Tag Sources for Recommendation in Collaborative Tagging Systems

Marek Lipczak, Yeming Hu, Yael Kollet, Evangelos Milios

Faculty of Computer Science, Dalhousie University, Halifax, Canada, B3H 1W5
lipczak@cs.dal.ca

Abstract. Collaborative tagging systems are social data repositories, in which users manage resources using descriptive keywords (tags). An important element of collaborative tagging systems is the tag recommender, which proposes a set of tags to each newly posted resource. In this paper we discuss the potential role of three tag sources: resource content as well as resource and user profiles in the tag recommendation system. Our system compiles a set of resource specific tags, which includes tags related to the title and tags previously used to describe the same resource (resource profile). These tags are checked against user profile tags – a rich, but imprecise source of information about user interests. The result is a set of tags related both to the resource and user. Depending on the character of processed posts this set can be an extension of the common tag recommendation sources, namely resource title and resource profile. The system was submitted to ECML PKDD Discovery Challenge 2009 for "content-based" and "graph-based" recommendation tasks, in which it took the first and third place respectively.

1 Introduction

The emergence of social data repositories made a fundamental change in the way information is created, stored and perceived. Instead of a rigid hierarchy of folders, collaborative tagging systems (e.g., BibSonomy[1], del.icio.us[2], Flickr[3], Technorati[4]) use a flexible folksonomy of tags. The folksonomy is created collaboratively by system users. While adding a resource to the system, users are asked to define a set of tags – keywords which describe it and relate it to other resources gathered in the system. To ease this process, some folksonomy services recommend a set of potentially appropriate tags. Proposing a tag recommendation system was a task of ECML PKDD Discovery Challenge 2009[5]. This paper presents a tag recommendation system submitted to the challenge.

[1] http://bibsonomy.org/help/about/
[2] http://del.icio.us/about/
[3] http://flickr.com/about/
[4] http://technorati.com/about/
[5] http://www.kde.cs.uni-kassel.de/ws/dc09/

1.1 Definitions

Collaborative tagging systems allow *users* ($u_i \in U$) to store *resources* ($r_j \in R$) in the form of *posts* ($p_{ij} \in P$). A post is a triple $p_{ij} = (u_i, r_j, T_{ij})$, where $T_{ij} = \{t_k\}$ is a *set of tags* assigned by the user to the resource. The data structure constructed by the collaborative tagging system (referred to as *folksonomy* [5]) is simply a set of posts. However, relations between three basic elements of the post allow us to represent the folksonomy as a tripartite graph of resources, users and tags. Each post can be then understood as a set of edges that form triangles connecting resource, user and tag. Projections of this tripartite graph can be used to examine the relations between folksonomy elements (e.g., two tags can be considered as similar when they are both linked to a large number of common resources, two users are similar when they are linked to the same tags).

Tag recommendation s is a pair (t, l), where t is a tag and l is a *recommendation score*, which is supposed to reflect the likelihood of the tag t being chosen by a user as a proper tag. A tag recommendation system returns a set of tag recommendations S. In this paper we use the term *tag recommendation set* (or simply *recommendation*) not only to refer to the final set of tags returned to the user, but also to denote the results of intermediate tag recommendation steps. In section 5 we define a set of operations on tag recommendation sets, which are used by our tag recommendation system.

User profile is a set of tags used by the user prior to the post that is being currently added to the system, $\mathbb{P}_u = \{t_k : u_i = u, r_j \in R, p_{ij} \in P, t_k \in T_{ij}\}$. The user profile is usually referred to as *personomy* [5]. We use a more general term, because it does not imply that the profile is personal. By analogy we can define a *resource profile*, which contains all tags that were attached to the resource (e.g., a scientific publication) by all users prior to the current post, $\mathbb{P}_r = \{t_k : u_i \in U, r_j = r, p_{ij} \in P, t_k \in T_{ij}\}$. Both user and resource profiles can serve as a simple tag recommendation set. For example, resource profile recommendation $S_{\mathbb{P}_r}$ is a set of tags from resource profile of r. Their score is the ratio of posts in which the tag was used to all posts of the resource (Eq. 1). The intuition behind this formula is that tags frequently used to describe a resource are likely to be used again, hence they are good recommendations.

$$l(t_k, r) = \frac{|\{p_{ij} : u_i \in U, r_j = r, t_k \in T_{ij}\}|}{|\{p_{ij} : u_i \in U, r_j = r\}|} \tag{1}$$

1.2 Tag recommendation tasks

The off-line evaluation of a tag recommendation system for challenge purposes is a complex task. Tags added to the resource are highly dependent on the state of the system and previous decisions of the user. It is not possible to create a large, realistic test dataset of posts, hiding at the same time the tags used in these posts. A test dataset which is large enough to objectively measure the quality of a recommendation system must cover a long period of time. If the tags in test data are hidden we lose access to the information about the state of the system, especially newly joined users, which make the dataset not representative. To ease this problem, the organizers of ECML PKDD Discovery Challenge 2009 divided the recommendation task into two subtasks which simulate two complementary recommendation approaches.

