The Art of Tagging: Measuring the Quality of Tags

Ralf Krestel and Ling Chen

L3S Research Center
Universität Hannover, Germany
krestel|lchen@L3S.de

Abstract. Collaborative tagging, supported by many social networking websites, is currently enjoying an increasing popularity. The usefulness of this largely available tag data has been explored in many applications including emergent semantics deriving, web resources categorization, and web search etc. However, since tags are supplied by users *freely*, not all of them are useful and reliable, especially when they are generated by spammers with malicious intent. Identifying tags of high quality, therefore, is critical in improving the performance of applications based on tags. In this paper, we propose TRP-Rank (Tag-Resource Pair Rank), an algorithm to measure the quality of tags by employing a *quality propagating* technique. The three dimensional relationship among users, tags and web resources is firstly represented by a graph structure. A set of seed nodes, where each node represents a tag annotating a resource, are then selected and their quality is assessed. The quality of the remaining nodes is calculated by propagating the known quality of the seeds through the graph structure. We evaluate our approach on a public data set where bad tags generated by suspicious spammers are manually labelled. The experimental results demonstrate the effectiveness of this approach in measuring the quality of tags.

1 Introduction

With the recent rise of Web 2.0 technologies, many social media applications like *Flickr*, *Del.ici.ous*, and *Last.fm* provide features which allow users to assign tags [1] to a piece of information such as a picture, blog entry, video clip etc. Web users from different backgrounds tag (annotate) resources on the Web at an incredible speed, which results in large volume of tag data obtainable from the Web today. The hidden value of tag data has been explored in many applications. For example, T.-Sutter et al [2] incorporated tags into collaborative filtering algorithms to enhance recommendation accuracy. In [3], the authors discussed using tags to lighten the limitation of the amount and quality of anchor text to improve enterprise search. The usage of tags in Web search has also been investigated in [4].

One notable reason which supports the increasing popularity of collaborative tagging is that users are permitted to enter tags at will, without referring to any pre-specified taxonomy or ontology. On one hand, this easy and flexibility utility boosts the widespread of collaborative tagging systems. On the other hand, allowing users to *freely* choose tags lead to the poor quality of tag data sometimes. For example, ambiguity and synonymy are two frequently cited problems. The tag "XP" is used to annotate both web pages about "Extreme Programming" and pages about "Windows XP". Synonymous tags, like "RnB" and "R&B", are also widely used. Such problems limit the applications built upon tags. Another problem which even damages the performance of applications using tags is tag spam, which refers to misleading tags generated maliciously in order to increase the visibility of some resources or simply to confuse users. *Therefore, measuring the quality of tags is an important problem so that good tags can be identified for effective applications.*

In [5], the authors discussed some properties a good tag combination (e.g., the set of tags annotating a common resource) should possess. For example, a good tag combination should cover multiple facets of the tagged resource; the set of tags should be used by a large number of people; and the number of resources identified by the tag combination should be small etc. They further proposed a tag suggestion algorithm based on the set of properties. In contrast to suggesting tags to users, our objective here is to assess the tags used by users. Koutria et al [6] proposed to combat tag spam by rank the results, returned from a query tag, based on the frequency co-occurrence between the tag and each resource. Thus, their approach is specially designed for tag based search. Our research objective is more general so that the results can be used in various applications of tags.

Note that, whether a tag is good or not makes sense only if the tag is assessed with respect to a particular resource. Hence, in our work, we study based on the unit of a tag-resource pair. We aim to measure the quality of each individual pair of tag and resource. For this purpose, we firstly construct a graph which models tag-resource pairs as nodes and co-user relationship as edges. We then select a set of seed nodes whose qualities are assessed manually. The qualities of the remaining nodes are calculated by propagating the qualities of seed nodes over the graph. In order to improve the performance of this approach, a set of various seed selection strategies are employed. We evaluate the effectiveness of our approach on a bibsonomy data set[1] labelled manually.

