Social Ranking: Finding Relevant Content in Web 2.0

Valentina Zanardi[1] and **Licia Capra**[2]

Abstract. Social (or folksonomic) tagging has become a very popular way to describe, categorise, search, discover and navigate content within Web 2.0 websites. Unlike taxonomies, which overimpose a hierarchical categorisation of content, folksonomies empower end users by enabling them to freely create and choose the categories (in this case, tags) that best describe some content. However, as tags are informally defined, continually changing, and ungoverned, social tagging has often been criticised for lowering, rather than increasing, the efficiency of searching, due to the number of synonyms, homonyms, polysemy, as well as the heterogeneity of users and the noise they introduce. In this paper, we propose a method to increase the efficiency of searches within Web 2.0 that is grounded on recommender system techniques. We measure users' similarity based on their past tag activity. We infer tags' relationships based on their association to content. We then propose a mechanism to answer a user's query that ranks (recommends) content based on the inferred semantic distance of the query to the tags associated to such content, weighted by the similarity of the querying user to the users who created those tags. We evaluate the effectiveness of this mechanism when performing searches on the CiteULike dataset.

1 INTRODUCTION

The advent of Web 2.0 has transformed users from passive consumers to active producers of content. This has tremendously increased the amount of information that is available to users (from videos on sites like YouTube and MySpace, to pictures on Flickr, to music on Last.fm, and so on). This content is no longer categorised according to pre-defined taxonomies. Rather, a new trend called *social* (or *folksonomic*) *tagging* has emerged and quickly become the most popular way to describe, categorise, search, discover and navigate content within Web 2.0 websites.

Unlike taxonomies, which overimpose a hierarchical categorisation of content, folksonomies empower end users by enabling them to *personally* and *freely* create and choose the categories (in this case, tags) that best describe a piece of information (a picture, a blog entry, a video clip, etc.). Tag clouds are then widely used to visualise a set of related tags that best describe either individual items or the content of a website as a whole, with the most frequently used tags being given more importance either in font size or color. Other visualisation techniques have been studied, in order to give more importance to tags' relationships rather than popularity [5, 11]. When users want to find content, they navigate, via hyperlinks, from a tag to a collection of items that are associated with that tag.

However, as tags are informally defined, continually changing, and ungoverned, social tagging has often been criticized for lowering, rather than increasing, the efficiency of searching [3]. This is due to the number of synonyms, homonyms, polysemy, as well as the heterogeneity of users, contexts, and the noise that they introduce.

In order to 'connect' users with content that they deem relevant with respect to their interests, efficient searching techniques have to be developed for this novel and unique domain. By efficient, we mean that the searching technique should be both *accurate* (i.e., the returned content does satisfy users' interests), and *complete* (i.e., if there is relevant content in the system, this should be found).

In this paper, we propose a technique, called *social ranking*, that aims to efficiently find, within a potentially huge dataset, content that is relevant to a user's query. In typical Web 2.0 fashion, we assume such content to have been described with an arbitrary number of tags and by an arbitrary number of users. We begin with a study of the key characteristics of a typical Web 2.0 website (Section 2). Based on these insights, we propose a mechanism to answer a user's query that is grounded on traditional recommender system techniques (Section 3): we measure users' similarity based on their past tag activity; we infer tags' relationships based on their association to content; finally, we rank (recommend) content based on the inferred distance of the query to the tags associated to such content, weighted by the similarity of the querying user to the users who created those tags. Preliminary experimental results demonstrate the good accuracy and coverage of social ranking (Section 4). We position ourselves with respect to other works in the area in Section 5, before discussing our plans for the future (Section 6).

2 DATASET ANALYSIS

In order to understand the key characteristics of the target scenario, and thus develop a query model that is grounded on its peculiarities, we started with the analysis of a typical Web 2.0 website, that is, CiteULike (http://www.citeulike.org). CiteULike is a social bookmarking website that aims to promote and develop the sharing of scientific references amongst researchers. Similarly to the cataloging of web pages within del.icio.us, and of photographs within Flickr, CiteULike enables scientists to organize their libraries with freely chosen tags which produce a folksonomy of academic interests. CiteULike runs a daily process which produces a snapshot summary of what articles have been posted by whom and with what tags. We downloaded one such archive in December 2007. The archive contained roughly 28,000 users, who had tagged 820,000 papers overall, using 240,000 distinct tags. A pre-analysis of the archive revealed the presence of a vast amount of papers and a vast amount of tags bookmarked/used by one user only. In order to make the dataset more manageable, we decided to prune it so to remove those papers and tags that had been bookmarked/used only once over the entire dataset. We were thus left with roughly 100,000 papers and 55,000 distinct tags (while keeping