The first task "content-based recommendation" focuses on the content of a resource that is tagged. In this task we assume that information about the resource and user profile is in most cases not available in the folksonomy. A recommender based on resource content is especially important for new users, which are in the early stage of building their profile. Although, as shown in Section 3.1, the need of creating the recommendation based only on the content is rare, the content based recommender can be a valuable starting point for more complex recommenders that use information gathered in the folksonomy. Such more complex recommenders are evaluated in the second task – "graph-based recommendation". The test set in this task contains only users, resources and tags that were present at least twice in the training data. To obtain this set the organizers extracted k-core of order 2 [2] of tripartite graph of users, resources and tags created from training data. The test set contained only posts for which user, resource and all tags can be found in the k-core. It is important to notice that the second task neglects the disproportion between the number of unique resources and users. It also greatly simplifies the recommendation task by removing posts with unique tags which are hardest to recommend in real systems. To improve the results for this task the system must follow some unrealistic assumptions. Although this paper describes an entry to the challenge, we aimed to present a general system which can be applied to a real folksonomy based repository of bookmarks or scientific publications. Each modification that was made to match the specific constraints created by the dataset and the second task of the challenge is clearly stated.

2 Related work

Most of the tag recommendation systems presented in the literature are graph-based methods. It is a natural choice for folksonomies in which textual content is hard to access. For example, a system by Sigurbjörnsson and van Zwol [9] uses co-occurrence of tags to propose tags that complement user-defined tags of photographs in Flickr. Jäschke et al. [6] proposed a graph-based recommendation system for social bookmarking services. The method is based on FolkRank, a modification of PageRank, which is suited for folksonomies. The evaluation on a dense core of folksonomy showed that the FolkRank based recommender outperforms PageRank and collaborative filtering methods.

Even if a tag recommendation system extracts tags from the resource content, usually it also uses the graph information. An example of a content-based recommender is presented by Lee and Chun [7]. The system recommends tags retrieved from the content of a blog, using an artificial neural network. The network is trained based on statistical information about word frequencies and lexical information about word semantics extracted from WordNet. Another system designed to recommend tags for blog posts is TagAssist [10]. The recommendation is built on tags previously attached to similar resources. Meaning disambiguation is performed based on co-occurrence of tags in the complete repository.

Finally, we would like to mention two somewhat similar systems which took the first and second place in the ECML PKDD Discovery Challenge 2008. The winning system was proposed by Tatu et al. [11], while the second place was taken by our submission [8]. Both systems utilize information from resource content and the folksonomy graph. The graph is used to create a set of tags related to the resource and a set of tags related to the user who is adding the resource to the system. The winning system bases these sets on tags gathered in the profile of resource or user. Natural language processing techniques are later used to extend the set of tags related to resource or user (i.e., WordNet based search for words that represent the same concept). Our system bases the resource related tags on the resource title, the set is extended by finding tags that co-occur with the base tags in the system. The user related tags are simply the tags from the user profile. The intersection of both sets creates a set of tags that are related to both resource and user. Our system tries to extend this set by finding more related tags in user profile. Finally, both systems extract tags from resource content and join the content tags with the resource and user related tags to create the final recommendation.

3 BibSonomy dataset

All presented experiments and the evaluation of proposed tag recommendation system were performed on a snapshot of BibSonomy [4], a collaborative tagging system, which is a repository of website bookmarks and scientific publications (represented by BibTeX entries). The training dataset contained posts entered to the system before January 1, 2009. The test data contained posts entered between January 1, 2009 and June 30, 2009. The snapshot was provided by the organizers of the ECML PKDD Discovery Challenge 2009. The preprocessing steps, applied prior to the release of the dataset, included removing useless tags (e.g., *system:unfiled*), changing all letters to lower case and removing non-alphabetical and non-numerical characters from tags. We decided to clean the dataset further by removing sets of posts that were imported from an external source. This preprocessing step involved posts for which one set of tags, defined by user or system, was assigned to a large number of imported resources. An example of such a set consists of $9,183$ posts tagged with tag *indexforum* by one user. Leaving that tag in the system would result in a biased profile of its author. Unfortunately, this cleaning step could not detect another type of imported posts, for which the system automatically defines tags and timestamps based on the information from an external source. An example of such posts is a set of bookmarks imported from a web browser, for which the collaborative tagging system can use the names of bookmark folders to automatically define tags. The second preprocessing step applied to the released data was separation of bookmark and BibTeX posts. We observed that the vocabulary used for both types of resources is different, even for individual users. Some of the tags (e.g., *free*) have different meaning when tagging websites or scientific publications. Finally, content based recommendation can be based on different metadata fields in both resource types.

