users.

To this end, we need to rank possible tags t, given a resource and a user. We rank based on a probabilistic approach. More formally, we estimate the probability $P(t|u, r)$ of a tag t given a resource r and a user u as follows:

$$P(t \mid r, u) = \frac{P(r, u \mid t)P(t)}{P(r, u)} \quad (1)$$

$$\approx \frac{P(r \mid t)P(u \mid t)P(t)}{P(r, u)} \quad (2)$$

$$= \frac{P(t \mid r)P(r)}{P(t)} \frac{P(t \mid u)P(u)}{P(t)} \frac{P(t)}{P(r, u)} \quad (3)$$

$$= \frac{P(t \mid r)P(t \mid u)}{P(t)} \frac{P(r)P(u)}{P(r, u)} \quad (4)$$

$$\propto \frac{P(t \mid r)P(t \mid u)}{P(t)} \quad (5)$$

Equation 1 applies Bayes' rule, Equation 2 splits $P(r, u|t)$ assuming conditional independence of r and u given t, Equation 3 again applies Bayes' rule to $P(r|t)$ and $P(u|t)$, Equation 4 simplifies, and Equation 5 discards the factors $P(r)$, $P(u)$, and $P(r, u)$, which are equal for all tags.

$P(t)$ can be estimated via the relative frequency of tag t in all bookmarks. For estimating $P(t|r)$ and $P(t|u)$ we investigate and combine two approaches. On the one hand, we use simple language models (Section 3.2), on the other hand, we use Latent Dirichlet Allocation (Section 3.3), in order to also recommend tags for new resources and users, which have only few bookmarks available.

The estimate in Equation 5 gives equal weight to $P(t|r)$ and $P(t|u)$. However, typically there are more tags available for a particular user u than for a resource r. Thus the estimate for $P(t|u)$ should be weighted more strongly than the estimate for $P(t|r)$. To this end, we smoothen $P(t|r)$ and $P(t|u)$ with the prior probability $P(t)$.

$$P'(t \mid r) \propto log_2(|r| + 1)P(t \mid r) + log_2(|u| + 1)P(t) \quad (6)$$

$$P'(t \mid u) \propto log_2(|u| + 1)P(t \mid u) + log_2(|r| + 1)P(t) \quad (7)$$

[5] projection π and selection σ operate on multisets without removing duplicate tuples

where $|r|$ is the number of tags available for a resource r, and $|u|$ is the number of tags available for a user u. When $|r|$ is smaller than $|u|$, $P(t|r)$ is smoothed more strongly, and thus influences $P(t|r,u)$ less than $P(t|u)$. Note that when $P(t|r)$ is zero, $P'(t|r)$ is proportional to $P(t)$. Consequently, the combined probability $P'(t|r) * P'(t|u)/P(t)$ is effectively proportional to $P'(t|u)$. Likewise, when a resource has no tags at all, $log_2(|r| + 1) = 0$, and the combined probability is again proportional to $P'(t|u)$.

The combination above is reminiscent of the popular "Product of Experts" approach, where our Experts are resources and users. We have also experimented with the popular mixture, which linearly interpolates $P(t \mid r)$ and $P(t \mid u)$. But this approach did not achieve competitive results.

3.2. Language Models

The most straightforward approach to tag recommendation is to simply recommend the most frequent tags for each resource. More formally, the probability for a tag t given a resource r is estimated as:

$$P_{lm}(t \mid r) = \frac{c(t,r)}{\sum_{t_i \in r} c(t_i, r)} \tag{8}$$

where $c(t,r)$ is the count of tag t in resource r. The probability $P_{lm}(t \mid u)$ of a user u using tag t is determined in a similar way from all tags the user has assigned.

Note that we do not need to smoothen the language models as usual, because we smoothen $P(t|r)$ and $P(t|u)$ with $P(t)$ via Equations 6 and 7.

3.3. Latent Dirichlet Allocation

Especially for new resources and users with only few bookmarks, the simple language model does not suffice for tag recommendation, because the tag vocabulary of the already available bookmarks may differ from the preferred tag vocabulary of the user. Smoothing with the global tag probability only effectively switches off tags that are not available for a resource or user.

To also recommend topically related tags, we use Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). The general idea of LDA is based on a simple generative model. Resources and users are modelled as mixtures of latent topics, which in turn consist of a mixture of words. When looking at a resource, for each tag, a user first chooses one of the topics of the resource and then chooses a tag from this topic. Likewise, from the perspective of the user, the user first chooses one of her topics of interest from which she chooses the tag.

More formally, the modeling process of LDA can be described as finding a mixture of topics z for each resource r, i.e., $P(z \mid r)$, with each topic described by tags t following another probability distribution, i.e., $P(t \mid z)$. This can be formalized as

$$P_{lda}(t \mid r) = \sum_{z=1}^{Z} P(t \mid z) P(z \mid r) \tag{9}$$

where $P_{lda}(t \mid r)$ is the probability of tag t for a given resource r and z ranges over the latent topics of the resource. $P(t \mid z)$ is the probability of tag t within topic z (see Equation 11). $P(z \mid r)$ is the probability of picking a tag from topic z in the resource (see Equation 12). The number of latent topics Z has to be defined in advance and allows to adjust the degree of specialization of the latent topics.