The rest of this paper is organized as follows. We discuss the background knowledge by reviewing related work in Section 2. In Section 3, we describe the approach which propagates the quality of tag-resource pairs and discuss improving the performance by employing different strategies to select a set of seeds. The evaluation results conducted on a public data set are presented and analyzed in Section 4. Finally, Section 5 concludes this paper with some summary remarks and future work discussions.

[1] http://www.kde.cs.uni-kassel.de/ws/rsdc08/dataset.html

2 Related Work

In this section, we review related work in two areas, collaborative tagging systems and spam detection.

A collaborative tagging system allows users of a web site to freely attach to a particular resource arbitrary tags which, in the opinion of the user, are somehow associated with the resource in question. The commonly noted structure of collaborative filtering systems is a tripartite model consisting of users, tags and resources. This model is developed as a theoretical extension of the bipartite structure of ontologies with an added "social dimension" in [7]. The dynamics of collaborative systems are examined in [8] using the tag data at the bookmarking site Del.ici.ous. According to this work, tag distributions tend to stabilize over time. Halpin et al. confirm these results in [9] and show additionally that tags follow a power law distribution. Considering the structure and stable dynamics of collaborative tagging systems, it seems likely that tag data would be a reliable source of semantic information reflecting the cultural consensus of a particular system's users. As a result, various applications of tag data have been researched. Mika [7] investigates the automatic extraction of ontological relationships from tag data and proposes the use of such emergent ontologies to improve currently existing ontologies which are less capable of responding to ontological evolution. Dmitriev et al. [3] explore the use of "annotations" for enterprise search to compensate for the lack of sufficient anchor text in intranet environments. In [4], tag data is exploited for the purpose of web search through the use of two tag based algorithms: one exploiting similarity between tag data and search queries, and the other utilizes tagging frequencies to determine the quality of web pages. Tso et al [2] incorporate the tag data into the collaborative filtering systems. Berendt and Hanser [10] demonstrate the benefits of using tag data for weblog classification by treating it as content instead of meta data.

Everywhere in the internet where information is exchanged, malicious individuals try to take advantage of the information exchange structure and use it for their own benefit. The largest amount of spam and historically the first field where spam was generated is the electronic communication system called e-mail. Afterwards, various internet applications are attacked by spammers such as search engine spam, blog spam, wiki spam etc, which triggered numerous research efforts in spam combating. For example, TrustRank [11] separates spam pages from non-spam pages based on the intuition that trustworthy pages usually link to also trustworthy pages and so on. They select a seed set of highly trusted pages first and then propagate the trust score of seed pages by following the links from these pages through the Web. A survey of approaches fighting spam on social web sites can be found in [12]. Comparing to spam detection from other web applications, studies on detecting spam from collaborative tagging systems are very limited. Koutrika et al [6] propose to combat spam in the particular situation when users query for resources annotated with certain tags. Their method ranks a resource higher if more users annotate it the queried tags, based on the assumption that tag spam may not be used by the majority. Our work is different in the way that our approach is not designed for a par-

ticular application. Consequently, the output of our algorithm can be used by any application based on tags. Xu et al [5] assign authority scores to users, and measure the goodness of each tag with respect to a resource by the sum of the authority scores of all users who have tagged the resource with the tag. Then, the authority scores of users are computed via an iterative algorithm similar to HITs [13]. Their approach treats every tag-resource pair used by a user equally even if a spam user may use good tag-resource pairs frequently and bad ones occasionally. Our approach addresses this problem by measuring the quality of every two tag-resource pairs used by a user independently.

3 Measure Tag Quality

The main purpose of tags is to support users in search tasks. But since there is no limitation on the vocabulary users are allowed to use for taggging, some tags describe a resource better than others and are thus more useful for search engines. Measuring the quality of a tag assigned to a resource can help search engines to improve their performance. Another challenge is the context in which a tag is used. Besides of tags that are obviously not useful for search like "myFavorite", "funny", "home", and other *personal*, or *subjective* tags (see [14, 15] for classification), there are tags that describe one resource very well, but are not suitable for other resources. This makes it necessary to always look at tag-resource pairs instead of only tags. Other approaches want to assess users instead of tags, for example to detect malicious users within a tagging system. Our approach can also be used to judge users by looking at their tag assignments. We will show in Section 4 that our approach can be used to identify spam users. The focus lies nevertheless on measuring the quality of a particular tag for a single resource and we argue that for search application the knowledge about malicious users is of less relevance than the knowledge of good tag assignments to resources.