3.1 General characteristics

According to the statistical information about the dataset presented on the Discovery Challenge website[6] the BibSonomy snapshot matches the usual characteristics of folksonomies, including large disproportion between the number of unique resources and users (Table 1). Among the posts in the BibSonomy snapshot 90% contained unique resources. These resources cannot be found in any other post, hence it is not possible to deduce tag recommendation based on resource profile. At the same time 0.8% of the posts, corresponding to 3,167 posts, were entered by users with no previous posts in the system. Except those posts, every time a post is added, the system is able to use the user profile to recommend tags. Similar proportions can be observed for the CiteULike[7] dataset.

	BibSonomy		CiteULike	
number of tags	1,401,104		4,927,383	
number of unique tags	93,756	(7% of tags)	206,911	(4% of tags)
number of bookmark posts	263,004		N/A	
number of unique bookmarks	235,328	(89% of posts)	N/A	
number of BibTeX posts	158,924		1,610,011	
number of unique BibTeX entries	143,050	(90% of posts)	1,390,747	(86% of posts)
number of users	3,617	(1% of posts)	42,452	(3% of posts)

Table 1. Statistics of BibSonomy training data compared to CiteULike dataset (complete dump up to February 27, 2009). Both datasets have similar proportion of unique tags, posts and users.

The disproportion between unique resources and users is ignored in the test data of "graph-based recommendation" task. All users and resources present in the dataset can be found in the training data at least twice. Despite this fact the differences in statistical characteristics of resource and user profiles should be taken into consideration while proposing a recommendation system for this task. The cumulative frequency distribution of resources shows that both for bookmark and BibTeX entries, even if we remove elements that occurred twice or less, most of the remaining elements still have a very small profile (Fig. 1). Looking at the same statistic for users we see that a significant fraction of them have over 100 posts in their profiles. Hence user profiles are likely to contain more

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

[7] http://www.citeulike.org/

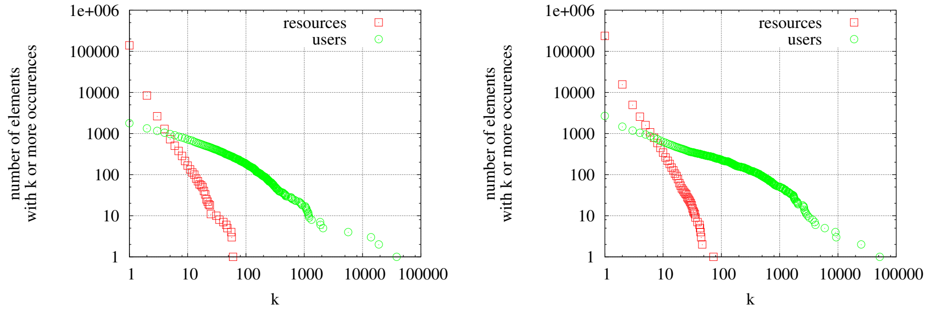


Fig. 1. Cummulative frequency distribution of resources and users for BibTeX (left) and bookmark (right) data. Much steeper curve for resources shows that we are much less likely to find a rich resource profile, comparing to the profiles of users.

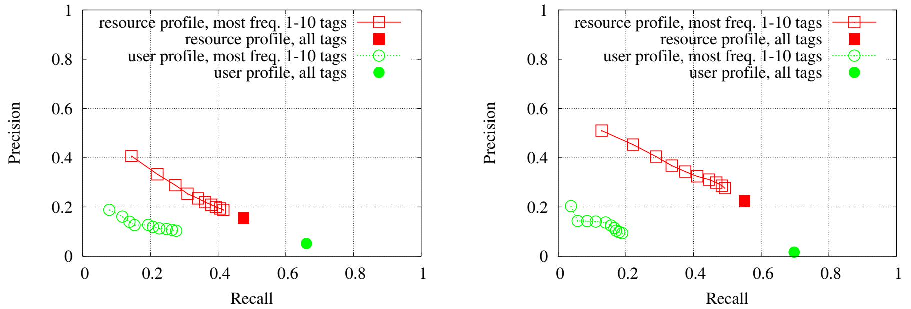


Fig. 2. Precision and recall of most frequent tags from resource and user profiles for BibTeX (left) and bookmark (right) "graph-based recommendation" task data.

potentially useful tags. To confirm this hypothesis we ran another experiment in which we simulated the test data of "graph-based recommendation" task and checked what is the precision and recall of basic recommenders that propose tags from resource/user profile sorted by frequency against real tags. To obtain a test set we divided the training data into training posts (entered before September 1, 2008) and test posts (entered later). We pruned them to be sure that all resources, users and tags occurred in the remaining part of the training set at least twice. Although this setting favours resource profiles, their overall recall is still lower than recall of the user profiles (Fig. 2). The fact that resource profiles are smaller makes them, however, a more precise source of tags. High recall of user profiles was observed by us repeatedly in many experiments. This is the reason why in our work we focused on user profiles, trying to increase the precision of this source of tags, while preserving reasonably high recall.