[1] Dept. of Computer Science, University College London, UK, email: v.zanardi@cs.ucl.ac.uk
[2] Dept. of Computer Science, University College London, UK, email: l.capra@cs.ucl.ac.uk

all 28,000 users). We believe that this pruning of the dataset has not compromised our analysis (and subsequent performance results), as there is little one can do to improve search efficiency on papers/tags nobody else knows.

We then analysed the remaining dataset more carefully in terms of *users' activity*, *papers' popularity*, and *tags' usage*.

Users' Activity. To begin with, we studied how many papers were tagged on average by each user in the system. As expected, there is a huge variance in users' activity, with roughly 70% of the users tagging less than 10 papers (low activity), while the remaining 30% bookmarks between 10 and 50 papers (medium activity), and between 50 and 200 papers (high activity). Note that even users with the most intense activity only bookmark a tiny portion of the whole paper set, thus suggesting a very focused and scoped interest within the broader scientific community.

We also analysed the size of the vocabulary these users spoke, that is, how many different tags were ever used by each user. We found that more than 70% of users only used less than 20 different tags, another 15% of users used between 20 and 60 tags, and the remaining used between 60 and 120 different tags. Once again, the extremely narrow proportion of tags used by each user suggests that user's interest is rather scoped in this domain, so that the vocabulary spoken by each of them is just a tiny proportion of the emerging folksonomy.

Papers' Popularity. We studied papers' relevance next, that is, we quantified how many users had bookmarked the same paper. The vast majority of papers (roughly 87%) were tagged by less than 5 users (low popularity); 12% were tagged between 5 and 15 times (medium popularity), and the remaining 1% more than 15 times (high popularity). This suggests that there is a small subset of highly popular papers who have been bookmarked by a significant proportion of the community, while there is a very long tail of less popular ones.

We also looked at how many different tags were used to describe each paper. 84% of papers had less than 10 different tags associated to them (and more than 54% of them with less than 5). The remaining 16% of papers used between 10 and 30 tags. This would suggest, in accordance with the analysis of users' activity previosuly done, that only a small subset of the whole folksonomy is needed to describe papers (and thus topics) - that is, users and tags are highly clustered around papers/topics.

Tags' Usage. Finally, we studied tags' usage, that is, to what extent the emerging vocabulary is shared among users. The vast majority of tags (roughly 70%) were used by less than 20 users, and an additional 12% by between 20 and 40 users. However, a non negligible 18% were actually shared by more than 40 users. As for papers' popularity, there exists a small subset of tags that are very widely used, and a very long tail of less popular ones.

We also studied how spread was the usage of tags, that is, to how many different papers was a tag associated. Confirming previous observations, we found the vast majority of tags (in excess of 70%) to be used on a tiny proportion of papers (less than 20), another 10% to be used for between 20 and 40 papers, with the remaining being used for more than 40 papers. Despite the huge number of tags in use in the CiteULike folksonomy, tags are thus shared by small communities of users and highly clustered around papers/topics.

2.1 Insights

Based on the dataset analysis summarised above, the following insights can be drawn.

Clustering of Users: users vary a lot in terms of activity; however, even the most active users bookmark a tiny proportion of the whole paper set. This suggests that users have clearly defined interests that map to a small proportion of the whole CiteULike content. This is confirmed by tags' usage: each user masters a small subset of the whole folksonomy, and users sharing part of the folksonomy form small clusters. We formulate the hypothesis that, by looking at users' tag activity, *users' similarity* can be quantified and exploited to answer content searches more accurately.

Clustering of Tags: despite the emergence of a rather broad folksonomy, each paper only needed a small set of tags to be described. This would suggest that there is a core of shared and agreed knowledge about tags within the communities who use them, and these are recurrently used to describe the same papers. We formulate the hypothesis that, by looking at what tags were associated to what papers, *tags' similarity* (or, rather, 'relationship') can be quantified and exploited to uncover relevant content.

In the next section, we describe how we used these hypothesis to develop our content search and recommendation technique.