3.1 Problem Specification

Given a set of tags $\mathcal{T}$, a set of resources $\mathcal{R}$, and a set of users $\mathcal{U}$, we define

- the tag assignment of tag $t \in \mathcal{T}$ to resource $r \in \mathcal{R}$ by user $u \in \mathcal{U}$ as tr_u and
- all tag assignments of a user u as $\mathcal{TR}_u$.
- $tr_{\text{res}} = r$ and $tr_{\text{tag}} = t \; \forall tr \in \mathcal{TR}$.
- $\mathcal{TR}$ is thus defined as $\mathcal{TR} = \cup_{u \in \mathcal{U}} \mathcal{TR}_u$.
- Further, our goal is a function to get a quality score for each tr: qual(x) $\forall x \in \mathcal{TR}$.
- To enable search engines to use the algorithm, we need two more functions:
- getTR$(x) = \{tr \subset \mathcal{TR} \mid \text{getR}(tr) = x \; \forall x \in \mathcal{R}\}$ and
- getR$(x) = \{r \subset \mathcal{R} \mid r = x_{\text{res}} \; \forall x \in \mathcal{TR}\}$.

3.2 Tagging System Model

We model a tagging system as a bidirected weighted graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of vertices with each $v \in \mathcal{V}$ represents a $tr \in \mathcal{TR} = \mathcal{T} \times \mathcal{R}$ and $\mathcal{E}$ a set of edges representing the co-occurence of tag-resource pairs tr for a user u. This means we have edges between the nodes tr_u^i and tr_u^j $\forall i, j \in \{1, \ldots, |\mathcal{TR}_u|\}, i \neq j$; $\forall u \in \mathcal{U}$. The weight $w_{1,2}$ of an edge $e_{1,2}$ between v_1 and v_2 is the number of users which have used both tag-resource assignments v_1 and v_2.

In Figure 1 we present a very simple tagging scenario: Suppose we have three users, three different tags and two resources. For example the tags "books", "cds", and "shopping", and as resources the URL of an online book store and a public library. Each user has a set of tagged resources, e.g. user u_1 has tagged resource r_1 with the tags t_1 and t_2, and the resource r_2 with tag t_1, which means $\mathcal{TR}_{u1} = \{A, B, D\}$.

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\};\ \mathcal{V} \in \mathcal{T} \times \mathcal{R};\ \mathcal{E} \in \mathcal{U}$
$\mathcal{U} = \{u_1, u_2, u_3\}$
$\mathcal{T} = \{t_1, t_2, t_3\}$
$\mathcal{R} = \{r_1, r_2\}$
$\mathcal{V} = \{A, B, C, D, E\}$

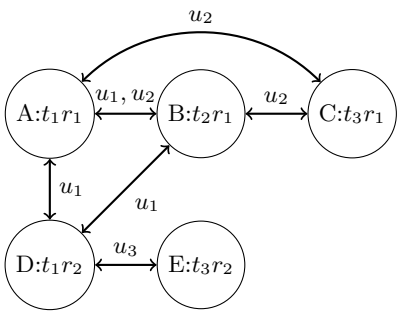

Fig. 1. A simple tagging scenario.

Based on this graph, we introduce a left stochastic transition matrix M, which is defined as:

$$M(i,j) = \begin{cases} 0 & \text{if } i,j \notin \mathcal{E} \\ \frac{w_{i,j}}{\sum_{n=0}^{|\mathcal{V}|} w_{i,n}} & \text{if } i,j \in \mathcal{E} \end{cases}$$

Since we have a bidirected graph this matrix is symmetric. Figure 2 shows the adjacency and transition matrix for our example.