LDA estimates the topic-tag distribution $P(t \mid z)$ and the resource-topic distribution $P(z \mid r)$ from an unlabeled corpus of documents using Dirichlet priors for the distributions and a fixed number of topics. Gibbs sampling (Griffiths and Steyvers (2004)) is one possible approach to this end: It iterates multiple times over each tag t_i in resource r, and samples a new topic z for the tag based on the probability $P(z|t_i, r, z_{-i})$ using Equation 10, until the LDA model parameters converge.

$$P(z \mid t_i, r, z_{-i}) \propto (C^{RZ}_{rz} + \alpha) \frac{C^{TZ}_{t_i z} + \beta}{\sum_t C^{TZ}_{tz} + T\beta} \tag{10}$$

C^{TZ} maintains a count of all topic-tag assignments, C^{RZ} counts the resource-topic assignments, z_{-i} represents all topic-tag and resource-topic assignments except the current assignment z for tag t_i, and α and β are the (symmetric) hyperparameters for the Dirichlet priors, serving as smoothing parameters for the counts. The complexity of Gibbs sampling is $O(n * k)$, with n the number of tokens, and k the number of topics. However, inference of the LDA model parameters can be easily parallelized, and well approximated by performing it incrementally on streams of data Yao et al. (2009). Thereby, costly inference can be performed once on a stable base corpus of tagged resources, and the resulting topic-tag distributions can be efficiently updated by folding in new resources.

Based on the counts the posterior probabilities in Equation 9 can be estimated as follows:

$$P(t \mid z) = \frac{C^{TZ}_{tz} + \beta}{\sum_{t_i} C^{TZ}_{t_i j} + T\beta} \tag{11}$$

$$P(z \mid r) = \frac{C^{RZ}_{rz} + \alpha}{\sum_{z_i} C^{RZ}_{rz_i} + Z\alpha} \tag{12}$$

The estimation of $P_{lda}(t \mid u)$ proceeds in the same way as the estimation of $P_{lda}(t \mid r)$ by operating on the individual tag sets of users rather than resources.

For resources, the resulting topics reflect a collaborative shared view of the resource, and the tags of the topics reflect a common vocabulary to describe the resource. Table 1 shows typical examples of resource topics. As can be seen, the topics group typically co-occuring tags, which often will not be used by the same user. E.g., one user may prefer the tag 'photography', another user may prefer 'photo' or 'photos'.

For users, the resulting topics reflect the topical interests of a user, and the tags of topics reflect the individual tagging vocabulary of the user and similar users. Table 2 gives examples of user topics. Note that latent topics are not necessarily disjoint. E.g. 'hardware' occurs in the 'mac' topic as well as in the 'do it yourself' topic, but most certainly these two interpretations of 'hardware' are rather disjoint.

By combining these two perspectives using Equation 5, the resource perspective serves as a selector of the topical content of the resources, while the user perspective takes into account the individual tagging practices of the user.

Tag	Prob.	Tag	Prob.
news	0.201	flickr	0.344
technology	0.182	photography	0.167
tech	0.118	photos	0.117
blog	0.082	photo	0.093
daily	0.070	tools	0.089
geek	0.067	web2.0	0.045
blogs	0.029	visualization	0.016
community	0.025	images	0.015
internet	0.023	pictures	0.012
computers	0.021	api	0.010
web	0.018	search	0.009
forum	0.018	internet	0.007
computer	0.015	applications	0.005
software	0.013	sharing	0.004

Table 1: Top tags composing the latent topics "tech news" and "flickr" based on resource profiles

Tag	Prob.	Tag	Prob.
mac	0.320	diy	0.234
osx	0.215	make	0.099
apple	0.191	hardware	0.084
software	0.170	creativity	0.080
video	0.025	hacks	0.072
quicktime	0.013	electronics	0.070
macintosh	0.012	crafts	0.063
mail	0.012	science	0.046
tv	0.009	mind	0.030
ipod	0.006	theory	0.027
gmail	0.005	photography	0.023
hardware	0.004	engineering	0.019
algorithm	0.003	tutorials	0.017
boot	0.002	language	0.008

Table 2: Top tags composing the latent topics "mac" and "do it yourself" based on user profiles

3.4. Combining LDA and LM

As $P_{lm}(t \mid r)$ and $P_{lda}(t \mid r)$ both constitute (normalized) probability distributions, we can combine these two by straightforward linear interpolation (likewise for $P(t \mid u)$):

$$P(t \mid r) = \lambda \cdot P_{lm}(t \mid r) + (1 - \lambda) \cdot P_{lda}(t \mid r) \tag{13}$$

We have experimented with a broad range for λ, and achieved consistently good results for λ in the range of [0.2..0.8]. We report results for $\lambda \in \{0.0, 0.5, 1.0\}$, where $\lambda = 0$, and $\lambda = 1$ practically switch off the estimates based on language models and latent topics respectively. Combined with the smoothing in Equations 6 and 7, we effectively use a two level smoothing of the simple language model $P_{lm}(t \mid r)$: First by the more general $P_{lda}(t \mid r)$ and then by the marginal tag probability $P(t)$.