3 SOCIAL RANKING

Let us consider a user $\overline{u}$ who is interested in finding some content of interest (in our specific case, papers). In a typical Web 2.0 scenario, $\overline{u}$ would submit a query $q_{\overline{u}}$ which consists of query tags $t_1, t_2, \ldots, t_n$. The system answering the query would normally rank results according to the following two criteria: the higher the number of query tags associated to the resource, the higher its ranking; and, the higher the number of users u_i who tagged the resource using (some of the) query tags, the higher its ranking. Intuitively speaking, the first criterion caters for accuracy of the result, the second caters for confidence in it. The formula used could look like:

$$R(p) = \sum_{u_i} (\#t_i \text{ used by } u_i \text{ on } p \mid t_i \in q_{\overline{u}}) \ , \qquad (1)$$

that is, the ranking of paper p is computed as the number of tags t_i that users u_i who bookmarked p used and that belonged to the query set $q_{\overline{u}}$.

While this simple technique could work well to find popular content described with popular tags (i.e., with reference to our previous analysis, papers that have been tagged more than 10 times using a small subset of popular tags), the technique would likely fail to address queries that look for the very long tail of medium-to-low popularity content, as a large amount of results would be returned, all scoring low. Accuracy would not be the only problem: if the user running the query used tags that also belong to the long tail, chances are that no content would be found at all, and coverage would then become a major issue.

To address these problems, we propose *social ranking*, a technique inspired by traditional Collaborative Filtering [2]: first, we identify who are the users with similar interests to the querying user $\overline{u}$; according to our analysis, such community should be easily identified by studying users' tag activity. Content tagged by these users should be scored higher in a way that is proportional to the quantified similarity. Second, even though tags can be broadly clustered in domains of knowledge, people tend to use slightly different subsets of them within each domain (as shown by the low number of tags used by each individual and on each paper). We thus identify the tags that are similar (or, rather, related) to the query tags, thus expanding the query to this enlarged set. We believe, and our evaluation will confirm, that *users' similarity improves accuracy* of the results, while *tags' similarity (i.e., query expansion) improves coverage*.

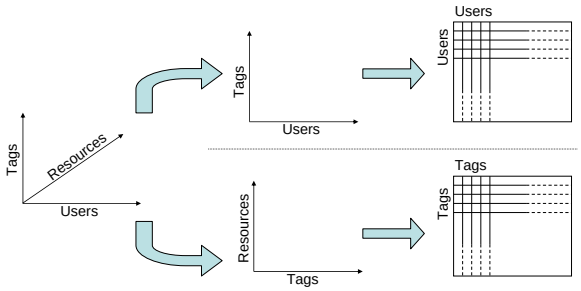

Figure 1. Transformation of the dataset

In the reminder of this Section, we illustrate how we compute users' similarity (Section 3.1), how we compute tags' similarity (Section 3.2), and how we combine these two techniques together (Section 3.3).

3.1 Users' Similarity

Social tagging typically provides a 3-dimensional relationship between users, resources and tags (users bookmark resources using a certain number of tags). Different definitions of users' similarity can be derived; here we consider a simple yet effective one: the more tags two users have used in common, the more similar they are, regardless of what resources they used it on. This definition projects our 3-dimensional space onto a 2-dimensional one, throwing away information about 'resources', and keeping only information about what tags a users has used and how often (Figure 1, top). While one may argue that, in so doing, we discard important information, we believe that, in scenarios where tags are highly clustered around topics, the information lost is not significant.

We thus describe each user u_i with a vector v_i where $v_i[j]$ counts the number of times that users u_i used tag t_j. Given two users u_i and u_j, we then quantify users' similarity $sim(u_i, u_j)$ as the cosine of the angle between their vectors:

$$sim(u_i, u_j) = cos(v_i, v_j) = \frac{v_i \cdot v_j}{||v_i|| * ||v_j||}$$

Various similarity measures can be used other than the cosine-based similarity [6]. For example, concordance-based similarity [1] could be used, so that the more tags two users share, the more similar they are (regardless of how many times they have used them). However, we believe tag frequency to be an important piece of information to determine a user's interests. Alternatively, Pearson Correlation (and its variations - e.g., weighted Pearson [19, 6]) could be used; as shown in [14], different similarity measures perform differently, both in terms of accuracy and coverage; we chose cosine-based similarity for its constantly good performance, although we plan to study the impact of other similarity measures in the future.