3.3 Quality Propagation

The idea of TRP-Rank is, like TrustRank [11] does for web pages, to manually assign quality scores to a subset of $\mathcal{TR}$, and propagate these trust values through the graph. The TrustRank algorithm is based on PageRank [16].

PageRank. PageRank is an algorithm that assigns scores to web pages based on link information. When important pages point to a page, this page should also

	A	B	C	D	E
A		2	1	1	
B	2		1	1	
C	1	1			
D	1	1			1
E				1	

$$M = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}$$

Fig. 2. Adjacency matrix (left) and transition matrix M (right) of our example graph.

be considered important. Thus importance information is propagated through the web graph via an iterative process:

$$\text{p-rank} = \alpha \cdot T \cdot \text{p-rank} + (1 - \alpha) \cdot \frac{1}{N} \cdot 1_N.$$

The α is a decay factor, T the transition matrix and N the number of web pages. The transition matrix is not weighted and all web pages get the same initial value. The iteration process goes on until the difference between two consecutive runs' results is negligible.

TrustRank. This formula was extended to identify web spam. Therefore the original PageRank algorithm was altered to be biased towards a seed set of high quality sites d, which were manually assessed with an oracle function $\mathrm{O}(x)$. d is then normalized: $d = d/|d|$ and $\text{trank}_0 = d$.

$$\text{t-rank}_{i+1} = \alpha \cdot T \cdot \text{t-rank}_i + (1 - \alpha) \cdot d.$$

Finding the seed set is done using an inverse PageRank algorithm to identify the nodes from where you can reach lots of other nodes, similar to the idea of Hubs [13], and rank them accordingly. The top-k are then manually assigned values 1, or 0 in case of a spam web site, and these initial values are stored in d.

TRP-Rank. For TRP-Rank $\text{qual}(x)$ is computed using the TrustRank formula to propagate initial tag quality scores through the graph. To allow for different degrees of quality of tag assignments we not only propagate scores for good tag assignments but also for explicitly bad ones. Distrust propagation was suggested as an extension to TrustRank to fight web spam before [17]. For TRP-Rank, we extended the manual seed set assessment to include both, high quality TRP and low quality ones. Therefore we populate our init vector d with:

$$d(i) = \begin{cases} \mathrm{O}(i) & \text{if } i \in SEED \\ 0 & \text{if } i \notin SEED, \end{cases}$$

with $\mathrm{O}(x) \in \{-1, 0, 1\}$ and $SEED$ as defined in Section 3.4. For our small example, the results of TRP-Rank are shown in Figure 3

$$\text{trp-rank}_{i+1} = 0.85 \cdot \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \cdot \text{trp-rank}_i + (1 - 0.85) \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$$

i	A	B	C	D	E
trp-rank(i)	-0.03341879	-0.03341879	-0.16368952	0.180295	0.05023218

Fig. 3. TRP-Rank computation and results after 10 iterations for our example.

3.4 Seed Selection Strategies

There are several seed selection strategies that could be applied. We experimented with three different ones, results can be found in Section 4.4. The two main challenges for seed set selections are finding an appropriate size for the seed set, i.e. having a good trade-off between accuracy on the one hand, and the expensive manual labeling process on the other hand, and on the other hand picking the *right* TRPs for the seed set.

We first computed PageRank scores for each TRP showing the connectivity of each node in the graph. The resulting list with the nodes ordered according to PageRank was the starting point for the three approaches we evaluated:

1. Top-k seed set,
2. Power law distributed seed set, and
3. Linear distributed seed set.

Figure 1 shows the different selection processes. Top-k seed set selection was also used in TrustRank to find highly connected nodes whose quality influences a lot of neighboring nodes. Since our graph is bidirected we can use PageRank directly and don't need to compute inverse PageRank scores.

The seed set size depends on the data and how connected each node is with others. One way to approximate the quality of tags is to assess not the tag-resource pairs directly but to judge users and then label their TRPs accordingly. This reduces the amount of necessary manual annotation significantly but means loosing some accuracy since "good" users might also have some low quality TRPs.

