Time based Tag Recommendation using Direct and Extended Users Sets

Tereza Iofciu and Gianluca Demartini

L3S Research Center
Leibniz Universität Hannover
Appelstrasse 9a D-30167 Hannover, Germany
{iofciu,demartini}@L3S.de

Abstract. Tagging resources on the Web is a popular activity of standard users. Tag recommendations can help such users assign proper tags and automatically extend the number of annotations available in order to improve, for example, retrieval effectiveness for annotated resources. In this paper we focus on the application of an algorithm designed for Entity Retrieval in the Wikipedia setting. We show how it is possible to map the hyperlink and category structure of Wikipedia to the social tagging setting. The main contribution is a time-based methodology for recommending tags exploiting the structure in the dataset without knowledge about the content of the resources.

1 Introduction

Tagging Web resources has become a popular activity mainly due to the availability of tools and systems making it easy to tag and also due to the advantage users see in tagging their resources. People can for example get better search results, or they can get new resources recommended based on tags other people assigned. One particular problem is the one of recommending relevant tags to users for resources they have introduced in the system.

Being able to effectively recommend tags would, firstly, simplify the tasks of the users on the web who want to tag resources (e.g., bookmarks, pictures, ...), and, secondly, would allow an automatic annotation of resources that enables, for example, a better search for resources or an improved resource recommendation.

When we want to assign a tag to a resource (or, to predict which tag a user would assign to a resource) a possible approach is to use the most popular tags for the given resource of the given user. Of course, this is not working well because users can tag resources which are different and people tag the same resource in different ways. For this reason most effective approaches look at the content of the resources and perform more complex analysis of the structure connecting users, resources, and tags.

Previous approaches focus on the content for resources (e.g., textual content of a web page) or on the structure of the tripartite graph composed of users, resources, and tags. The approaches we propose in this paper do not take into account the content of the resources but only the connection structure in

the graph. Additionally, we put more importance on more recent tags with the assumption that users' interests might change over time.

We adapt an algorithm proposed for ranking entities in Wikipedia [1] based on a set of initial relevant examples (e.g., already tagged resources) and on the structure of hyperlinks connecting pages and categories containing them. As we defined hard links between documents and categories they belong to and soft links between documents and categories containing linked documents, so we define these types of links between resources/users and tags in the tag recommendation setting.

The rest of the paper is structured as follows. In Section 2 we describe the proposed algorithms also showing the correspondence to the Wikipedia setting. In Section 3 we describe the experimental setting and results. In Section 4 we compare our work with previously proposed approaches and, finally, in Section 5 we conclude the paper.

2 Graph Based Algorithms

In this section we describe the algorithms we designed and used for the graph based task that have been run at Discovery Challenge (DC) 2009.

2.1 Using the Resource-User Graph

In both submitted approaches, starting from the input *query post* (i.e., the posts from the test file) we retrieve the resource it refers to. We call this resource the *query resource*. For the query resource we retrieve, using the train data, all the users that have annotated it in different posts. We call this set of users the *direct user set*. We then use this set of user as an input for the algorithm and retrieve all tags the users have assigned. In the second algorithm, in addition to the set of direct users, we also retrieve the user neighborhood (i.e., users that used at least once a tag in common with the given user). We then use the reunion of the two user sets as input for recommending tags. We call the reunion of the two user sets, the *extended user set*. As a third approach we have also retrieved just the tags that have previously been assigned to the resource as baseline for comparison.

As seen in Figure 1, by traversing the post - resource - users graph, we obtain the set of direct users that have annotated the resource given in the query post. The extended user set is obtained by adding also the neighborhood users to the direct user set, see Figure 2. We considered two users as being neighbors if they had common tags.

As a baseline approach we considered the recommendation of the most popular tags for a resource, where we only kept the tags assigned by the *direct users* to the resource of the query post.

