Measuring the influence of tag recommenders on the indexing quality in tagging systems

Article · June 2012

DOI: 10.1145/2309996.2310009

2 authors, including:

Steffen Staab
Universität Stuttgart
661 PUBLICATIONS **26,034** CITATIONS

SEE PROFILE

Measuring the Influence of Tag Recommenders on the Indexing Quality in Tagging Systems

Klaas Dellschaft
Inst. for Web Science and Technologies (WeST)
Universität Koblenz-Landau
Koblenz, Germany
klaasd@uni-koblenz.de

Steffen Staab
Inst. for Web Science and Technologies (WeST)
Universität Koblenz-Landau
Koblenz, Germany
staab@uni-koblenz.de

ABSTRACT

In this paper, we investigate a methodology for measuring the influence of tag recommenders on the indexing quality in collaborative tagging systems. We propose to use the inter-resource consistency as an indicator of indexing quality. The inter-resource consistency measures the degree to which the tag vectors of indexed resources reflect how the users understand the resources. We use this methodology for evaluating how tag recommendations coming from (1) the popular tags at a resource or from (2) the user's own vocabulary influence the indexing quality. We show that recommending popular tags decreases the indexing quality and that recommending the user's own vocabulary increases the indexing quality.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative Computing, Evaluation*

Keywords

Collaborative Tagging, Indexing Quality, Tag Recommenders

1. INTRODUCTION

Collaborative tagging systems allow users to organize resources, e. g. photos, bookmarks or BibTeX entries, by assigning tags or keywords to them. Users can freely choose the tags which they want to use for indexing resources. Over time, the tag assignments of the different users lead to the emergence of a loose categorization system for resources, often called a *folksonomy* [11]. One key aspect of tagging systems is the uncontrolled nature of the community's vocabulary. Nevertheless, it has been observed in [5, p. 205] that the combined tag assignments of users "give rise to a stable pattern in which the proportions of each tag are nearly fixed". This is typically taken as an indicator that tagging is successful in collaboratively indexing resources despite of its uncontrolled nature.

During adding tags to resources, in many tagging systems, e. g. in Delicious and Bibsonomy, the users see a set of tag recommendations. It is an often posed question how these tag recommendations influence the users in their tagging decision and whether this influence is rather positive or negative. In the related work (see Section 2), there exist several approaches for analyzing the influence of tag recommendations. In many cases, it is analyzed how tag recommendations influence the inter-indexer consistency [3, 5, 7, 10, 14]. It corresponds to the degree to which the users have agreed on a common vocabulary for describing resources. It is assumed that increasing the inter-indexer consistency is important for dealing with the uncontrolled nature of the vocabulary in tagging systems.

But what does a high inter-indexer consistency mean in terms of indexing quality as it might also be measured by precision and recall during querying resources? The inter-indexer consistency only measures the average consensus of the indexers at the single resources. But precision and recall are influenced by how indexers use the indexing terms across a set of resources. For example, a high recall is achieved if related resources are linked to each other by indexing them with terms which express their common aspects. Furthermore, a high precision is achieved if indexing terms are discriminative enough, i. e. if they only link related but not also unrelated resources.

None of these aspects which influence precision and recall are directly measured by the inter-indexer consistency. Nevertheless, the inter-indexer consistency may be positively correlated with these aspects by introducing the assumption that the indexers reach the same consensus for related resources and a different consensus for unrelated resources.

In contrast, the inter-resource consistency is a more direct way of measuring the aspects which lead to high precision and recall. It measures in how far the indexers are successful in linking related resources by indexing their common aspects. Thus, the inter-resource consistency is directly correlated with the indexing quality and a high precision and recall of query results (cf. Subsection 3.1). Measures of inter-indexer consistency are also positively correlated with the inter-resource consistency, if it can be assumed that the indexing terms are "selected individually and independently by each of the indexers" [19].