4 Evaluation

To evaluate our algorithm we chose an indirect approach: Since there is no manually annotated corpus – of which we are aware of – that could be used to compare our results for the quality of tags with a gold standard, we used tag data compiled for a competition[2] to detect spam users. This data consists of 221354 tag assignments by 1328 users of the BibSonomy[3] system for publications. Out of

[2] `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`
[3] `http://www.bibsonomy.org`

Algorithm 1 Different Seed Selection Strategies

Input:
N: a set of graph nodes, K ($K < |N|$): the number of seeds
Output:
$SEED$: A set of selected seed nodes

order N as $\hat{N} = \{x_1, x_2, \cdots, x_n\}$ such that $PR(x_i) \geq PR(x_{i+1})$
for each $x_i \in \hat{N}$ **do**
  if Seed Selection with Top-K **then**
    if $|SEED| < K$ **then**
      $SEED = SEED \cup \{x_i\}$
    end if
  end if
  if Seed Selection with Power-Law **then**
    if $i \in \{a_n\};\ a_n = n + \lfloor b^n \rfloor;\ b = e^{\frac{\ln(|N|-K-1)}{K-1}};\ \forall n \in \{0, \ldots, K-1\}$ **then**
      $SEED = SEED \cup \{x_i\}$
    end if
  end if
  if Seed Selection with Linear Function **then**
    if $\exists n \in \mathbb{N} \mid \lfloor an = i \rfloor; a = \frac{|N|}{K}$ **then**
      $SEED = SEED \cup \{x_i\}$
    end if
  end if
end for

these users, 118 were marked manually as spammers and 1210 as non-spammers. 195198 individual tag-resource pairs (TRPs) were identified out of which 132520 TRPs were considered. The other were discarded because they were made by users only having one tag assignment. In Section 4.3 we reduced this number even more. Table 1 shows the number of TRPs and by how many users these were shared. Although these numbers seem to imply that the graph is not highly connected, a rather small seed set is sufficient to reach most of the nodes, as we will show in Section 4.4.

Table 1. Number of users sharing a TRP.

Number of TRPs	175619	15767	2664	641	197	115	55	41	24	75
Shared by # of Users	1	2	3	4	5	6	7	8	9	$\geq$10

We needed to adopt our algorithm to automatically evaluate our results, mapping our quality score for each tag assignment to scores for each users. This was achieved by adding up the individual tag assignment scores for each user:

Given a user u and his tag assignments $\{\mathrm{tr}_u^1, \ldots, \mathrm{tr}_u^n\}$,

then the score for this user is $\mathrm{spamScore}(u) = \frac{1}{n} \sum_{i=0,\ldots,n}^{n} \mathrm{qual}(tr_u^i)$

and $\mathrm{isSpammer}(u) = \begin{cases} 1 & \text{if } \mathrm{spamScore}(u) < 0 \\ 0 & \text{otherwise.} \end{cases}$

Before using the data we did some preprocessing. Since the data set consists of the raw BibSonomy data we had to give IDs to each individual TRP. To allow for the semantic relationship between certain tags, we decided to use stemming and ignore capital letters to assign one ID to a group of tags (e.g. "Book", "book", or "Books").

For the seed set generation we made use of the available user information in the annotated dataset: spammer$(u) \in \{-1, 1\}$. Thus, each $tr \in SEED$ was assigned the score:

$$\mathrm{O}(tr) = \begin{cases} 1 & \text{if } \frac{1}{|\mathcal{U}_{tr}|} \sum_{u \in \mathcal{U}_{tr}} \text{spammer}(u) > 0 \\ -1 & \text{if } \frac{1}{|\mathcal{U}_{tr}|} \sum_{u \in \mathcal{U}_{tr}} \text{spammer}(u) < 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{U}_{tr} = \{u \in \mathcal{U} \mid tr \in \mathcal{TR}_u\}$. This might lead to some false judgements regarding the tag quality of each single tag assgnment done by a user, since also a spam user might have done a few high quality tag assignemnts. To get better results, a manual annotation of each tag-resource pair would have been necessary. To get reliable data, the annotation process would need to be repeated by different persons, because of the subjective character of the notion "tag quality".