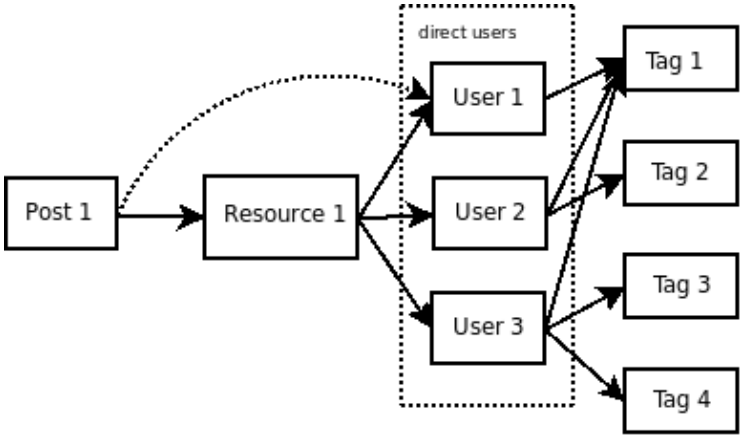

Fig. 1. Tags recommended based on the set of users who have annotated the query resource.

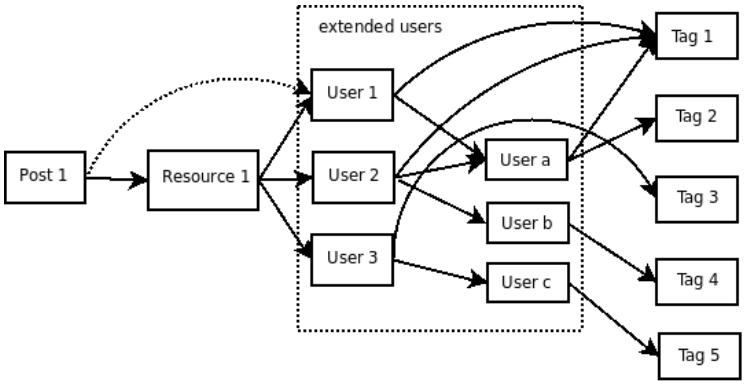

Fig. 2. Tags recommended based on the set of users who have annotated the query resource and users in the immediate neighborhood of the *direct user set*.

2.2 Comparison to the Wikipedia scenario

The algorithms described in this paper are adapted from those developed for finding relevant results for Entity Retrieval queries in the Wikipedia Setting [1]. This work was performed in the context of the Entity Ranking track at the evaluation initiative INEX 2008 [2]. In the following we describe how we can map the Entity Ranking setting with the tag recommendation one.

In the Wikipedia setting we have as input a set of example entities. The goal is to extend such set with other relevant entities. If, for example, the initial set for the query "European Countries" contains Italy, Germany, and France, then the goal is to extend this list with entities such as Spain, Slovenia, Portugal, ... Our approach is to retrieve other entities based on common assigned Wikipedia categories. We extract two sets of categories, hard categories as direct categories (similarly to the *direct user set*) and soft categories from the neighboring entities (i.e., following hyperlinks between Wikipedia articles). As neighboring entities we considered the most frequent entities the example entities linked to (similarly to the *extended user set*). In the Wikipedia setting entities link to entities via hyperlinks, and each entity has several categories assigned to it.

2.3 Time dependent tag ranking

Following the intuition that tags can get outdated over the years, and, thus, older assigned tags should be weighted less for recommendation, we introduced a time decaying function of posts. Scores are assigned to posts based on the time when they have been issued compared to the time the latest test post has been issued. The time decaying function is defined by the following formula:

$$postScore_i = \lambda^{\Delta Time_i} \tag{1}$$

with the decaying factor lambda being smaller than 1 and the time difference being calculated in years. The tag scores are computed based on the tag specificity (i.e., how often they have been assigned) defined as:

$$tagSpecificity_i = \log(50 + tagCount_i) \tag{2}$$

Given the different user sets for a query post, we extract from the training data the most frequent common tags the users have assigned. The tag score is computed based on the formula:

$$tagScore_i = \frac{\sum_j (postScore_j)}{tagSpecificity_i} \tag{3}$$

where a post j was considered only if it was posted by one of the users from the *direct user set* for the first approach and from the *extended user set* for the second approach. The tags are sorted based on this score and the top five tags are kept and recommended.

As a baseline, we ranked the tags based on popularity within the resource (i.e., how often a tag has been assigned to a resource) also keeping into account when they had been assigned to the resource, based on the formula:

$$tagScore_i = \sum_j (postScore_j) \tag{4}$$

3 Experiments

Experiments were performed on the DC 2009 benchmark[1] in order to evaluate the proposed algorithms.