However, our investigations show that the assumption of the positive correlation between inter-indexer consistency on the one hand and inter-resource consistency or indexing quality on the other hand does not hold when it comes to investigating the influence of tag recommenders. The rea-

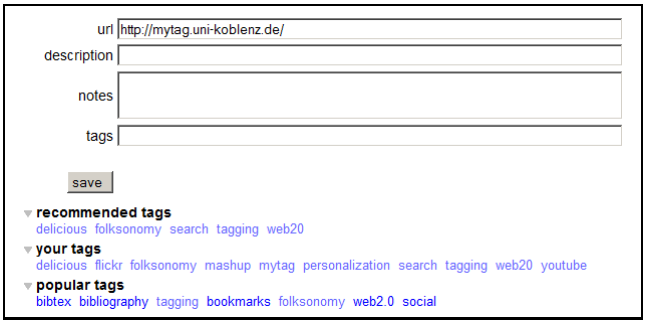

Figure 1: Tagging interface of Delicious. Users can enter free tags in the *tags* input field and/or they can select some of the suggested tags.

son is that the users no longer apply their tags individually and independently of each other. We thus argue that only the inter-resource consistency can be used as an indicator of indexing quality in tagging systems. We support this argument by measuring the inter-resource and the inter-indexer consistency for two exemplary tag recommenders, showing that the two measures are not positively correlated with each other. The rest of this paper is structured as follows:

In Section 3, we provide a methodology for measuring the inter-resource consistency in tagging systems. Furthermore, we take from the related work a method for measuring the inter-indexer consistency. In Section 4, we derive the hypotheses that these two measures are not positively correlated with each other for two exemplary baseline tag recommenders. Then, in Section 5 we describe the user experiment during which we collect the necessary data for measuring the inter-resource and the inter-indexer consistency. The results from Section 6 support our hypotheses that the two measures are not positively correlated with each other if tag recommendations are used and that thus only the inter-resource consistency is a valid measure of indexing quality.

2. RELATED WORK

Given an individual user who is about to tag a given resource, e. g. a web page, there are three basic paradigms of suggesting tags to this user [8]: One can suggest (1) tags based on the tag assignments of other users (either extracted from the tag assignments associated with the current resource or from all tag assignments), (2) tags based on the previous tag assignments of the current user, and (3) tags based on the content of the current resource, e. g. by extracting keywords from the content or title of a web page.

Simple tag recommendation algorithms suggest tags only based on one of the three paradigms. For example, in the tagging interface of Delicious (see Fig. 1) the user sees amongst others the seven most popular tags at the current resource and all his previously used tags. More sophisticated tag recommendation algorithms suggest tags based on several of the paradigms. For example, the recommender in [8] first extracts candidate tags from the local vocabulary and the content of the current resource. Then, the candidate tags are checked against the vocabulary of the current user.

But which effect does a given tag recommendation algorithm have on the indexing quality in collaborative tagging systems? In the introduction, we have explained that one central aspect of indexing quality is the inter-resource consistency, i. e. in how far the indexers are successful in linking related resources by indexing their common aspects [18]. For measuring the inter-resource consistency, an indicator of relatedness of resources is required which is independent of the indexing to be tested [18]. In [18], the authors use topical clusters of resources for measuring in how far the resources within a cluster are linked by their indexing terms. The higher the inter-resource consistency, the better are precision and recall of queries which use the indexing terms.

Because this independent indicator of resource relatedness is difficult to acquire, many studies concentrate on the inter-indexer consistency instead. The inter-indexer consistency does not require such additional data. But this approach is only valid if the indexing terms are "selected individually and independently by each of the indexers" [19]. An example of a recent study using inter-indexer consistency where this assumption holds is available in [13]. In this study, the authors compare the indexing quality between professional indexers and laymen.

But also in the literature about tagging systems, often some kind of inter-indexer consistency is measured. For example, in [5] it is studied how long it takes until the frequencies of the most popular tags at a resource have reached a stable state. A faster convergence process is ascribed to be an indicator for a higher inter-indexer consistency. Furthermore, in [14] the inter-indexer consistency is measured in terms of the tag reuse rate, i. e. by the average number of users who apply a tag. Finally, in several other studies [3, 6, 7, 10] a smaller size of the final vocabulary is taken as an indicator for the consensus among the users. In all these studies, it has been shown that recommending tags based on the tag assignments of other users leads to a higher inter-indexer consistency. This higher inter-indexer consistency is then taken as an indicator of an improved indexing quality, ignoring the fact that the assumption of individual and independent tag selection does not hold.