4. Evaluation

Evaluating personalized tag recommendation algorithms is not a trivial task. To get precise performance statistics a good way would be to compare two recommendation algorithms in a live tagging system, which allows for a direct user evaluation. Since this scenario is unfeasable or means interfering with a running system other approaches are prefered.

One popular way to evaluate is to take existing data from a tagging system and conduct tests on a hold-out set of tags, resources, or users (Herlocker et al. (2004)). This approach has a promising characteristic: All tags which were used for a resource by a particular user are definitly known. The drawback is that these tags have been added by the user after being suggested by some automatic algorithm within the tagging system. This can bias the tag assignments towards the used tag recommendation algorithm, see Golder and Huberman (2006). Another disadvantage is that only a small set of correct, good tags are actually picked by the user making no distinction between totally unsuited tags and suitable recommended tags which were not picked by the user for whatever reason. Note that this leads to an underestimation of the actual tag recommendation quality.

To extenuate these disadvantages the test dataset has to be designed thoroughly. The strength of a recommendation algorithm can only be judged in comparison with other algorithms run on the same dataset. Thus we need to compare directly state-of-the-art algorithms with the proposed tag recommendation algorithm on the same dataset.

Golder and Huberman (2005) observe that tag distributions for a resource tend to stabilize after around 100 bookmarks. This makes tag recommendation especially challenging for resources having a lot less bookmarks. This so-called cold start problem gives the most discriminative results for different algorithms.

Before we report our results, we have a detailed look into the used datasets, performance metrics, and the used baseline.

4.1. Datasets

We performed experiments on two datasets. The first one is based on a crawl from Delicious. It consists of diverse urls tagged by Delicious users. The second dataset was provided in the context of a tag recommendation challenge held in conjunction with the ECML/PKDD conference 2009. It consists of data from the bookmarking system Bibsonomy, which not only includes tagged urls but also tagged Bibtex entries.

Delicious Dataset. We use a crawl from Delicious provided by Wetzker et al. (2008). The dataset consists of nearly 1 million users crawled between December 2007 and April 2008. The retrieval process resulted in about 132 million bookmarks or 420 million tag assignments that were posted between September 2003 and December 2007. Almost 7 million different tags are contained in the dataset and about 55 million different urls were annotated.

To do the computations in memory and in a reasonable time we were forced to use only a sample of the whole dataset. The huge amount of data and the fact that no spam filtering was done also results in a very sparse overlap between tags, resources and users. To get a dense subset of the sampled data we computed

p-cores described by Batagelj and Zaversnik (2002) for different levels.

For $p = 20$ we get enough bookmarks for each resource to split the data based on resources into meaningful training and test sets (90%:10%). The 20-core ensures that each tag, each resource and each user appears at least 20 times in the tag assignments. For the 10% resources in the test set, we only include the bookmarks for the first n users ($n \in 1, 3, 5, 7, 10, 20$) who annotated a resource into the training set. This results in a setting close to real life situations where users often annotate a resource previously annotated by only a few (n) other users. As soon as a resource is annotated by many users, tag recommendation can exploit the stabilized tag distribution (Golder and Huberman (2005)) for resources and recommending good tags becomes less challenging.

The proposed setup allows to analyze how well different algorithms can generalize from relatively few tags available for a resource similuating the cold start problem in tagging environments.

Parameter settings were tested on 1/256 of the data. We have five test sets containing 10% of the resources with different numbers of "known" bookmarks. On this set, the only preprocessing of the tag assignments performed was the decapitalization of the tags. No stemming or other simplifications were applied. More sophisticated preprocessing improve the results but would complicate the evaluation of the algorithms and the comparison of different methods.

Bibsonomy Dataset.. This dataset consists of the provided training and test data for the Discovery Challenge 2009 held in conjuction with ECML/PKDD 2009 (Eisterlehner et al. (2009)). The training set consists of 253,615 tag assignments done by 1,185 individual users, 14,443 distinct URLs and 7,946 distinct BibTeX posts, and 13,276 distinct tags. This dataset was cleansed before by removing spammers and automatically added tags (like "imported", "public", "systemimported", etc.) and a post-core at level 2 was computed, that is, all users, tags, and resources which appeared in only one post were removed.

Three different tasks were provided aiming at different capabilities of the participating systems. Along with content-based and graph-based tag recommendation, one task dealt with online tag recommendation. Since our approach works without any additional content solely on the tags assigned by users to resources, the second task (graph-based tag recommendation) is predestined to test our algorithms on.