Starting from the query posts in the test file we recommended for each post the top five tags using the two described approaches and the baseline. In Figure 3 it is possible to see effectiveness values for the two approaches when a different number of retrieved tags is considered. We can see that the direct user approach performs better. Figure 4 shows the same result with Precision/Recall curves of the two proposed approaches.

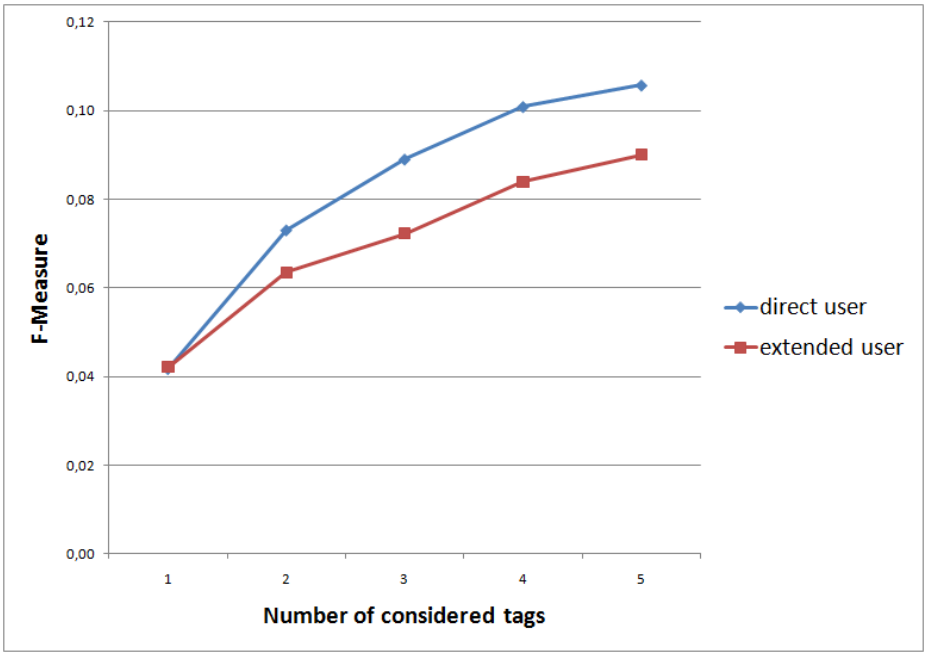

Fig. 3. F-Measure values for *Direct user* approach and *Extended user* approach (λ=0.9)

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

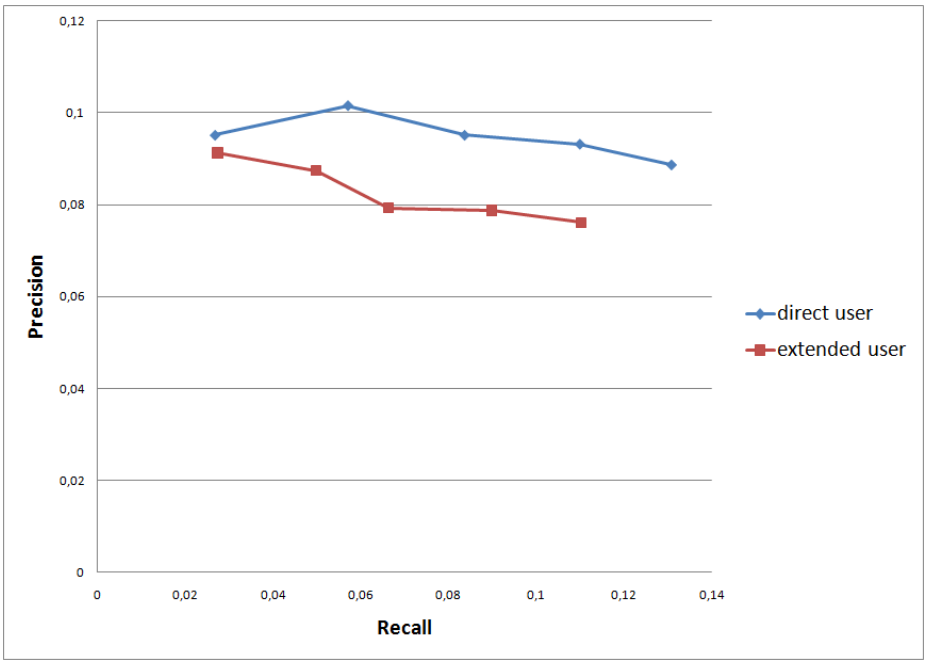

Fig. 4. Precision/Recall curves for *Direct user* approach and *Extended user* approach (λ=0.9)