But besides the approaches for measuring the inter-indexer consistency, also some alternative measures for indexing quality have been proposed in the literature about tagging systems. For example, in [15, 17] it has been proposed to compare the resulting tag assignments against the true preferences of the users. In [15, 17], the true preferences of the users are defined to be the tag assignments which would occur without the influence of tag suggestions. This methodology judges any deviation from the uninfluenced behavior of users as negative. Thus, from our perspective it seems unsuitable for measuring positive effects of tag recommenders because positive effects can only occur if users deviate from their uninfluenced behavior.

Furthermore, in the literature about tag recommendation algorithms, the quality of tag recommenders is measured by the precision and the recall of the set of recommended tags (see [8, 16] for examples). In these studies, precision and recall compare the tag recommendations against a gold standard. It depends on the chosen gold standard how to interpret precision and recall. For example, if precision and recall are measured in a live system[1], then they measure how often users accept a recommendation. If precision and recall compare the recommendations with the uninfluenced tag assignments of the individual users then the recommendations are compared to the true preferences of the users

[1]See the *Online Tag Recommendations* track of the ECML PKDD Discovery Challenge 2009 http://www.kde.cs.uni-kassel.de/ws/dc09/

(like in [15, 17]). The former approach measures in how far users prefer one tag recommender over another. But this aspect of tag recommendations is distinct from the resulting indexing quality. The latter approach of comparing with uninfluenced tag assignments has the same drawbacks as the approach in [15, 17] (see above).

3. MEASURING THE INDEXING QUALITY

In this section, we describe the inter-resource and inter-indexer consistency measures in more detail. Furthermore, we explain in Subsection 3.1 how an improved inter-resource consistency might also lead to an improved precision and recall for queries. In Section 4, we then present our hypotheses how inter-resource and inter-indexer consistency are influenced by two exemplary tag recommenders.

3.1 Measuring the Inter-Resource Consistency

In general, the inter-resource consistency measures in how far indexers are successful in linking related resources by indexing their common aspects. We follow the approach from [18], in which the relatedness of resources according to their tag vectors V is compared to their relatedness according to a set of topical clusters C. Given V and C, the idea of inter-resource consistency is as follows: (1) If two resources are contained in the same topical cluster $c \in C$ then this should be reflected by a higher similarity of their tag vectors. (2) If two resources are contained in different topical clusters c_1 and c_2 then this should be reflected by a lower similarity of their tag vectors. The higher the ratio of the similarity within a cluster and the similarity between distinct clusters, the better is the inter-resource consistency.

During ranked retrieval, the similarity of two tag vectors v_1 and v_2 has an important influence on the relevance ranking of the corresponding resources with regard to a query. The more similar two tag vectors, the more likely the corresponding resources will get a similar relevance value and the closer together they will be in the ranked result list. Thus, the first criterion from above ensures that resources from the same topical cluster are in average ranked closer together. The second criterion ensures that resources from distinct clusters are in average ranked farther away from each other.

Overall, combining both criteria leads to a better separation of resources from the different topical clusters in a ranked result list. Thus, resources from the topical cluster which is most relevant for answering a query are intermixed with fewer resources from other, less relevant topical clusters. Assuming the match between indexing terms and query terms which has been shown in [15], then an improved inter-resource consistency finally leads to an improved precision and recall for the top-k results for a query.

Measuring the inter-resource consistency is a two-step process: In a first step, we have to measure the pairwise similarities of the tag vectors in V. In the second step, we then measure the ratio between the similarity of tag vectors within a cluster and the similarity of tag vectors from distinct clusters. In the following, we propose measures applicable for a ranked retrieval model. In [18], measures for a boolean retrieval model are available.