The test dataset for task 2 consists of 775 userId-contentId tuples extracted from the running Bibsonomy bookmarking system. For each userId-contentId tuple the participating systems had to recommend 5 tags. The actual tags assigned by the users are used as ground truth.

4.2. Evaluation Measures

We use standard information retrieval evaluation metrics to report and compare the performance of the algorithms.

- *P@1 — precision at one:* Percentage of test cases where the first recommended tag was actually used by the user to annotate the resource. This is the same as success at one (S@1).
- *P@5 — precision at five:* Percentage of tags among the first five recommended tags that where actually used by the user. Averaged over all test cases.
- *S@5 — success at five:* Percentage of test cases where at least one of the first five recommended tags was used by the user.
- *S@uAVG — success at user average:* Percentage of test cases where at least one of the recommended tags was used by the user. The number of recommended tags is the average number of tags per (other) bookmark for the user.
- *P@uAVG — precision at user average:* Percentage of tags among the top n recommended tags that where actually used by the user, where n is again the average number of tags per bookmark.
- *R@uAVG — recall at user average:* Percentage of user tags among the top n recommended tags, n as above.
- *Fma@5 — f1 macro average at five:* The harmonic mean of averaged precision and recall for the first five recommended tags.
- *Fmi@5 — f1 micro average at five:* The averaged harmonic mean of precision and recall for the first five recommended tags.
- *MRR — mean reciprocal rank:* The average over all test cases of the multiplicative inverse of the rank of the first correct tag.

4.3. Baseline

To get a good estimation of the performance of our tag recommendation algorithms we compare the results with the results from FolkRank by Hotho et al. (2006). FolkRank (FR) is one of the state-of-the-art tag recommender algorithms. Its recommendations are very accurate but this comes with high computational costs. In contrast to our approach, FolkRank does not make use of latent topics but relies on a graph representation of the folksonomy.

The basic idea is to adapt PageRank by Page et al. (1998) to get scores for tags. A graph $G = (V, E)$ is constructed from the folksonomy $F = (U, R, T, X)$, where the vertices are users, resources, and tags ($R \cup U \cup T$) and the edges are co-occurences of tags and users, tags and resources, and users and resources within tag assignments $(u, t, r) \in X$.

The symmetric characteristic of the graph G would lead to scores biased towards "popular", i.e., highly connected entities within the graph when employing the adapted PageRank (ap). Thus, folkrank uses a differential approach and computes the scores for each node based on the difference between a regular PageRank computation and a "personalized" PageRank, like, e.g., Gyöngyi et al. (2004), using a preference vector.

For tag recommendation this preference vector is highly biased towards two entries: the user and the resource for whom the recommendation is computed, see Jäschke et al. (2007). To compare FolkRank with LDA and LM employed only on resources or only on users, we only boost one entry in the preference vector — the resource or the user in question. This gives either resource-centered or user-centered FolkRank results.

Rec. based on		P@1	S@5	S@10	S@uAVG	P@uAVG	R@uAVG	P@5	R@5	F1@5	MRR
User	Resource										
FR	–	0.271	0.537	0.673	0.442	0.190	0.231	0.168	0.262	0.205	0.400
LDA	–	0.279	0.571	0.689	0.475	0.194	0.229	0.178	0.262	0.212	0.407
LM	–	0.284	0.584	0.717	0.482	0.204	0.246	0.187	0.282	0.225	0.424
LDA&LM	–	0.288	0.596	0.715	0.486	0.208	0.250	0.190	0.287	0.228	0.428
–	FR	0.488	0.768	0.824	0.657	0.294	0.376	0.281	0.420	0.337	0.601
–	LDA	0.496	0.815	0.880	0.675	0.328	0.416	0.310	0.460	0.370	0.635
–	LM	0.493	0.762	0.806	0.661	0.352	0.363	0.335	0.411	0.369	0.604
–	LDA&LM	0.560	0.826	0.894	0.718	0.358	0.454	0.334	0.492	0.397	0.678
LM	LM	0.547	0.813	0.882	0.726	0.353	0.441	0.319	0.478	0.382	0.667
LDA	LDA	0.561	0.847	0.908	0.738	0.370	0.467	0.336	0.507	0.404	0.689
LDA	LM	0.532	0.812	0.885	0.722	0.340	0.425	0.313	0.467	0.375	0.653
LM	LDA	0.566	0.859	0.924	0.759	0.386	0.488	0.343	0.526	0.416	0.703
FolkRank		0.570	0.840	0.906	0.734	0.354	0.452	0.325	0.499	0.393	0.689
LDA&LM	LDA&LM	0.610	0.890	0.934	0.795	0.415	0.529	0.372	0.564	0.448	0.733

Table 3: Results for one known bookmark and different algorithms on the Delicious dataset

The FolkRank scores are computed iteratively and finally combined:

$$R^{ap}_{i+1} = c(\alpha R^{ap}_i + \beta A R^{ap}_i) \quad (14)$$

$$R^{pref}_{i+1} = c(\alpha R^{pref}_i + \beta A R^{pref}_i + \gamma P) \quad (15)$$

$$R = R^{pref} - R^{ap} \quad (16)$$

where α, β, γ are constants and c is a normalization factor. A is a row-stochastic version of the adjacency matrix of G.