In the following experiments we focused on four questions:

1. Is it possible in general to identify spammers using our graph-based approach?
2. Can negative score propagation be used beneficially?
3. Can we make the process more effective by reducing data without loosing too much accuracy?
4. What is a good seed set size and what function is suited for the seed set selection process?

4.1 Graph-based Approach

The data set contains only 9.75% spam users and only 0.88% spam TRPs, where we consider a TRP spam if more spam users than non-spam users use it. To get an idea what we can expect as a maximum for our algorithm we assume to have a seed set consisting of all the tag-resource pairs in the data set: $SEED = trp \in \mathsf{TRP}$. Since our algorithm is meant to assess the quality of tag-resource assignments and not primarily to detect spam users, 100% accuracy can not be achieved. In some cases, due to the properties of the data, we actually prefer to differ from the manual assessment: Suppose we have a user with lots of good TRPs, but one spam TRP. Our algorithm will not identify this user as spammer, since its judgement is based on looking at the individual tags. The manual assessor will mark this user as spammer because the malicious intention of the user is in the foreground.

Nevertheless do we get only 2.03% wrong as shown in Table 2 on top.

4.2 Quality Propagation

We did several experiments evaluating the spread of positve scores through the net and negative scores. If we use all the information we have about spam users and assign negative scores to the TRPs they have we can identify more spammers correctly than using only the positive scores assigned to TRPs used by known non-spam users. Table 2 shows the results for this theoretical scenario having the spammer/non-spammer information for all users. When we only look at one kind of scores – positive or negative – we need to decide what to do with users having a score of 0.0. In general, we consider them as non-spammers. Looking at only positive scores, we consider them as spammers. For higher accuracy they should be considered *unknown* until these users add more TRPs and thus can be captured by our algorithm. For the specific task of identifying users as spammers we have to make a commitment.

Table 2. Confusion matrix for theoretically achievable maximum using differnt kinds of propagation scores.

Positive and Negative spread information			
True Positives:	**1210**	True Negatives	**89**
False Positives:	**29**	False Neagatives	**0**
Only positive spread information			
True Positives:	**1079**	True Negatives	**114**
False Positives:	**4**	False Neagatives	**131**
Only negative spread information			
True Positives:	**1210**	True Negatives	**91**
False Positives:	**27**	False Neagatives	**0**

4.3 Data Reduction

For large data sets the matrix of our algorithm can become very large. To reduce the amount of data to process, we examined the effect of considering only TRPs with tags that were used by at least x users. This seems to be justified at least for the case of measuring the quality of a certain tag for a certain resource. But also for detecting spam users, leaving out these TRPs is an option. The accuracy for the theoretically achievable maximum using the whole data set as seed set drops by only 2.94 % when considering only tags that were used by at least 10 users:

- Minimum 10 Users → Accuracy 94.80 %
- Minimum 3 Users → Accuracy 95.63 %
- Minimum 0 Users → Accuracy 97.67 %

For minimum occurence of 10 this means a reduction of the transition matrix size by more than 50%.

4.4 Seed Set Selection

Selection Functions. The function for seed set selection is highly dependent on the data to analyze. We started using the top-k TRPs as ranked by PageRank. Figure 4 shows the PageRank scores for all TRPs in our data set.