3.1.1 Measuring the Similarity of Tag Vectors

A common model in information retrieval is the vector space model which forms the basis for ranked retrieval. It is the fundamental model for several information retrieval tasks like scoring documents on a query, document classification and document clustering [9]. According to this model, the tag vector of a resource captures the relative importance of tags for this resource. In our case, the tag vector of a resource contains how often the tags $t_1, \ldots, t_n$ have been assigned to it by the users. A standard way for calculating the similarity of resources in the vector space model is the cosine similarity [9]. Given two tag vectors v_i and v_j, their similarity is measured as follows:

$$cosim(v_i, v_j) = \cos\Theta = \frac{v_i \cdot v_j}{\| v_i \| \cdot \| v_j \|} \quad (1)$$

The calculation of the cosine similarity is based on the angle Θ between two tag vectors. Θ itself can also be used for measuring the *dissimilarity* between the tag vectors of two resources.

3.1.2 Within Cluster vs Distinct Clusters Similarity

In order to measure the ratio between the similarity of tag vectors in the same topical cluster and the similarity of tag vectors in distinct topical clusters, we propose to use the Silhouette Coefficient. The Silhouette Coefficient was first introduced in [12] for evaluating clustering algorithms. Given a set of resources $R = \{r_1, \ldots, r_n\}$ and a set of clusters $C = \{c_1, \ldots c_k\}$, so that each resource is contained in only one of the clusters, and a measure of dissimilarity between the resources, the Silhouette Coefficient s_i for a resource r_i is computed as follows:

First, the average dissimilarity a_i of r_i to all other resources in its cluster c is computed. Second, from all clusters not containing r_i, we identify the cluster c' whose resources have in average the lowest dissimilarity to r_i. We call this minimal average dissimilarity b_i. In our case, the average dissimilarity scores a_i and b_i correspond to the average angle Θ between the tag vectors of the resources. Finally, a_i and b_i are set into relation to each other as follows:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (2)$$

The Silhouette Coefficient s_i ranges between -1 and 1. s_i will take a positive value if resource r_i is closer to the resources in the same cluster c than to resources in the closest other cluster c'. It reaches its maximal value if the dissimilarity of r_i to the resources in c is 0. In contrast, s_i will take a negative value if r_i is farther away from the resources in c than from the resources in c'. It reaches its minimal value if the dissimilarity of r_i to the resources in cluster c' is 0.

In general, the following relationship between the inter-resource consistency and the precision and recall of queries for resources in cluster c holds: The lower the s_i value, the more likely it gets that resources from cluster c' are ranked as more relevant for the query than the resource r_i. This decreases the precision and recall of queries for resources in cluster c. The same holds for querying resources from cluster c': The lower the s_i value, the more relevant is r_i for such a query, thus leading to a decreased precision.

Given these definitions, we can now use the average Silhouette Coefficient $E(s_i)$ for measuring the inter-resource consistency of a set of tag vectors $V = \{v_1, \ldots, v_n\}$ of the resources in R. The higher the $E(s_i)$-value, the higher the inter-resource consistency of the tag vectors. The $E(s_i)$-values for two sets of tag vectors V_1 and V_2 can be compared given that they describe the same set of resources R

and that they are compared to the same set of clusters C. Only if these two preconditions are fulfilled, we can be sure that a difference in the two $E(s_i)$-values also indicates a difference in the inter-resource consistency for V_1 and/or V_2.

3.2 Measuring the Inter-Indexer Consistency

According to our analysis of the related work in Section 2, many authors assume that a high inter-indexer consistency also indicates a high indexing quality. We argue that this assumption does not hold if the users are influenced by tag recommendations during tagging. In order to support our argument, we compare during our evaluation a traditional measure of inter-indexer consistency to our measure of inter-resource consistency. If our argument is correct, we expect to see no positive correlation between the two measures.