5. Experiments

We have systematically applied the described approaches for estimating $P(t|r)$, $P(t|u)$, and $P(t|u,r)$, and compared them to the corresponding results using FolkRank.

5.1. Results for Delicious Dataset

Table 3 gives a complete overview for tag recommendation when there is only one bookmark available for the resource.

The first four rows give the results for taking only the user perspective into account, i.e., tags are predicted based on $P(t|u)$ only (or for FolkRank the preference vector is only biased towards the user). We see that generally the probabilistic approach introduced in this paper outperforms FolkRank (FR) w.r.t. all measures. The simple (smoothed) language model approach (LM) slightly outperforms LDA. The linear interpolation of LM with LDA achieves a very slight improvement over LM and LDA alone. It is also clear that the user perspective in isolation performs worse than the resource perspective (next four rows). This is to be expected. The mixture of topics that a user is interested in is typically much more diverse than the mixture of topics a particular resource is about. Thus, no matter how the tag probabilities are estimated, just recommending the most likely tags for a user, disregarding the resource, will often go astray.

The general trends for tag recommendation based on resources only (second four rows) are slightly different. FR and LM are rather clearly outperformed on all measures, but among them, for some measures FR is better than LM and vice versa. LDA comes on a clear second place, and a very clear winner is is again the linear interpolation of LM and LDA.

It is interesting that LDA outperforms LM on resources, while LM outperforms LDA on users. The strength of LDA is to generalize from the tagging practices of individual users who have assigned tags to a particular resource (such as "photography"), in order to also include semantically related tags (such as "photo"). This strength turns out to be a slight weakness for the user perspective, possibly because users tend to stick to a particular vocabulary, and thus the generalization by LDA does not help.

The next four rows inspect the performance of the individual approaches to estimate the probability of a tag when combined for personalized tag recommendation. The most simple (and scalable) approach by just combining the smoothed language models already achieves significant improvements[6] compared to tag recommendation based on LMs for the resource only. Combining only LDA for the user and resource yields further improvement, and the best combination of individual models is achieved by using LM for the user perspective and LDA for the resources. This is consistent with the results for user-based (LM best) and resource-based recommendation (LDA best).

[6] all improvements are significant well beyond a confidence of 0.99 based on a 2-tail paired t-test.

Known Bookmarks	F-Macro Average	
	FolkRank	LDA+LM
1	0.393	0.448
3	0.447	0.476
5	0.462	0.477
10	0.475	0.491

Table 4: F-macro average for different number of known bookmarks on the Delicious dataset

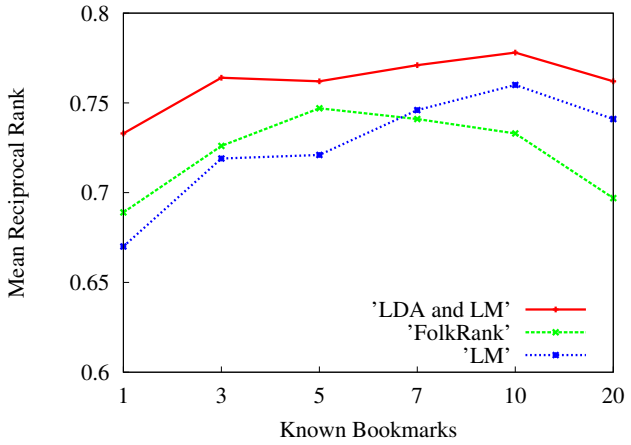

Figure 1: Mean reciprocal rank for different numbers of known bookmarks on the Delicious dataset

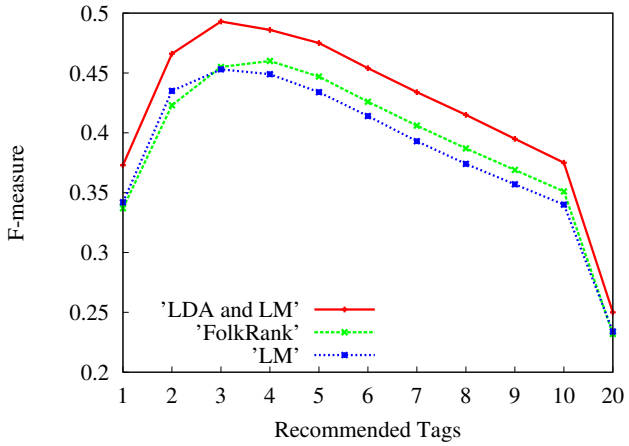

Figure 2: Macro f-measure for different numbers of recommended tags on the Delicious dataset

Finally, the last two rows compare full FolkRank with a complete combination of user-based recommendations using an interpolation of LDA and LM with resource-based recommendation. This full combination outperforms FolkRank signifantly on all measures. With one exception (S@10, with only 3 % improvement), all relative improvements are in the range of 7 % to 17 % with 11 % average.