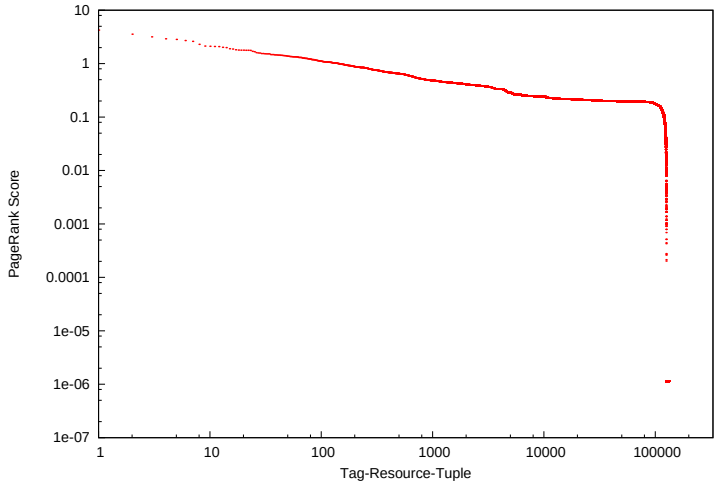

Fig. 4. Log-log graph of PageRank scores for the whole data set.

The PageRank scores of spam TRPs are mostly in the lower area. This leads to the problem that we don't get seeds from the spam partitions of the graph if we only consider top ranked TRPs.

The next approach we evaluated was using a power law distributed seed set, taking advantage of the high connectivity of the higher ranked TRPs but also include possible lower ranked spam seeds.

And as a last selection function we examined a simple linear one, picking every xth PageRank-sorted TRP. The result for the different selection strategies is shown in Table 3. The top-k approach does not find any spam users, even for larger seed sets and thus couldn't outperform a baseline approach regarding all users as non spammers. In [18] this effect was investigated for TrustRank, whose seed set might not cover all topics equally well.

Seed Set Size. The seed set size is a crucial factor for the algorithm. Since we need an oracle function that gives us qual(x) $\forall x \in SEED$ we want our seed set as small as possible. In the general case, the oracle function can be considered a human assessing the $SEED$. This is usually quite expensive and we will discuss alternative approaches in 5.

For the task of identifying spam users, the graph approach limits the achievable accuracy based on connectivity of the TRPs on the one hand, and the

Table 3. Accuracy for different seed set selection strategies with seed set size 10000/20000.

Strategy	Accuracy Seed Set Size	
	10000	20000
Top-k	91.11 %	91.11 %
Powerlaw	94.58 %	96.39 %
Linear	94.88 %	96.31 %

tagging behavior of the users on the other hand. As mentioned before, a user with a lot of high quality tags and one or two spam tags would still be considered a good, non-spam user by our system, but might be considered a spammer by other systems. In Table 4 accuracy is presented for different seed set sizes, also using the whole information about spammers to use all TRPs as seed set.

Table 4. Results for different sized seed sets using linear seed set function. Users with score=0.0 are considered non-spammers (TP=true positives, TN=true negatives, FP=false positives, FN=false negatives).

Seed Set Size	TP	FP	TN	FN	Accuracy
132520	1210	29	89	0	97.82 %
50000	1210	29	89	0	97.82 %
20000	1210	49	69	0	96.31 %
10000	1210	68	50	0	94.88 %
5000	1210	87	31	0	93,45 %

4.5 Discussion

It looks like our algorithm performs quite well on identifying users as spammers based on the quality of their TRPs. After looking at the data into more detail, it seems that our approach could even do better when modifying the definition of "spammer". For example users with only one "test" tag assignment are considered non spammers in the data set. Since these are no malicious users this might be an accaptable classification. From the tag quality point of view these users would be considered spammers because they use bad quality TRPs.

We saw that actually negative score is more effective when looking for spammers. For the quality assessment, using both, positive and negative scores is the best solution. This means we have two seed sets, one containing good TRPs and one containing bad ones. Depending on the data set, and especially on the percentage of spam/low quality TRPs in the data, it is difficult to find negative seeds. In a working tagging system, the majority are good/non-spam tags, and thus negative seeds are ranked rather low by PageRank which makes them hard to find.

Regarding the size of the whole data set, we saw that the accuracy drops only little when putting some restrictions on the tags which are allowed for valid TRPs. Especially for the quality assessment of TRPs and large data sets, it makes sense to only include TRPs that have tags, which are used at least by n different users.

The seed set selection process is also dependent on the data set. To get a well balanced, representative seed set, the linear function provides the best results. A ranking based on PageRank is neccessary in all cases to get seeds from as many partitions of the graph as possible.