4 Tag recommendation sources

The presented recommendation system is the evolution of the work on the system [8] submitted to the ECML PKDD Discovery Challenge 2008[8]. In this section we summarize the results of experiments conducted during the work on the previous version of the system. Their main objective was to evaluate the quality of three basic sources of tags – words from resource title, tags assigned to the resource by other users (resource profile) and tags in user profile.

Resource title We tested most of the metadata fields looking for potential tags. Among them the resource title appears to be the most robust source of tag recommendations. The title is a natural summarization of web page or scientific publication, which means it plays a similar role as tags. In addition, the title is present on the resource posting page, which means it can possibly suggest tags to the user. It is easy to notice the evidence for this observation in the example posts of *User B* and *User C* shown in Table 2. Both of them used the tags *prediction* and *social* for "Social tag prediction" paper, which became the only occurrence of these tags in their profiles, unlike tag *recommender* which was used by them around fifty times, probably to describe the general area of interests. The number of words in the title is comparable to the number of tags, hence no additional cleaning steps are needed the achieve fairly high precision comparing to other examined tag sources (around 0.1). The drawback of this source is low recall (around 0.2), which makes the title inappropriate as a stand-alone tag recommender. For bookmark posts the web page URL appears to be another valuable source of tags. Although URL tags are less precise than title tags, their union can increase the recall of recommendation.

Posts:	*Social tag prediction*	*Towards the Semantic Web: Collaborative Tag Suggestions*
User A	Heymann 08 *tag recommendation*	Xu 06 *tag recommendation*
User B	**prediction** *tag recommender* **social** tagging	*tag recommender* tagging
User C	folksonomy **prediction** *recommender* **social** tag *tagging* toread	folksonomy *recommender* summerschool *tagging*

Table 2. Example posts of three users tagging two publications related to the tag recommendation problem (two tags were removed to increase anonymity of posts). **Bold** tags seem to be suggested by the title. Tags in *italics* likely represent the concept of tag recommendation problem in users' profiles.

Resource profile Tags assigned to the resource by other folksonomy users are not a good source of tag recommendations. One of the reasons is the sparsity of

[8] http://www.kde.cs.uni-kassel.de/ws/rsdc08/

data; 90% of resources were added to the system only once. This fact significantly limits the possible recall of this source of tags. The other issue is the personal character of posts and tags, which hurts the precision of retrieved tags. Given the example of two resources about the same concept, we see that users cannot agree on tags describing it: *tag recommendation*, *tag recommender*, *tagging recommender* (Table 2). The variety of tags attached by users creates, however, another application of resource tag sets. Mining relations between tags attached to the same resource can result in a graph of relations between tags. Using a relationship graph the system can identify tags which are also potential recomendations. The graph consists of general relations between tags and can be used independently of the resources, which reduces the negative impact of data sparsity. In our work we use two types of graphs. *TagToTag graph* is a directed graph which captures the co-occurence of tags. The weight of an edge is analogous to the confidence score (Eq. 2) in association rule mining [1], where $support(\{t_1 \cap t_2\})$ is the number of co-occurrences of tags t_1 and t_2 and $support(\{t_1\})$ is the number of occurrences of tag t_1. The second graph (*TitleToTag*) is created specifically for the resource title as the base of the recommendation. Using the same model it captures the relations between words from resource title and its tags.

$$confidence(t_1,\ t_2) = \frac{support(\{t_1 \cap t_2\})}{support(\{t_1\})} \tag{2}$$

User profile For cognitive simplicity and effcient retrieval, a typical user employs the same limited set of tags to describe resources of the same topic (Table 2). This pattern is the reason for high recall of user tags. On the other hand the user profile is a combination of tags related to many user interests and activities, which makes it a very imprecise source of tags. The most frequent tags from the user profile are likely to be related to the most central interests of the user. In our system we try to utilize the potential of user profile tags to extract user's tags that are related to the interests specific to the posted resource.

5 Tag recommendation system

Our tag recommendation system is a composition of six basic tag recommenders (Fig. 3). The result of each recommender is a tag recommendation set with scores in the range $[0, 1]$. The recommender makes a decision based on the resource content, resource related tags and user profile tags. However, its design makes it applicable to all posts even if the resource or user profile cannot be found in the system database. In such cases, the corresponding basic recommenders are not active. The following sections and Algorithm 1 give the detailed description of each basic recommender and the data flow in the system.