3.2 Tags' Similarity

We define tags' similarity as follows: the more resources have been tagged with the same pair of tags, the more similar (related) these tags are, regardless of the users who used them. This definition projects our 3-dimensional space onto a 2-dimensional one, as shown in Figure 1, bottom part. Similarly to what we said before, in scenarios where users' interests are a rather small and consistent subset of the broader range of topics in the whole website, we believe that the information thrown away during the projection is not significant.

We thus describe each tag t_i with a vector w_i where $w_i[j]$ counts the number of times that tag t_i was associated to paper p_j. Given two tags t_i and t_j, we then quantify tags' similarity $sim(t_i, t_j)$ as the cosine of the angle between their vectors:

$$sim(t_i, t_j) = cos(w_i, w_j) = \frac{w_i \cdot w_j}{||w_i|| * ||w_j||}$$

3.3 Two-Step Query Model

The query model we propose exploits the two similarity measures discussed above (on users and on tags) in the following way. When user $\overline{u}$ submits a query $q_{\overline{u}} = \{t_1, t_2, \ldots, t_n\}$ to discover content that can be described by query tags $t_1, t_2, \ldots, t_n$, two steps take place:

1. **Query Expansion**: the set of query tags q is expanded so to include all $t_i \mid t_i \in q_{\overline{u}}$ (for which $sim(t_i, t_i) = 1$), plus those tags $t_{n+1}, \ldots, t_{n+m}$ that are deemed most similar to the query tags (for which $0 < sim(t_i, t_j) \leq 1$, with $i \in [1, n]$ and $j \in [n+1, n+m]$). We call this set q^*. This set is constructed so that, for each $t_i \in q_{\overline{u}}$, its top k most similar tags are included, in a fashion similar to the top k Nearest Neighbour (kNN) strategy in recommender systems. Different choices of k will have an impact on both accuracy and coverage; we will discuss preliminary results for different values of k in Section 4.
2. **Ranking**: all resources that have been tagged with at least one tag from the extended query set are retrieved. Their ranking depends on a combination of: the relevance of the tags associated to the paper with respect to the query tags (papers tagged with $t_i, i \in [1, n]$ should count more than those tagged with $t_j, j \in [n+1, n+m]$); and, the similarity of the taggers with respect to the querying user $\overline{u}$ (papers tagged by similar users should be ranked higher, as these users are more likely to share interests with $\overline{u}$ than others, and thus are in a better position to recommend relevant content).

The ranking of a paper p would then be computed as:

$$R(p) = \sum_{u_i} \left(\sum_{\substack{t_i \\ t_j \in q^*}} sim(t_i, t_j) \right) * (sim(\overline{u}, u_i) + 1) \quad (2)$$

where, for each user u_i who tagged p, $\sum_{\substack{t_i \\ t_j \in q^*}} sim(t_i, t_j)$ quantifies how relevant the tags t_i associated by u_i to p are with respect to the tags t_j belonging to expanded query q^*; note that, in the basic case of formula 1, this simply meant counting how many tags from q user u_i associated to p. Moreover, the relevance is then magnified (i.e., papers are pushed higher up in the ranking) in a way that is proportional to user's similarity $sim(\overline{u}, u_i)$.

We call this approach social ranking as it exploits information coming from the emergent social network of users and social network of tags to rank content in a way that is meaningful to the querying user. We are now ready to evaluate this approach.

4 EVALUATION

In this section, we present preliminary results of the ongoing evaluation of social ranking. We begin with a brief description of the portion of the dataset we have been experimenting with, together with a characterization of the properties that are mostly relevant to social filtering (Section 4.1). We then describe how we have conducted

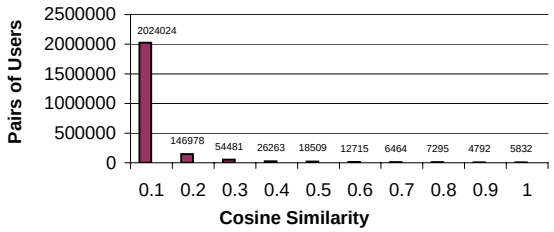

Figure 2. Distribution of users' similarity

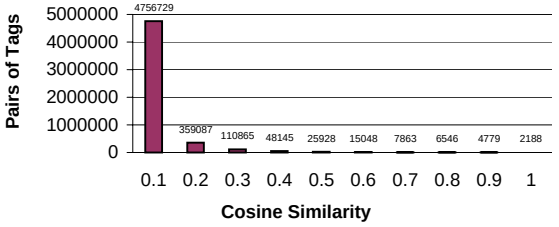

Figure 3. Distribution of tags' similarity

the experiments (Section 4.2), before analysing social ranking performance, both in terms of accuracy and of coverage (Section 4.3).