By looking at the literature about tagging systems (see Section 2), two measures related to the inter-indexer consistency can be identified: The tag reuse rate from [14] and the size of the vocabulary [3, 6, 7, 10]. The global vocabulary size is not a good measure for the inter-indexer consistency because it is not only influenced by the overlap of the users' vocabularies or the inter-indexer consistency respectively but also by the average size of the users' vocabularies. Thus, we will only use the tag reuse rate for measuring the inter-indexer consistency in the following. In [14], it is defined as "the average number of users who apply a tag". In our case, we first compute the tag reuse rate tr_i for each resource r_i. The overall tag reuse rate then corresponds to the average $E(tr_i)$ over all resources.

4. RESEARCH HYPOTHESES

In this section, we present two exemplary tag recommenders which are actually used in Delicious. For these two tag recommenders we derive the hypotheses that in their case the inter-resource consistency and the inter-indexer consistency are not positively correlated with each other. The first recommender shows that the inter-resource consistency may increase even if the inter-indexer consistency decreases. The second recommender shows that the inter-resource consistency may decrease even if the inter-indexer consistency increases. If we are able to show in Section 6 that these correlations between inter-resource and inter-indexer consistency hold for the two recommenders then this would support our argument that one has to directly measure the inter-resource consistency instead of the inter-indexer consistency for making conclusions on the indexing quality in tagging systems.

4.1 Increasing the Inter-Resource Consistency

One important way for increasing the inter-resource consistency in a tagging system is to increase the inter-resource consistency of the tag assignments of the individual users. It is the objective of the *User Tags*-based recommender to help the individual user in establishing a consistent tagging vocabulary and to consistently apply it to all resources in his personal collection which have the respective aspects in common. This objective is tried to be achieved by recommending the user all his previously used tags. This recommender is also used in Delicious (see the *Your Tags* suggestions in Fig. 1). These considerations about the *User Tags*-based recommender lead to the following testable hypothesis:

HYPOTHESIS 1. *Suggesting the user his/her own tags in the user interface increases the inter-resource consistency and/or indexing quality in tagging systems.*

Of course, it is unreasonable to assume that the *User Tags* recommender increases the inter-indexer consistency in a tagging system. In reverse, it can be assumed that the inter-indexer consistency either remains unchanged or that it is even decreased. Both cases support our argument that one can not assume a positive correlation between inter-resource and inter-indexer consistency.

HYPOTHESIS 2. *Suggesting the user his/her own tags in the user interface leads to an unchanged or decreased inter-indexer consistency in a tagging system.*

4.2 Increasing the Inter-Indexer Consistency

One important way for increasing the inter-indexer consistency in a tagging system is to show the individual users the tags of the other users. It is a common assumption in the literature that such suggestions reduce the uncontrolled nature of the vocabulary in tagging systems (see Section 2). But in how far do they also help in increasing the inter-resource consistency in a tagging system? A positive correlation between both measures can no longer be automatically assumed because the individual users no longer select the tags individually and independently of each other.

In the following, we argue that in case of the *Popular Tags*-based recommender a decreased inter-resource consistency may be observed although it increases the inter-indexer consistency. Our *Popular Tags* recommender suggests the individual user the seven most popular tags of a resource, , i. e. it mimics the behavior of the corresponding recommender in Delicious (see the *Popular Tags* suggestions in Fig. 1).

In [17], it has been argued with a theoretical model that a recommender based on popular tags may distort the true tagging preferences of a user. Thus, the user applies different tags then he would do without seeing the suggestions. But in itself, distorting the true tagging preferences is not a negative thing: The *User Tags* recommender changes the actual tag assignments of a user, nevertheless we assume in Hypothesis 1 that it helps to increase the indexing quality. But in case of the *Popular Tags* recommender, the tag frequencies converge to a random limit (see [5]). Thus, the tag frequencies no longer only express the important aspects of a resource but they are also influenced by a random process. This influence of the random process decreases the inter-resource consistency:

HYPOTHESIS 3. *Suggesting the user a list with the most popular tags at a resource decreases the inter-resource consistency and/or indexing quality in tagging systems.*

Furthermore, getting feedback about the tags used by other users for describing the same resource increases the inter-indexer consistency:

HYPOTHESIS 4. *Suggesting the user a list with the most popular tags at a resource increases the inter-indexer consistency in a tagging system.*

5. USER EXPERIMENT

In this section, we describe the web-based user experiment which we use for testing the hypotheses from Section 4. The results of the experiment are presented and discussed in Section 6 and 7. The experiment is divided into two phases: During the first phase, screenshots of ten web pages were shown to the users in a random order. To each web page, the

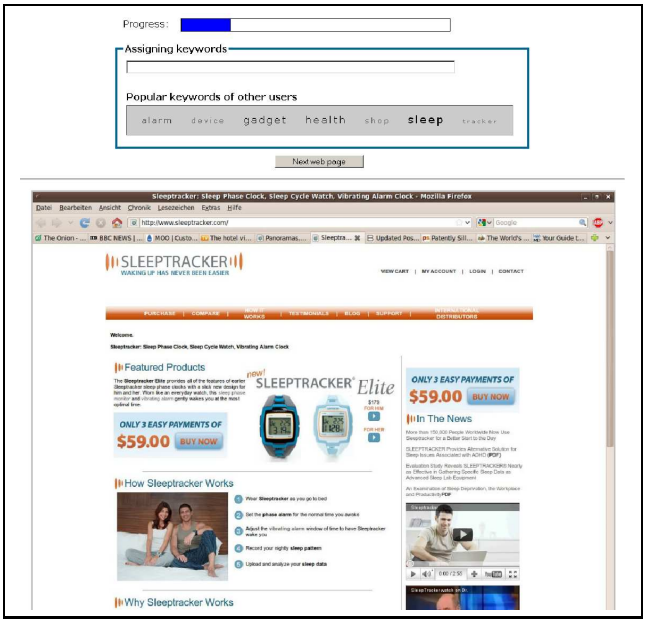

Figure 2: The user interface for assigning keywords to the 10 web pages. The web pages were shown in random order. During tagging, the users saw a screenshot of a browser window showing the web page. The screenshot also included the URL and the title of the web page. Depending on the experimental condition, a tag cloud with the tag suggestions was displayed below the input field for the keywords. Here, the interface for the *Popular Tags* condition is shown. By clicking on one of the suggested tags, the user added it to the input field.

participants should assign any number of tags (see Fig. 2). In a second phase, we asked the participants to group the web pages into topical clusters (see Fig. 3).

For the experiment, we used the same set of web pages as in a previous experiment by Bollen and Halpin [1]. The URLs of the used web pages are shown in Tab. 1. The URLs were selected so that their topics appeal to the general public and not only to participants with a specialized background. In [1] more details are available about how the specific URLs were selected. From the set of eleven web pages used in [1], one web page was removed because a pretest showed that participants had problems in understanding the topic of the web page based on a screenshot of it.

5.1 Experimental Conditions

In order to test our hypotheses from Section 4, we have to distinguish the following three experimental conditions:

Under the **No Suggestions** condition, the users do not get any tag suggestions while tagging the ten web pages. This user group is the control group to which we compare the results of the other two experimental conditions.

Under the **Popular Tags** condition, the users gets suggested the seven most popular tags for the current web page. The most popular tags are based on the tag assignments of the previous users under the same experimental condition for the same web page. Prior to the experiment, each of the web pages got initialized with the tags of a random user from Delicious for the same web page (see Tab. 2). For

Figure 3: The user interface for grouping the web pages into topical clusters. In the left column, screenshots of the 10 web pages are shown. On the right side, all clusters of the current user are shown. The users were allowed to create any number of clusters. When creating a new cluster, the users were asked to provide a name for it.

ID	URL
1	http://www.theonion.com/
2	http://news.bbc.co.uk/2/hi/uk_news/6057734.stm
3	http://uk.moo.com/
4	http://www.tvtrip.com/
5	http://www.panoramas.dk/
6	http://www.sleeptracker.com/
7	http://blisstree.com/feel/what-happens-to-your-body-if-you-drink-a-coke-right-now/
8	http://www.patentlysilly.com/
9	http://www.whfoods.com/
10	http://www.webmd.com/balance/features/your-guide-to-never-feeling-tired-again/

Table 1: URLs of the 10 web pages used during the experiment.

the German variant, the same tags have been translated to German.