5.1 Recommendation based on resource content

The process starts with the extraction of potential tags from the content of resource. For BibTeX posts the **title** of publication is used, for bookmarks the

Algorithm 1: Tag recommendation system

```
Data: a resource $r$ and user $u$
Result: a tag recommendation set $S_{final}$
begin
    /*Step 1 – Extraction of content based tags*/
    $Words_{title} \longleftarrow extractTitleWords(r)$
    $S_{title} \longleftarrow \emptyset$
    foreach $w \in Words_{title}$ do
        $S_{title}$ add $makeTag(w, getPriorUsefullness(w))$
    $removeLowQualityTags(S_{title}, 0.05)$
    if $isBookmark(r)$ then
        $Words_{URL} \longleftarrow extractUrlWords(r)$
        $S_{URL} \longleftarrow \emptyset$
        foreach $w \in Words_{URL}$ do
            $S_{URL}$ add $makeTag(w, getPriorUsefullness(w))$
        $removeLowQualityTags(S_{URL}, 0.05)$
        $rescoreLeadingPrecision(S_{title}, 0.2)$
        $rescoreLeadingPrecision(S_{URL}, 0.1)$
    $S_{content} \longleftarrow mergeSumProb(S_{title}, S_{URL})$
    /*Step 2 – Retrieval of resource related tags*/
    $S_{TitleToTag} \longleftarrow \emptyset$// related tags from TitleToTag graph
    $S_{TagToTag} \longleftarrow \emptyset$// related tags from TagToTag graph
    $S_{\mathbb{P}_r} \longleftarrow getProfileRecommendationBasic(\mathbb{P}_r)$
    foreach $s_k \in S_{title}$ do
        $S_{s_k,TitleToTag} \longleftarrow \emptyset$
        foreach $t \in getRelated(g_{TitleToTag}, s_k)$ do
            $S_{s_k,TitleToTag}$ add $makeTag(t, s_k.l * confidenceTitleToTag(s_k.t, t))$
    foreach $s_k \in S_{content}$ do
        $S_{s_k,TagToTag} \longleftarrow \emptyset$
        foreach $t \in getRelated(g_{TagToTag}, s_k)$ do
            $S_{s_k,TagToTag}$ add $makeTag(t, s_k.l * confidenceTagToTag(s_k.t, t))$
    $S_{TitleToTag} \longleftarrow unionProb(T_{s_1,TitleToTag}, \ldots, T_{s_n,TitleToTag})$
    $S_{TagToTag} \longleftarrow unionProb(T_{s_1,TagToTag}, \ldots, T_{s_n,TagToTag})$
    $S_{r\ Related} \longleftarrow unionProb(S_{TitleToTag}, S_{TagToTag}, S_{\mathbb{P}_r})$
    /*Step 3 – Retrieval of resource and user related tags*/
    $S_{\mathbb{P}_u} \longleftarrow getProfileRecommendationByDay(\mathbb{P}_u)$
    $S_{r,u\ Related} \longleftarrow indersectionProb(S_{r\ Related}, S_{\mathbb{P}_u})$
    /*Final recommendation*/
    $rescoreLeadingPrecision(S_{title}, 0.3)$
    $rescoreLeadingPrecision(S_{\mathbb{P}_r}, 0.3)$
    $rescoreLeadingPrecision(S_{r,u\ Related}, 0.45)$
    $S_{final} \longleftarrow unionProb(S_{title}, S_{\mathbb{P}_r}, S_{r,u\ Related})$
end
```

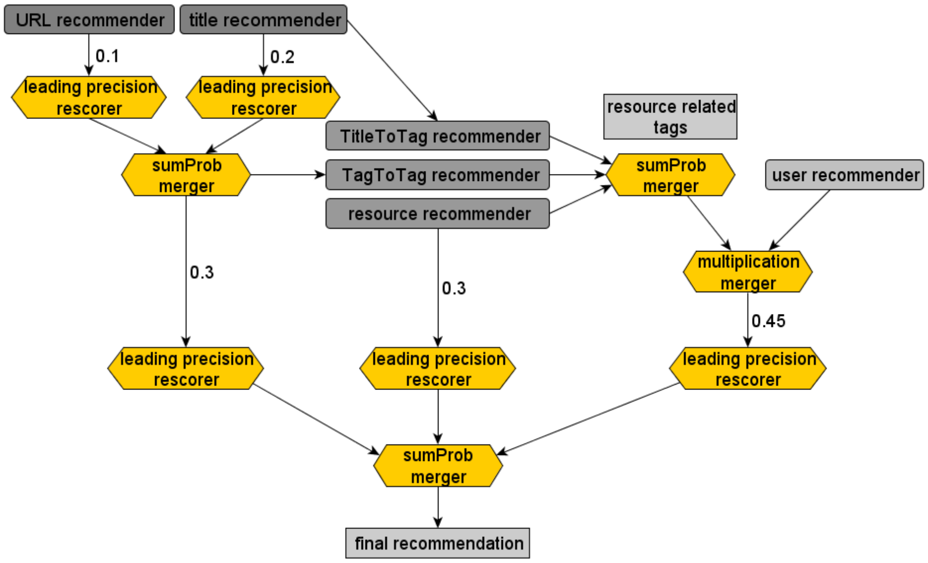


Fig. 3. Data flow in proposed tag recommendation system.