4.1 The Dataset

Based on the pre-analysis of the CiteULike dataset described in Section 2, we have performed a further cut of the dataset, in order to obtain a small yet meaningful subset to experiment with. In particular, we have considered only those tags that have been used on at least 15 different papers, and by at least 20 users. This has left us with a dataset consisting of roughly 12,000 users, 83,000 papers, and 16,000 tags. Note that the long tail phenomenon still dominates in the pruned dataset:

Long tail of users' similarity: as shown in Figure 2, the vast majority of users' pairs have very low value of similarity (below 0.1), while there exists a long tail of higher similarity pairs. This would suggest users are highly focused (and clustered) around topics, and thus only a small portion of users are indeed good recommenders to each other.

Long tail of tags' similarity: as shown in Figure 3, each tag is related to only a small subset of other tags, again suggesting that only a handful of tags are used (and thus need to be learned) to describe specific categories of content.

We believe that the results we are going to present in this section generally hold for datasets that exhibit similar characteristics.

4.2 Simulation Setup

In order to quantify accuracy and coverage of social ranking, we have been conducting the following experiment. We randomly picked up a user $\overline{u}$, randomly "hid" one of his bookmarked papers p, and then performed a query q with the tags that $\overline{u}$ had associated to p. Since p was bookmarked by $\overline{u}$ (before we hid it), $\overline{u}$ is obviously interested in it, so a recommender system should be able to return p (coverage). Note that, in our pruned dataset, it was always the case that, even after hiding $\overline{u}$'s bookmark for p, at least another bookmark made by a user u' for p existed, as we only kept in the dataset those papers that had been bookmarked by at least one user; it should thus be possible, in principle, to locate and return p. Moreover, the highest the position of p in the ranked list of returned papers, the better the accuracy of the ranking algorithm.

In all experiments, we compare the output of our social ranking algorithm (formula 2) with the simple benchmark presented in Section 3 (formula 1). Given the high variability of users' behaviour and papers' popularity in the dataset, we have been conducting 6 different sets of experiments where we varied:

- the level of activity of the querying user, distinguishing heavy taggers HT (users who tagged more than 50 papers), medium taggers MT (users who tagged between 10 and 50 papers), and low taggers LT (users who tagged less than 10 papers);
- the level of popularity of the hidden bookmark, distinguishing popular papers PP (those that had been bookmarked by at least 5 users), and unpopular ones UP (those that had been bookmarked by less than 5 users).

Our goal is to investigate the impact of these two characteristics onto the efficiency of the querying model.

4.3 Results

The first set of experiments we conducted aimed to analyse the impact of users' similarity alone on the ranking of results. We thus compared the basic query model with the advanced query model where tag expansion had been disabled. For each query, the list of returned papers is thus the same, but ranked differently. For each user in each group (heavy/medium/low taggers), three bookmarks where removed for each paper category (popular/unpopular), and their corresponding tags searched. As the number of users in each group varies, so does the total number of queries performed (from 2,200 for the small group of HT/PP, to 14,000 for the much larger group of LT/UP). Table 1 summarises the results obtained, in terms of percentage of times the advanced model does better/same/worse than the basic query model. We also report the percentage of queries for which the target paper remained uncovered.