In Figures 5 and 6 we measure the impact of using the time information when recommending the most popular tags for a resource. With a value of 0.9 for λ, in the time decaying function, the scores were slightly lower than when using just the popularity information (Figure 5). When using a value of 0.95 for λ, there is a small improvement over the baseline when considering 4 and 5 tags (see Figure 6). We ran experiments also with values smaller than 0.9 for λ which have shown that Precision and F-measure decrease quite a lot (3% for F-measure with $\lambda = 0.1$).

4 Related Work

Previous work on tag recommendation mainly distinguish between those looking at the content of the resources and those looking at the structure connecting users, resources, and tags.

Approaches looking at content of resources for tag recommendations are, for example, [5] which looks at content-based filtering techniques. In [6] the authors also look at collaborative tag suggestion in order to identify most appropriate tags.

A specific area of this field looks at recommending tags focusing on an individual user rather than providing general recommendation for a resource. In [4] they first create a set of candidate tags to be recommended and then they

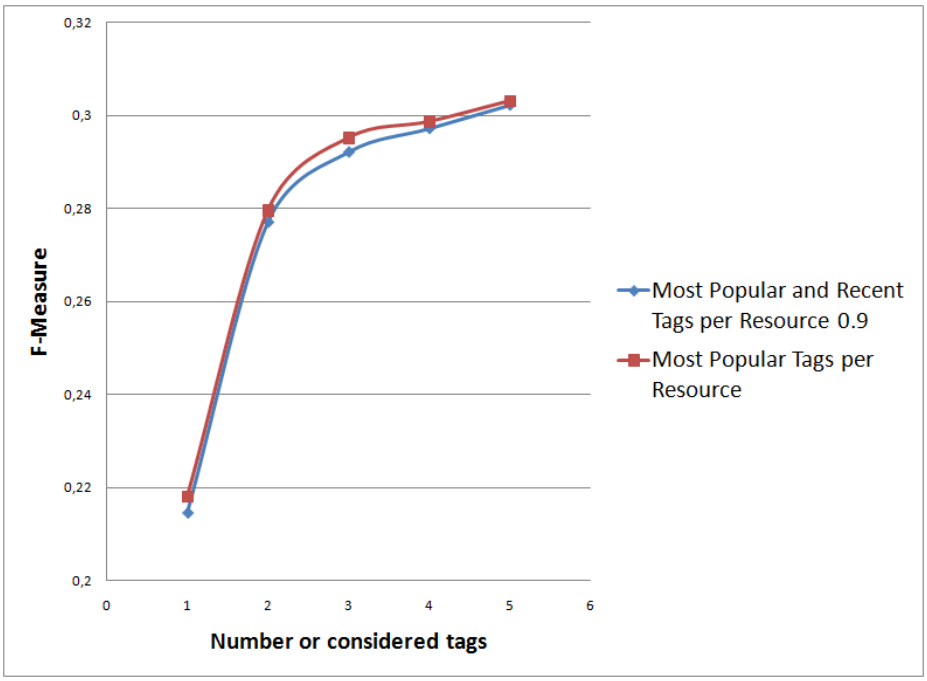

Fig. 5. F-Measure values for *Most Popular Tags per Resource* approach and *Most Popular and Recent Tags per Resource* approach (λ=0.9)

filer it based on the previous tag a particular user has assigned in the past. In [3] the FolkRank algorithm is evaluated and compared with simpler approaches. This is a graph based approach that computes popularity scores for resources, users, and tags based on the well-known PageRank algorithm exploiting the link structure. The assumption is that resources which are tagged with important tags by important users becomes important themself. Similarly to FolkRank, our approach exploits the link structure between users, resources, and tags, but rather looks at the vicinity of a post (i.e., a [resources,user] pair) in order to compute a weight for the most appropriate tags.