title recommendation is combined with tags extracted from the resource **URL**. Each word extracted from the title (or URL) is scored based on the usage of this word in previous posts. The score is the ratio of the number of times the word was used in the title (or URL) and as a tag to the total number of occurrences of the word in the title (or URL). Low-frequency words (i.e., words that were used in the title less than 50 times) are assigned an arbitrary score 0.1 which is the estimated probability of using a low-frequency word as a tag. To improve precision, content based recommender tags with score lower than 0.05 are removed from the recommendation set. This step serves also as a language independent stop-words remover. Preliminary experiments indicated that the bookmark title is more precise source of tag recommendation than its URL. This observation should be reflected in the way both tag recommendation sets are merged for bookmark posts. We tested a few rescoring functions, the best results were observed for the *leading precision* rescorer (Eq. 3), which sets the average precision (based on training data) as the score of first tag l_1 and modifies the scores of following tags l_i to preserve the proportion between all tag scores. Based on the tests on training data, the average precision of the title tag with the highest score is 0.2, while for URL it is 0.1.

$$l'_i = \frac{avgPrecisionAt1 * l_i}{l_1} \tag{3}$$

5.2 Extraction of resource related tags

The result of title recommender is later used to propose title related tags in **TitleToTag recommender**. The related tags are extracted for each title word independently. The relation score, multiplied by the score of the word from the title recommender, becomes the score of the tag. This process produces a set of related tags for each title word. These sets are later merged, the scores of tags that can be found in more than one set are summed as they were probabilities of independent probabilistic events (Eq. 4). **TagToTag recommender** processes tags analogously, however, the input of this recommender is a complete content based tag recommendation set (title and URL for bookmarks). The aim of these recommenders is to produce a large, but likely not precise set of tags related to the resource. The third recommender that is able to produce a similar set is the **resource recommender**, which returns a set of tags from resource profile. The score of resource tag is the number of its occurrences divided by the number of occurrences of the resource. Although for most real posts this recommender would not return any tags, it plays a significant role in the "graph-based recommendation" task, where the resource of each tested post can be found in the system database at least twice. The scores of the results of three recommenders are summed in a probabilistic way (Eq. 4). This union of tags represents all the tags that are somehow related to the resource, and we refer to them as *resource related* tags.

$$l_{merged} = 1 - \prod_{i:t_i = t_{merged}} (1 - l_i) \tag{4}$$

5.3 Recommendation based on user profile

The **user recommender** produces a set of tags that were used by the user prior to the current post. Issues related to the construction of user profiles (i.e., import of posts, possible change of user interests) make a simple frequency value not a good score for user profile based recommendation. Tags most likely to be reused are the ones that were steadily assigned to posts while the user profile was built. To capture these tags we counted the number of separate days in which a tag was used by the user. To obtain the tag score we divided the number of days the tag was used by the total number of days in which the user was adding posts to the system. This approach allows a decrease in the importance of tags that were assigned by the user in a short period of time only; however, it only partially solves the problem of imported posts. For some of imported posts the system automatically produces low-quality tags and assigns time stamps copied from an external repository (e.g., importing web browser bookmarks, the system copies the time they were created). The combination of artificial tags and real time-stamps makes these posts very hard to detect. Removing such artificial posts is likely to improve the accuracy of the user profile recommender in a real recommendation system; however, it can have undesired consequences when applied to the challenge datasets. If the user imported posts before both training and test data were collected, it is possible that some of them can be found in both datasets. Hence we should train the system for tags from these posts, because it is possible that they can be found in test data as well. Even if we modify the frequency score the representation of user profile still contains tags related to various user interests. Checking the tags extracted from user profile against resource related tags allows us to extract tags that are particularly important for the processed posts. The intersection of both sets of tags produces tags related both to user as well as resource. The score of a tag is the product of scores from both source sets.

Finally the results of title recommender, resource recommender and the intersection of resource related tags and user profile are merged. As all three sets are results of independent recommenders, tags must be rescored to ensure that tags from more accurate recommenders will have higher score in the final tag recommendation set. Again the *leading precision* rescorer was used for the three input tag recommendation sets. The top ten tags of this set create the final recommendation set. The challenge organizers proposed to limit the recommendation set size to five tags, which seems to be a good number to be presented to a user, however, for evaluation purposes it is interesting to observe more tags.

6 Evaluation

This section presents the results of the off-line system evaluation based on the available BibSonomy snapshot. The evaluation approach assumed that all and only relevant tags were given by the user. Although this method simplifies the problem, it is robust and objective. The quality metrics were precision and recall, commonly used in recommender system evaluations [3].