Table 4 shows the F-macro average comparing the performance having differing prior knowledge about an item.

Figure 1 compares mean reciprocal rank (MRR) of the main approaches when varying the number of available bookmarks between 1 and 20. FR stands for FolkRank, LM for a combination of language models for the user and the resource, LDA and LM for the full combination of language models and LDA on users and resources. The full combination clearly outperforms the other two approaches, but notably the scalable combination of simple language models outperforms FolkRank for more than seven bookmarks. The reason why MRR degrades with all approaches at least for 20 bookmarks is due to the experimental setup. When using 20 bookmarks, much fewer test data are available in the post-core at 20, and the few remaining test data may be the most difficult ones.

Finally, Figure 2 shows the progression of F-Measure depending on the number of recommended tags for resources with three known bookmarks. For all approaches, recommending three tags appears to provide the best balance between recall and precision. This also reflects the tagging behaviour of users who on average assigned 4.3 tags to one resource in our dataset. Again the approach presented in this paper clearly outperforms FolkRank and smoothed language models, with the latter two being more or less on par.

To get an impression of the actual tags recommended, Table 5 gives a randomly picked example of tags recommended by FolkRank and by the approach introduced in this paper. The correctly predicted tags are in bold. We see that our approach correctly predicts 4 tags in top 6, and 6 correct predictions in the top 20,whereas FolkRank predicts only 4 in top 20. But of course a single example can only provide anecdotal evidence. For tag recommendation both tag sets make intuitively sense and the underestimation of the hold-out strategy can be observed.

5.2. *Results for Bibsonomy Dataset*

An overview of the performance of the algorithms on the Bibsonomy dataset is presented in Table 6. Since the official task required five tags to be recommended for each user–resource pair in the testset, we only report precision@5, recall@5, and f1@5. The first four lines in the table show the results for only using the user profiles to recommend tags. As with the Delicious dataset, the results for using only the resource profiles are nearly three times as good. Using language models only on the resource information is already enough to beat FolkRank. Adding LDA and the knowledge about the tagging behaviour of a user in the past just slightly improves the results (from 0.301 to 0.308). This is due to the characteristics of this dataset. The resources are already tagged by many users and the tag distribution for a single resource has already stabilized. We are not dealing with a cold start problem in this dataset which lowers the probability that LDA can find very relevant tags that have not been used by users for a resource so far.

Another dataset dependent parameter is the number of latent topics. Table 7 depicts the f1-measure for using between 100 and 5,000 latent topics for the corpus. A maximum of the performance is reached by 1,000 topics. The table also indicates, that recommending 4 tags gives better recall and precision values then recommending 5 tags. The reason for this is the average number of tags a user assigns to a resource in the Bibsonomy system (3.96 tags/resource). By adapting our algorithm to the average number of tags a user assigns, i.e. recommending not 5 tags for all users but less if the average number of tags a user assigned to a resource was less than 5, we can even improve the F1@5 score. Of course recall will drop a little, but precision will be significantly higher.

To see the impact of the user profile information and the resource profile information, we plotted the f1 values for different

Original tags from user	FolkRank		LDA&LM	
	tags	score	tags	score
	microformats	0.0138	**microformats**	50.6
	howto	0.0078	howto	15.1
	standards	0.0070	**tutorial**	12.8
	tutorial	0.0068	standards	11.5
	collection	0.0066	**programming**	9.6
	information	0.0064	**reference**	8.4
webdev	resources	0.0061	semantic	7.2
programming	tutorials	0.0060	development	5.3
reference	**webdev**	0.0024	software	4.0
web2.0	tool	0.0021	web	3.4
webdesign	development	0.0020	xml	3.2
xhtml	**programming**	0.0018	**webdesign**	3.0
microformats	web	0.0013	code	2.7
tutorial	html	0.0011	tool	2.4
	code	0.0009	**webdev**	2.3
	software	0.0007	information	2.3
	javascript	0.0006	css	2.0
	python	0.0005	design	2.0
	snippets	0.0004	tips	2.0
	optimization	0.0004	tutorials	1.6

Table 5: Recommended tags for user 800 and resource "http://www.xfront.com/microformats/" from the Delicious dataset

weights α in Figure 3. The weighting is done similar to Equation 13:

$$P(t \mid r, u) = \alpha \cdot P_{lm\&lda}(t \mid r) + (1 - \alpha) \cdot P_{lm\&lda}(t \mid u) \quad (17)$$

where α defines the weight for combining the resource with the user information. A maximum for recall as well as for precision is found for $\alpha = 0.7$. But in general the resource information is much more valuable in this setting than the user profiles.