The initialization of the *Popular Tags* condition with a random user is necessary in order to introduce a comparable level of randomness to the tag assignment process as in a real system like Delicious. In a real system, the resources would also be first tagged by different users. Without initializing the resources with a random posting, the first assignments at the resources would all come from the first participant of the *Popular Tags* condition.

Under the **User Tags** condition, each user sees all tags which he/she previously used in the experiment. For the first web page, the users do not get any suggestions.

5.2 Recruiting and Instructing Participants

We used several channels for recruiting participants for the experiment: (1) We approached colleagues and friends. (2) We promoted the experiment during the Web Science Conference 2011. (3) We distributed the call for participation over twitter and several public mailing lists about informa-

ID	URL
1	theonion, news, america
2	bbc, news, evolution, human
3	moo, business cards, post cards, printing
4	tvtrip, travel, hotels, reviews
5	panorama, background image
6	sleep, alarm, shop
7	health, coke, diet
8	funny, patents
9	health, food
10	sleep, health, guide

Table 2: Tags used for bootstrapping the English *Popular Tags* condition.

German	|Users|	|Tags|	|TAS|	|TAS|/|User|
No Suggestions	74	706	2,134	28.84
User Tags	79	466	1,507	19.08
Popular Tags	78	531	2,228	28.56
English	|Users|	|Tags|	|TAS|	|TAS|/|User|
No Suggestions	115	973	3,150	27.39
User Tags	118	819	2,919	24.74
Popular Tags	118	550	3,003	25.45

Table 3: Sizes of the experimental data sets. Only participants who finished tagging all ten web pages are included. (TAS = tag assignments)

tion retrieval. (4) We distributed the call in an internal news group of the University of Koblenz.

All in all, 877 users participated of which 582 finished tagging all 10 web pages. For 530 users, also the grouping of the web pages according to their similarity is available. In Section 6, we only use the tag assignments and groupings of those 582 users who finished tagging all 10 web pages. According to a questionnaire at the end of the experiment, approximately 53% of the participants use tagging systems for searching regularly or sometimes. The rest tried it either once or not all. Furthermore, 45% of the participants upload content to tagging systems regularly or sometimes.

Due to our recruiting strategy, we expected to observe a homogeneous subgroup of native German speakers. Thus, we decided to not only offer an English variant of our experiment but also a German variant. In both variants, the same English web pages were shown but in the German variant we asked the participants to preferably use German keywords. Thus, German participants were able to use their larger and more accurate active German vocabulary during tagging. Each participant decided on his own whether to participate in the German or English variant.

All in all, 231 users finished the experiment in the German variant and 351 users in the English variant (see Tab. 3). It was the objective of our recruiting strategy to recruit around 100 participants for each of our experimental conditions because usually this number of participants is required until the tag vectors reach a stable state (cf. [5]). For the German variant, we recruited slightly less participants than our target value of 100 participants per experimental condition. For the English variant, we recruited slightly more participants. But an a posteriori analysis of our results showed that we nevertheless reached for all experimental conditions the primary objective of having stable tag vectors.

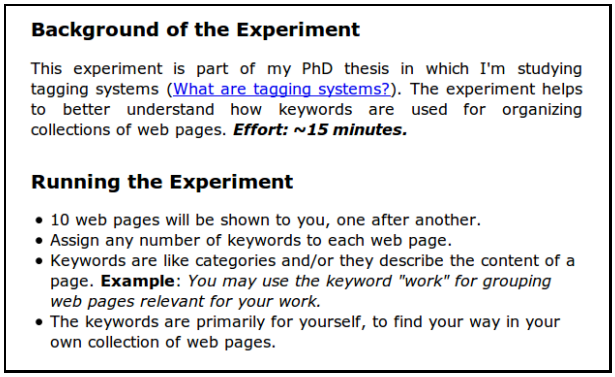

Figure 4: Instructions given to the participants of the English experiment variant.