6.1 Methodology

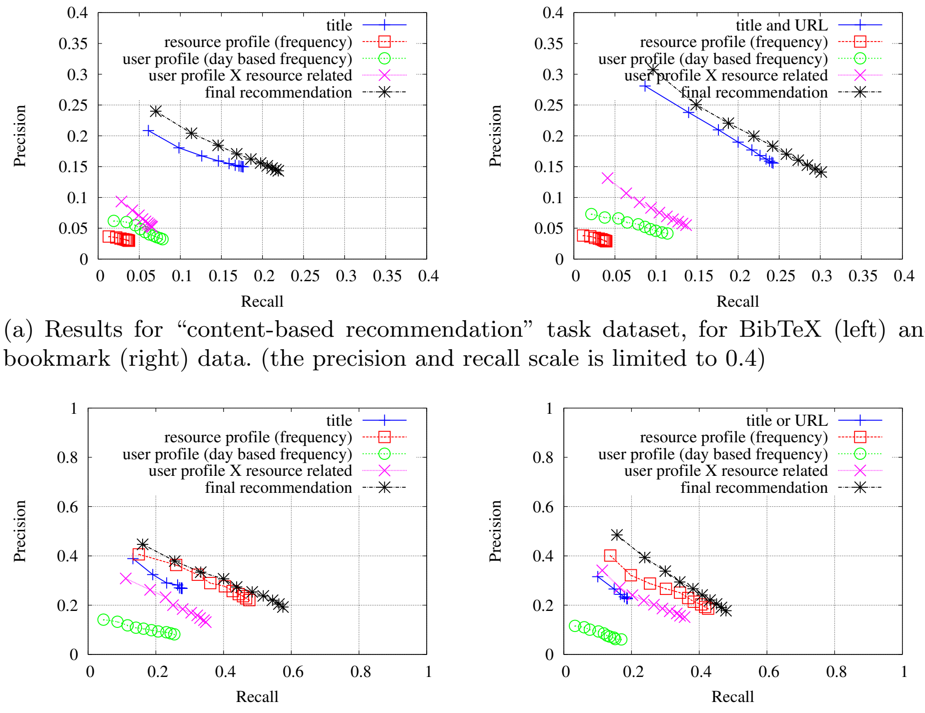
To keep the list of correct tags secret during the contest the organizers kept strict division between training and test set. The test data contained posts entered to to BibSonomy between January 1, 2009 and June 30, 2009. Each post of which user, resource and all tags could be found in k-core of order 2 of training data was used as test post for the "graph-based recommendation" task. The remaining posts were used for the "content-based recommendation" task. Comparison of training and test data for both tasks is presented in Table 3.

As we decided to separate the processing of BibTeX and bookmark posts we present the results for two post types separately. The final recommendation is presented together with the intermediate steps of the system: tags extracted from the resource title (and URL), the most frequent tags from resource profile and user profile and the combination of resource related tags and user profile tags. As each tag from the tag recommendation set can be ranked by its score it is straightforward to present any selected number of recommended tags. The plots (Fig. 6.1) present consecutive results for the top n tags, where $1 \leq n \leq 10$. For the "graph-based recommendation" task the tags that could not be found in

	training	test - Task 1	test - Task 2	test total
BibTex	158,924	26,104 (98.7% of test total)	347 (1.3% of test total)	26,451
bookmark	263,004	16,898 (97.5% of test total)	431 (2.5% of test total)	17,329
posts total	421,928	43,002 (98.2% of test total)	778 (1.8% of test total)	43,780

Table 3. Number of posts in training and test dataset. Sparsity of folksonomy graph causes large disproportion between test set for "content-based recommendation" task (Task 1) and "graph-based recommendation" task (Task 2). Another interesting fact is a different ratio of BibTeX and bookmark posts in training and test data.

the k-core of training data were removed from each recommendation set before calculating precision and recall.



(a) Results for "content-based recommendation" task dataset, for BibTeX (left) and bookmark (right) data. (the precision and recall scale is limited to 0.4)

(b) Results for "graph-based recommendation" task dataset, for BibTeX (left) and bookmark (right) data.

Fig. 4. Precision and recall of proposed tag recommendation system and intermediate steps. Test data was divided into BibTeX and bookmark part.

6.2 Results

As expected, precision and recall of the recommendation results in the "content-based recommendation" task are mostly driven by the content tags. Low score of user profile recommenders for BibTeX data is likely caused by a large number of posts by users who started to use the system after the training set was built. According to the rules set by the organizers the precision score was averaged over all posts in the test set, even if a recommender returned no tags for some of them. Whenever a user profile was available the user based recommender obtained significantly better results than content based recommender only.