In this approach, we are only using the tag assignment information and no meta-information or content in any way. This makes our approach universal with respect to the underlying tagging system. On the other hand, different systems, like for example the bookmarking system Bibsonomy, offer more information that could be used. This information can be very valuable when recommending tags. The best tag recommendaton systems at the discovery challenge exploited this additional information, such as content of the tagged resource or meta information like a resource description. A recent study by Lipczak and Milios (2010) analyzed the use of resource titles to tag these resources. It shows the benefits that can be gained by recommender algorithms taking these tagging system dependent information into account.

6. Conclusion

In this paper we have explored user-centered and resource-centered approaches for personalized tag recommendation. We compared and employed a language modeling approach and an approach based on Latent Dirichlet Allocation. We furthermore thoroughly investigated the use of language models and

Rec. based on		Prec@5	Rec@5	F1@5
User	Resource			
FR	–	0.079	0.124	0.096
LM	–	0.083	0.125	0.100
LDA	–	0.084	0.130	0.102
LDA&LM	–	0.084	0.129	0.102
–	LDA	0.209	0.318	0.252
–	FR	0.238	0.365	0.288
–	LM	0.258	0.361	0.301
–	LDA&LM	0.253	0.384	0.305
LM	LM	0.218	0.334	0.264
LDA	LDA	0.218	0.336	0.265
LM	LDA	0.215	0.339	0.263
LDA	LM	0.230	0.350	0.270
FolkRank		0.241	0.376	0.294
LDA&LM	LDA&LM	0.252	0.394	0.308

Table 6: Results on Bibsonomy dataset for five recommende tags

LDA for tag recommendation showing that simple language models built from users and resources yield competitive performance while consuming only a fraction of the computational costs compared to more sophisticated algorithms. We showed that the combination of both methods (LDA and LM) tailored to users and resources outperforms state-of-the-art tag recommendation algorithms with respect to a broad variety of performance metrics.

For future work we want to investigate the use of these methods for item or user/community recommendation in the context of tagging systems. Especially for item recommendation, the extension of our approach to incorporate content information might be beneficial. Even for non-textual resources like videos or audio, additional metadata could be exploited. It would also be interesting to see whether the behavior of the current algorithms changes when applied to a photo or video tagging system instead of bookmarking systems. One question in this context woulb be whether users tag videos differently then web pages and whether LDA and LM can be employed in the same manner. Finally, we plan to investigate how additional contextual knowledge such as time, location, and current task can be used to further personalize tag recommendation. A starting point to this end could be to have a multi-lingual aware, personalized tagging system dealing with identification of users' native languages and possibly automatic translation of tags.

No. Tags	# LDA topics					
	100	200	500	1,000	3,000	5,000
1	0.219	0.208	0.230	0.221	0.222	0.214
2	0.286	0.280	0.292	0.295	0.288	0.277
3	0.301	0.309	0.311	0.313	0.301	0.306
4	0.308	0.313	0.311	0.318	0.307	0.310
5	0.303	0.311	0.305	0.315	0.303	0.306

Table 7: F-measure for different number of recommended tags and different number of LDA topics using LM&LDA on resources and LM&LDA on users for the Bibsonomy dataset

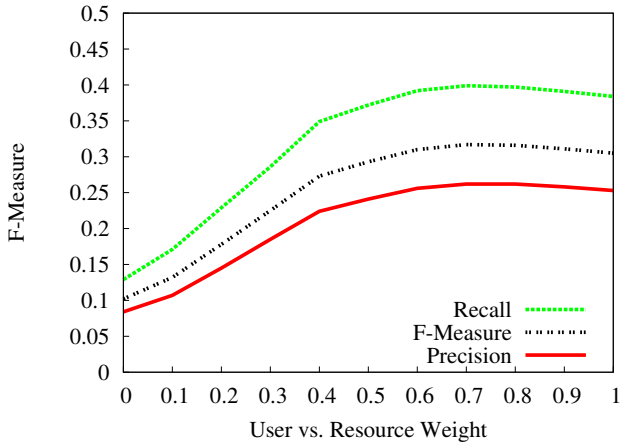

Figure 3: Precision, recall, and f-measure for different weights on user and resource information on the Bibsonomy dataset

Acknowledgment

This work is partially supported by the European Union FET project LivingKnowledge (FP7–ICT–231126).

References

H. Halpin, V. Robu, H. Shepherd, The complex dynamics of collaborative tagging, in: Proceedings of the 16th International Conference on World Wide Web (WWW 2007), ACM, 2007, pp. 211–220.

K. Bischoff, C. S. Firan, W. Nejdl, R. Paiu, Can all tags be used for search?, in: CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, ACM, New York, NY, USA, 2008, pp. 193–202.

S. A. Golder, B. A. Huberman, The structure of collaborative tagging systems, CoRR abs/cs/0508082 (2005).

D. Bollen, H. Halpin, An experimental analysis of suggestions in collaborative tagging, in: 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009, Milan, Italy, 15-18 September 2009, Main Conference Proceedings, IEEE, 2009, pp. 108–115.