Experiment	Better	Worse	Tied	Not Found
HT/PP	25%	28%	30%	17%
HT/UP	49%	12%	15%	24%
MT/PP	25%	15%	42%	18%
MT/UP	42%	13%	9%	36%
LT/PP	26%	13%	45%	16%
LT/UP	39%	13%	8%	40%

Table 1. Impact of Users' Similarity on the Ranking of Result

In all scenarios but the first one, the advanced query model outperforms the basic query model, and it does so more dramatically when considering unpopular papers, where the gap between the two approaches (the difference between the 'better' and 'worse' column) reaches 37%. This result confirms the importance of weighting the recommendations coming from similar users more, when looking for less 'mainstream' content. The ranking of results is slightly better (28% against 25%) when using the basic query model in the first scenario instead: when focusing on more mainstream content (i.e., the

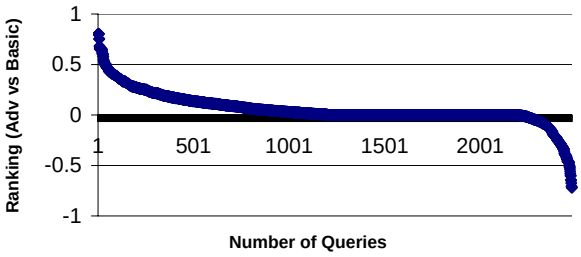

Figure 4. Improvement obtained for unpopular papers

hidden paper has been tagged many times by different users), simple searches based on exact tag matching work well enough.

In order to quantify the improvement obtained, we have computed the difference of the ranking at which a paper is found using the advanced model versus the basic model, normalised by the total number of results returned (i.e., a tiny difference in the ranking between the two techniques, for example, position 12 versus position 14, is not really appreciable by an end user). The higher the positive difference, the better the performance of the advanced model and viceversa. Figure 4 illustrates the results for heavy taggers and unpopular papers. As shown, there is a considerable set of results for which the performance of the advanced approach is not simply better than basic (positive difference), but significantly so (more than 20% of the queries ran, returned the hidden paper at a position that is between 80% and 20% higher up with the advanced model than it was with the basic model).

An important observation can be drawn when looking at the column labeled 'Not Found' in Table 1: searching techniques purely based on the matching of the query tags with those associated to papers reveal a substantial amount of uncovered resources; this percentage is approximately 16-17% for popular papers, and it increases up to 40% for unpopular ones. Low coverage is an indication that users bookmark resources differently; in order to uncover resources of interest, query tags must be expanded to include semantically related ones. We have thus conducted a second set of experiments, using the full social ranking model against the basic one, when extending each query tag with the top kNN tags. The goal of these experiemnts is to quantify the improvement obtained by social ranking on coverage, and its consequences on accuracy. We have focused on the long tail of unpopular papers, as this was the most problematic scenario, and the one where social ranking should bring the most benefits. So far we have obtained results for the heavy taggers/unpopular papers scenario for $k = 5$: the items not found are reduced from 24% to 14%; in cases where both techniques find the hidden paper, social ranking performs better in 40% of the cases, while in 31% it does worse. These inital results would confirm that it is possible, with simple techniques, to automatically learn tags' similarity and use it to boost coverage, without giving accuracy away. We are running experiments with higher values of k: preliminary results, obtained on a smaller set of queries, would indicate that coverage keeps improving up to $k = 20$, while higher values begin to have a non-negligible negative impact on accuracy. This is aligned with the pre-analysis we have conducted: each paper usually receives 10 tags or less, up to a maximum of 30 tags; expanding the query tags to larger sets thus injects too much noise within results.

5 RELATED WORK

Research in the area of social tagging has proliferated in recent years, due to the increasing popularity of such systems. Studies have been conducted both to understand tag usage and evolution (e.g., [22, 4]), and to learn and exploit their hidden semantics. In [8], a large study of social tagging on the popular del.icio.us bookmarking system is presented, aimed at characterizing users' activity, pages' popularity, and tags' distribution; the knowledge base (in this case, the whole Web) is so large and dynamic that the authors are quite pessimistic on the benefits that social bookmarking can bring to web searches. In [7], the same authors have shown how searches on del.icio.us can be improved if a navigable hierarchical taxonomy of tags is derived from tag usage, to help users broadening/narrowing the set of tags that best describe their interests. Our approach takes a different stance, and rather than offering users an organised tag navigation system, it aims to transparently improve users' searches based on emergent tags semantics and query expansion. In [18], tags are related back to a fixed ontology of concepts, thus exploiting both techniques to enhance information retrieval capabilities. Differently from this approach, our goal is to autonomically derive tags' relationships, which can then be fitted into an effective query search algorithm, without relying on a prefixed ontology. In [20], semantics that specifically relate to places and events are inferred for resources within the Flickr dataset; their approach is highly tied to location information, and thus not easily generalizable to other domains. In [24], a probabilistic generative model is proposed to describe users' annotation behavior, and to automatically derive tags emergent semantics; during searches, their approach is capable of grouping together synonymous tags, while it calls for user's intervention when highly ambiguous tags are found. Very early work, but with similar goals, is presented in [25], where a simpler technique, based on an analysis of the relationship between users, tags and resources, is proposed to disambiguate tags. Tag systems have recently revealed their susceptibility to tag spam, that is, malicious annotations generated to confuse users. The problem has been well analysed in [13], where they tried to identify misused tags, and quantify the extent to what tagging systems are robust against spam. Robust solutions to tag spamming are still being investigated.