5 Conclusions and Further Work

In this paper we presented our first approaches for tag recommendation using graph information. We proposed two approaches, where, given a query post, we retrieve two sets of users. Based on the tags assigned by users in these sets we recommend new tags. The first set of users, the direct user set, consists of the users that have tagged the resource referred to by the query post. The second set of user, the extended user set, consists of the direct user set as well as the users who are neighbours based on commonly assigned tags to the users in the direct set. The tag scores have been computed keeping into account also the time when they have been assigned.

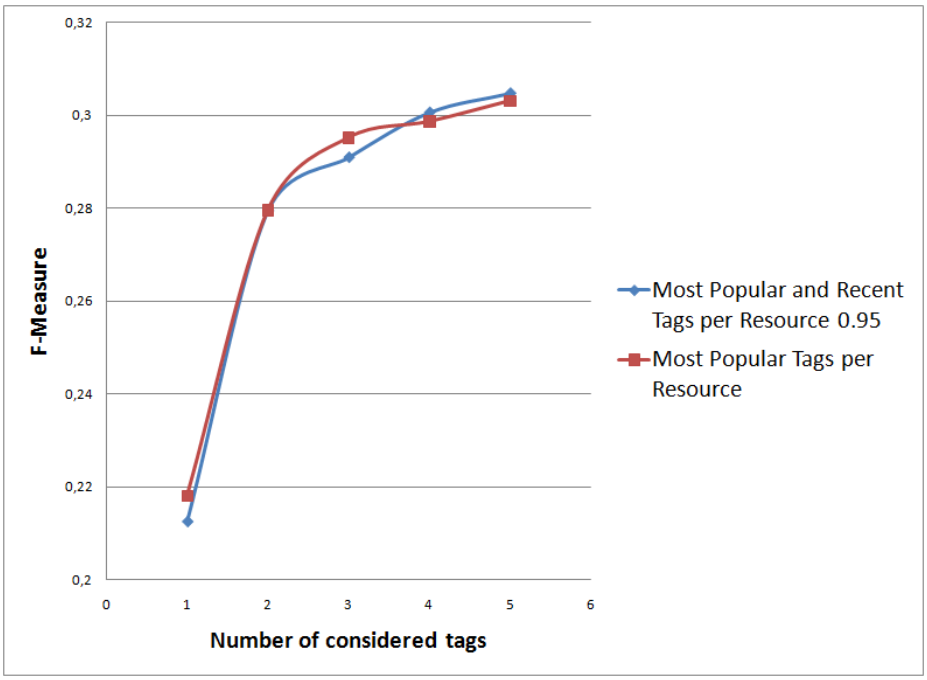

Fig. 6. F-Measure values for *Most Popular Tags per Resource* approach and *Most Popular and Recent Tags per Resource* approach (λ=0.95)

With the proposed approaches, we evaluated the effect of the tag posting time. We compared a time dependent ranking to a tag popularity. In the future, we aim at giving a higher importance to the user given in the query post than to the rest of the direct users.

Acknowledgments. This work is partially supported by the EU Large-scale Integrating Projects OKKAM[2] - Enabling a Web of Entities (contract no. ICT-215032), and LivingKnowledge[3] (contract no. 231126)

References

1. Nick Craswell, Gianluca Demartini, Julien Gaugaz, and Tereza Iofciu. L3s at inex 2008: Retrieving entities using structured information. In *INEX*, 2008.
2. Gianluca Demartini, Arjen P. de Vries, Tereza Iofciu, and Jianhan Zhu. Overview of the inex 2008 entity ranking track. In *INEX*, 2008.
3. Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag recommendations in social bookmarking systems. *AI Commun.*, 21(4):231–247, 2008.

[2] `http://fp7.okkam.org/`
[3] `http://livingknowledge-project.eu/`

4. Marek Lipczak. Tag recommendation for folksonomies oriented towards individual users. In *Proceedings of ECML PKDD Discovery Challenge (RSDC08)*, pages 84–95, 2008.
5. Gilad 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*, pages 953–954, New York, NY, USA, 2006. ACM.
6. Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, 2006.