Y. Song, L. Zhang, C. L. Giles, Automatic tag recommendation algorithms for social recommender systems, ACM Trans. Web 5 (2011) 4:1–4:31.

A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, Information retrieval in folksonomies: Search and ranking, in: The Semantic Web: Research and Applications, Springer, Heidelberg, Germany, 2006, pp. 411–426.

F. Eisterlehner, A. Hotho, R. Jschke (Eds.), ECML PKDD Discovery Challenge 2009 (DC09), volume 497 of *CEUR-WS.org*, 2009.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. 22 (2004) 5–53.

G. Mishne, Autotag: a collaborative approach to automated tag assignment for weblog posts, in: WWW '06: Proceedings of the 15th international conference on World Wide Web, ACM, New York, NY, USA, 2006, pp. 953–954.

P. A. Chirita, S. Costache, W. Nejdl, S. Handschuh, P-tag: large scale automatic generation of personalized annotation tags for the web, in: WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM, New York, NY, USA, 2007, pp. 845–854.

R. Jäschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, G. Stumme, Tag recommendations in folksonomies, in: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer, Heidelberg, Germany, 2007, pp. 506–514.

Z. Xu, Y. Fu, J. Mao, D. Su, Towards the semantic web: Collaborative tag suggestions, in: Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference.

N. Garg, I. Weber, Personalized, interactive tag recommendation for flickr, in: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, ACM, New York, NY, USA, 2008, pp. 67–74.

B. Sigurbjörnsson, R. van Zwol, Flickr tag recommendation based on collective knowledge, in: WWW '08: Proceeding of the 17th international conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 327–336.

P. Heymann, D. Ramage, H. Garcia-Molina, Social tag prediction, in: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2008, pp. 531–538.

R. Krestel, P. Fankhauser, W. Nejdl, Latent Dirichlet Allocation for Tag Recommendation, in: RecSys '09: Proceedings of the third ACM conference on Recommender systems, ACM, New York, NY, USA, 2009, pp. 61–68.

R. Wetzker, C. Zimmermann, C. Bauckhage, S. Albayrak, I tag, you tag: translating tags for advanced user models, in: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining, ACM, New York, NY, USA, 2010, pp. 71–80.

P. Symeonidis, A. Nanopoulos, Y. Manolopoulos, Tag recommendations based on tensor dimensionality reduction, in: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, ACM, New York, NY, USA, 2008, pp. 43–50.

S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in: WSDM '10: Proceedings of the third ACM international conference on Web search and data mining, ACM, New York, NY, USA, 2010, pp. 81–90.

Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, C. L. Giles, Real-time automatic tag recommendation, in: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2008, pp. 515–522.

X. Si, M. Sun, Tag-lda for scalable real-time tag recommendation, Journal of Computational Information Systems 6 (2009) 23–31.

R. Krestel, P. Fankhauser, Tag recommendation using probabilistic topic models, in: F. Eisterlehner, A. Hotho, R. Jäschke (Eds.), ECML/PKDD Discovery Challenge (DC'09), Workshop at ECML/PKDD 2009), Bled, Slovenia, pp. 131–141.

M. Bundschus, S. Yu, V. Tresp, A. Rettinger, M. Dejori, H.-P. Kriegel, Hierarchical bayesian models for collaborative tagging systems, in: ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2009, pp. 728–733.

M. Harvey, M. Baillie, I. Ruthven, M. J. Carman, Tripartite hidden topic models for personalised tag suggestion, in: Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings, volume 5993 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 432–443.

D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

T. L. Griffiths, M. Steyvers, Finding scientific topics., Proc Natl Acad Sci U S A 101 Suppl 1 (2004) 5228–5235.

L. Yao, D. Mimno, A. McCallum, Efficient methods for topic model inference on streaming document collections, in: KDD '09: Proceedings of the 15th ACM SIGKDD conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2009, pp. 937–946.

S. Golder, B. A. Huberman, Usage patterns of collaborative tagging systems, Journal of Information Science 32 (2006) 198–208.

R. Wetzker, C. Zimmermann, C. Bauckhage, Analyzing Social Bookmarking Systems: A del.icio.us Cookbook, in: Proceedings of the ECAI 2008 Mining Social Data Workshop, pp. 26–30.

V. Batagelj, M. Zaversnik, Generalized cores, CoRR cs.DS/0202039 (2002).

L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford Digital Library Technologies Project, 1998.

Z. Gyöngyi, H. Garcia-Molina, J. O. Pedersen, Combating web spam with trustrank, in: Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 - September 3 2004, Morgan Kaufmann, 2004, pp. 576–587.

M. Lipczak, E. Milios, The impact of resource title on tags in collaborative tagging systems, in: HT '10: Proceedings of the 21st ACM conference on Hypertext and hypermedia, ACM, New York, NY, USA, 2010, pp. 179–188.