5 Conclusions and Future Work

In this paper, we focus on the problem of measuring the quality of tags which are supplied by users to annotate resources on the Web. Due to the intrinsic feature of existing collaborative tagging systems that users are allowed to supply tags freely, the resulting tags can have great disparity in quality. Consequently, measuring the quality of tags appropriately is important toward effectively exploiting the usefulness of tags in many other applications. The main characteristics of our algorithm are represented by the data model we adopted and the seed selection functions we investigated. By decoupling the relationship between users and tag-resource pairs, we model the tag-resource pairs as nodes and co-user relationship as edges of a graph. Different from existing models, this structure allows every two tag-resource pairs used by the same user to have different quality, which complies with the practical situation better. Our algorithm, which propagating quality scores iteratively across the graph, needs to be initialized with the scores of a set of seed nodes. We investigate various seed selection strategies which aims to not only minimize the size of the seed set but also minimize the error of resulted quality scores. The effectiveness of our algorithm is evaluated on a manually labelled data set and demonstrated by the promising experimental results.

For future work we have a couple of ideas on how to pursue:

- Investigate the use of discrete values between -1.0 and 1.0 instead of only 1,0 and -1 for seeds.
- Make use of Web 2.0 and let users generate the seed set.
- Compile a data set to evaluate tag quality directly.
- Model users as nodes and TRPs as edges in the graph to directly find spam users.

6 Acknowledgements

This work is supported by the EU project IST 45035 - Platform for searcH of Audiovisual Resources across Online Spaces (PHAROS).

References

1. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In Wiil, U.K., Nürnberg, P.J., Rubart, J., eds.: Hypertext, ACM (2006) 31–40
2. Tso-Sutter, K.H.L., Marinho, L.B., Schmidt-Thieme, L.: Tag-aware recommender systems by fusion of collaborative filtering algorithms. In Wainwright, R.L., Haddad, H., eds.: SAC, ACM (2008) 1995–1999
3. Dmitriev, P.A., Eiron, N., Fontoura, M., Shekita, E.J.: Using annotations in enterprise search. [20] 811–817
4. Bao, S., Xue, G.R., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. [19] 501–510
5. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: WWW2006: Proceedings of the Collaborative Web Tagging Workshop, Edinburgh, Scotland (2006)
6. Koutrika, G., Effendi, F., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: AIRWeb. (2007)
7. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: International Semantic Web Conference. Volume 3729 of Lecture Notes in Computer Science., Springer (2005) 522–536
8. Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems. CoRR **abs/cs/0508082** (2005)
9. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. [19] 211–220
10. Berendt, B., Hanser, C.: Tags are not metadata, but just more content - to some people. In: ICWSM. (2007)
11. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.O.: Combating web spam with trustrank. In Nascimento, M.A., Özsu, M.T., Kossmann, D., Miller, R.J., Blakeley, J.A., Schiefer, K.B., eds.: VLDB, Morgan Kaufmann (2004) 576–587
12. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing **11**(6) (2007) 36–45
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5) (1999) 604–632
14. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: tagging, communities, vocabulary, evolution. In: Proceedings CSCW, New York, NY, USA, ACM (2006) 181–190
15. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science **32**(2) (2006) 198–208
16. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. techreport (1998)
17. Wu, B., Goel, V., Davison, B.D.: Propagating trust and distrust to demote web spam. In Finin, T., Kagal, L., Olmedilla, D., eds.: MTW. Volume 190 of CEUR Workshop Proceedings., CEUR-WS.org (2006)
18. Wu, B., Goel, V., Davison, B.D.: Topical trustrank: using topicality to combat web spam. [20] 63–72
19. Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J., eds.: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007. In Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J., eds.: WWW, ACM (2007)

20. Carr, L., Roure, D.D., Iyengar, A., Goble, C.A., Dahlin, M., eds.: Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006. In Carr, L., Roure, D.D., Iyengar, A., Goble, C.A., Dahlin, M., eds.: WWW, ACM (2006)