Research has been very active also in relating tag activity to users, in order to discover their interests and consequently users' communities. Work within the Semantic Web domain has tried to classify users into categories and describe the key features of such categories [15]. More recently, users have been classified according to their explicitly stated profile [10], based on a probabilistic model which takes into account users's interest to topics [26], and based on their level of tagging activity and breadth of interests [12]. In [16], users' common interests are discovered based on patterns of frequently co-occurrent tags, using a classical association rule algorithm, which however does not take into account considerations about user's activity. All these works, including our attempt to find similar users, are based on the observation that real world networks exhibit a so-called community structure [21]; defining the set of characteristics that would enable the best fitting and natural clustering of taggers and tags is an open research question.

In this paper, we have been combining the two research streams highlighted above (i.e., automatic learning of tag semantics and users' interests) in order to improve query searches and ranking. Other research groups have been conducting research in the same area. In [23, 17], the integration of tag information within standard recommender system's algorithms has been proposed, in order to give better recommendations to users; although very promising, at

present such works do not take into account the 'activity' of users, in terms of amount of resources being tagged, and number of tags being used. We believe this information to be crucial to extract users' interests and thus improve the efficiency of searches. Tag activity has been combined with a PageRank-like algorithm, in order to improve the ranking mechanism, in situations where resources are not linked together as in a typical web graph structure [9]; their approach, called FolkRank, provides good results when querying the folksonomy for topically related elements, while it is easily subverted if less related/popular tags are being used, due to the size and sparsity of folksonomies on the web. In this work, we have tried to ameliorate the sparsity problem in folksonomy; further improvements could be achieved by clustering users within better scoped communities; we intend to explore this aspect next.

6 FUTURE DIRECTIONS

In this paper we have presented social ranking, a technique that aims to improve content searches in Web 2.0 scenarios, by exploiting users' similarity and tags' similarity. The former is used to gain confidence in the relevance of the retrieved content: the higher the similarity between the querying user and the user that has bookmarked it, the higher the chances that the paper is of relevance, thus reducing the amount of uninteresting content being presented to users. The latter is used to tackle the problem of heterogeneity, sparsity and lack of structure in folksonomy instead: by implicitly learning tags' similarity from their usage, we can increase the amount of relevant yet unpopular content being uncovered.

Ongoing work spans different directions. First, we are conducting a variety of experiments to better assess the current ranking model: we are varying the value of k during query tag expansion, the similarity function used to compute both user' and tags' similarity, and we plan to experiment with a different dataset too (namely, Last.fm). The technique we currently use to populate the user-by-tag matrix, and the tag-by-resource matrix, is rather simplistic (a basic counter of how many times a user has used a tagged, and how many times a tag has been used on a resource). More advanced techniques could be used, which could then lead to more accurate similarity results: for example, including time information, to cater for the most frequently used tags, the most recently used tags, etc.

In terms of model, our plan is to refine the techniques we use to find both similar users and similar tags. We have started analysing the impact of a variety of clustering techniques to identify communities of users; beyond similarity in the tags' usage, there exist other parameters of relevance, including level of activity (to distinguish active users who contribute to the knowledge base, from passive consumers), variety of tags used (unpopular tags may reveal more about a user's interests than popular ones), and so on. In parallel, we are studying more refined techniques to learn relationships between tags. The ultimate goal is to enrich Web 2.0 applications with accurate and robust techniques to give users what they are really looking for.

REFERENCES

[1] A. Agresti, *Analysis of Ordinal Categorical Data*, J.Wiley & Sons, 1984.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, 'Using Collaborative Filtering to Weave an Information Tapestry', *Communications of the ACM*, **35**, 61–70, (1992).

[3] Scott Golder and Bernardo A. Huberman, 'Usage patterns of collaborative tagging systems', *Journal of Information Science*, **32**(2), 198–208, (2006).