The results for the "graph-based recommendation" task show surprisingly high accuracy of resource profile tags (which was not observed to such a degree on training data). For the test dataset in this task the intersection of resource related tags and user profile has lower precision than resource profile tags. This is an unexpected result, comparing to the previous results on training dataset, where the intersection of resource related tags and user profile had comparable or higher precision and recall to resource profile. Despite this unexpected behaviour the tags from the user profile are able to increase the f1 score by 0.02 for tag recommendation set of size 5. The open question is how representative the results of this dataset are, considering the fact that less than 2% of test posts matched the conditions of this task.

For both tasks there is a noticeable difference between the results for both types of data. However, it is not clear if it is caused by some fundamental differences between BibTeX and bookmark posts, or the differences between the two particular test datasets used. It is important to notice that the high number of tested posts has no impact on the statistical validity of results. The way the test data was prepared makes it very dependent on the behavior of users in the period of time the data was collected.

Finally we present the results of the final recommendation for combined BibTeX and bookmark posts, which were submitted to the challenge (Table 4). The systems were ranked based on the f1 score (Eq. 5) for the tag recommendation set of size 5. Based on that criterion the presented tag recommendation system took the first place in the "content-based recommendation" task (out of 21 participants) and the third place in the "graph-based recommendation" task (again, out of 21 participants).

$$f1 = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

7 Conclusions and future work

In creating the presented tag recommendation system we considered the title of a resource as a natural starting point of the recommendation process. We tried to extend the set by tags related to the title as well as tags present in the profiles of resource and user. Our main aim was to extract valuable tags

Table 4. The results of the presented tag recommendation system. In the challenge the systems were ranked based on the f1 score for the tag recommendation set of size 5.

	content-based recommendation			graph-based recommendation		
#result tags	recall	precision	f1	recall	precision	f1
1	0.0805	0.2664	0.1236	0.1587	0.4679	0.2370
2	0.1275	0.2224	0.1621	0.2465	0.3869	0.3012
3	0.1626	0.1987	0.1788	0.3143	0.3361	0.3248
4	0.1885	0.1821	0.1852	0.3682	0.2998	0.3305
5	0.2080	0.1705	**0.1874**	0.4070	0.2700	**0.3246**
6	0.2218	0.1616	0.1869	0.4425	0.2468	0.3169
7	0.2323	0.1549	0.1858	0.4701	0.2282	0.3072
8	0.2403	0.1495	0.1843	0.4929	0.2116	0.2961
9	0.2467	0.1457	0.1832	0.5092	0.1967	0.2838
10	0.2515	0.1424	0.1819	0.5220	0.1842	0.2723

from user profile which is a very rich but imprecise source of tags. Designing the system we mostly focused on the precision of the recommended tags. To avoid the risk of recommending tags less precise than tags extracted from the title we decided to leave it as the only recommendation whenever the user profile was unavailable. This was a frequent case in "content-based recommendation" task, which gives us hope that the system will be able to achieve even better results for the final "on-line recommendation" task. The system is now connected to BibSonomy and recommends tags to each newly added post in real time. This evaluation setting will give a realistic assessment of system quality.

In our future work on this project we plan to focus on tagging patterns of individual users which would allow us to tune the recommendation for each specific user. Discovering strong patterns, like user who uses author name and year of publication for each BibTeX post, can greatly increase the accuracy of recommender for this specific user. Another interesting issue is handling of multi-word concepts (e.g., is a user going to use two tags "information" "retrieval" or one "information.retrieval"?). Finally, we hope that evaluation settings like "on-line recommendation" task would allow us to investigate short temporal patterns when a user adds a sequence of posts related to the same problem.

References

1. Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.
2. V. Batagelj and M. Zaveršnik. Generalized cores, 2002. cite arxiv:cs.DS/0202039.
3. Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.
4. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proc. the First*

Conceptual Structures Tool Interoperability Workshop at the 14th Int. Conf. on Conceptual Structures, pages 87–102, Aalborg, 2006. Aalborg Universitetsforlag.

5. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006. Springer.
6. Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4702 of *LNCS*, pages 506–514. Springer, 2007.
7. Sigma On Kee Lee and Andy Hon Wai Chun. Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ANN semantic structures. In *ACOS'07: Proc. the 6th Conf. on WSEAS Int. Conf. on Applied Computer Science*, pages 88–93, Stevens Point, Wisconsin, USA, 2007. WSEAS.
8. Marek Lipczak. Tag recommendation for folksonomies oriented towards individual users. In *Proceedings of the ECML/PKDD 2008 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008.
9. Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proc. the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.
10. S.C. Sood, K.J. Hammond, S.H. Owsley, and L. Birnbaum. TagAssist: Automatic tag suggestion for blog posts. In *Proc. the International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.
11. Marta Tatu, Munirathnam Srikanth, and Thomas DSilva. RSDC08: Tag recommendations using bookmark content. In *Proceedings of the ECML/PKDD 2008 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008.