[4] Harry Halpin, Valentin Robu, and Hana Shepherd, 'The complex dynamics of collaborative tagging', in *Proceedings of the 16th International Conference on World Wide Web*, pp. 211–220, New York, NY, USA, (2007). ACM Press.

[5] Y. Hassan-Montero and V.Herrero-Solana, 'Improving tag-clouds as visual information retrieval interfaces', in *Intl. Conference on Multidisciplinary Information Sciences and Technologies*, Merida, Spain, (2006).

[6] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, 'An Algorithmic Framework for Performing Collaborative Filtering', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 230–237, (1999).

[7] Paul Heymann and Hector Garcia-Molina, 'Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems', Technical Report 2006-10, Stanford University, (April 2006).

[8] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina, 'Can Social Bookmarking Improve Web Search?', *Resource Shelf*, (2007).

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, *Information Retrieval in Folksonomies: Search and Ranking*, 411–426, 2006.

[10] William H. Hsu, Joseph Lancaster, Martin S.R. Paradesi, and Tim Weninger, 'Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach', (March 2007).

[11] O. Kaser and D. Lemire, 'Tag-cloud drawing: Algorithms for cloud visualization, tagging and metadata for social information organizatio', in *Intl. Conference on the World Wide Web*, Alberta, Canada, (2007).

[12] Shreeharsh Kelkar, Ajita John, and Doree Seligmann, 'An Activity-based Perspective of Collaborative Tagging', (March 2007).

[13] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, 'Combating spam in tagging systems', in *Proc.of the 3rd Intl. Workshop on Adversarial Information Retrieval on the Web*, pp. 57–64, New York, NY, USA, (2007).

[14] N. Lathia, S. Hailes, and L. Capra, 'The effect of correlation coefficients on communities of recommenders', in *Proceedings of 23rd Annual ACM Symposium on Applied Computing*, (2008).

[15] K. Faith Lawrence and M. C. Schraefel, 'Bringing Communities to the Semantic Web and the Semantic Web to Communities', in *Proceedings of the 15th International Conference on World Wide Web*, (2006).

[16] X. Li, L. Guo, and Y. E. Zhao, 'Tag-based Social Interest Discovery', in *Proc. of the 17th Intl. World Wide Web Conference*, (2008).

[17] Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, and Shunsuke Uemura, 'Tag-based Contextual Collaborative Filtering', in *18th IEICE Data Engineering Workshop*, (2007).

[18] Alexandre Passant, 'Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs', in *Proceedings of International Conference on Weblogs and Social Media*, (2007).

[19] H. Polat and W. Du, 'Privacy-Preserving Collaborative Filtering using Randomized Perturbation Techniques', in *The Third IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, FL, (November 2003).

[20] T. Rattenbury, N. Good, and M. Naaman, 'Towards automatic extraction of event and place semantics from flickr tags', in *Proc.of the 30th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 103–110, New York, NY, USA, (2007).

[21] Jianhua Ruan and Weixiong Zhang, 'Identifying network communities with a high resolution', *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, **77**(1), (2008).

[22] S. Sen, S. K. Lam, Al M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, M. F. Harper, and J. Riedl, 'tagging, communities, vocabulary, evolution', in *Proc.of the 20th Conference on Computer Supported Cooperative Work*, pp. 181–190, New York, NY, USA, (2006).

[23] Karen H. L. Tso-Sutter, Leandro Balby Marinho, and Lars Schmidt-Thieme, 'Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms', in *Proceedings of 23rd Annual ACM Symposium on Applied Computing*, pp. 16–20. ACM Press, (2008).

[24] X. Wu, L. Zhang, and Y. Yu, 'Exploring social annotations for the semantic web', in *Proc. of the 15th Intl. Conference on World Wide Web*, pp. 417–426, New York, NY, USA, (2006).

[25] C. Man Au Yeung, N. Gibbins, and N. Shadbolt, 'Mutual Contextualization in Tripartite Graphs of Folksonomies', in *Proc. of the 6th Intl. Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, pp. 960–964, (2007).

[26] D. Zhou, E. Manavoglu, J. Li, L. C. Giles, and H. Zha, 'Probabilistic models for discovering e-communities', in *Proc. of the 15th Intl. Conference on World Wide Web*, pp. 173–182, New York, NY, USA, (2006).