After choosing between the German or the English variant of the experiment, each participant was randomly assigned to one of the three conditions described in Subsection 5.1. The experimental condition with the most participants was excluded from the random assignment, if it already contained at least 5 participants more than the condition with the fewest participants. This ensured a balanced distribution of participants over the experimental conditions.

The participants were not aware that different experimental conditions exist and that they have to create topical clusters at the end of the experiment. They were only told that the experiment analyses how keywords are used for organizing collections of web pages (see Fig. 4).

6. RESULTS

In this section, we are showing the results of our user experiment. The results help us in validating the hypotheses from Section 4. In a first step, we evaluate in Subsection 6.1 in how far the users from the different experimental conditions have identified similar topical clusters. The identification of similar topical clusters is a precondition for comparing the inter-resource consistency between the different experimental conditions in Subsection 6.2. Finally, in Subsection 6.3 we compare the inter-indexer consistency between the different experimental conditions.

6.1 Similarity of the Topical Clusters

In Subsection 3.1.2, we have described how to use the average Silhouette Coefficient $E(s_i)$ for measuring the inter-resource consistency of the tag assignments. But before we can apply this method on our data, we have to verify that the participants of the different experimental conditions have in average identified the same topical clusters during the second phase of the experiment (see Fig. 3). Otherwise, the differences in the $E(s_i)$-values may not only be caused by the influence of the respective experimental condition but also by differences in the topical clusters.

During the second phase of the experiment, we received feedback from 530 of our participants. A user was only able to finish the second phase if every web page was assigned to one cluster. The participants were allowed to provide a name for each cluster in order to make it easier for them to keep track of their clusters. On average, each participant separated the 10 web pages into 4.76 clusters, i. e. 2,521 clusters have been created. Together, the users identified 140 distinct clusters. Two topical clusters are considered as equal if they contain the same web pages.

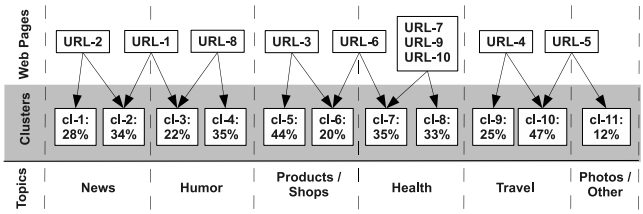

Figure 5: Visualization of the 11 most frequently identified clusters of web pages. Each box in the gray area corresponds to one cluster. Within the box of each cluster it is given, by how many participants the cluster has been identified. For example, 28% of all experiment participants put the *BBC* web page (URL-2) alone into a cluster, leading to cluster *cl-1*. Another 34% of the participants instead decided to group *BBC* (URL-2) together with *The Onion* (URL-1), leading to cluster *cl-2*. The remaining 38% of the participants have put URL-2 into other, less frequent clusters. Nevertheless, an analysis of the names used for *cl-1* and *cl-2* reveals that both clusters are seen as related to the *News* topic.

In Fig. 5, the eleven most frequently identified clusters from the second experiment phase are shown. Altogether, the eleven clusters from Fig. 5 represent 70.25% of all identified topical clusters. According to the names of the clusters, the 10 web pages are roughly related to 6 different topics. URL-1, URL-5 and URL-6 are each on the border between two topics. For example, the web page *The Onion* (URL-1) publishes satirical news articles. 34% of the users think that it is more related to the *News*-topic and thus they group it with an article from the *BBC* web page (URL-2), leading to cluster *cl-2*. In contrast, 22% emphasize more the *Humor*-topic and thus group it with *Patently Silly* (URL-8) which lists funny and strange patents, leading to cluster *cl-3*.

In Fig. 5, the reported cluster probabilities are based on all 530 participants who completed the second phase of the experiment. But in Fig. 6 it can be seen that in the German variant the vast majority of the participants perceive *The Onion* (URL-1) as related to *BBC* (URL-2) and the *News* topic. In contrast, participants of the English variant more prefer to cluster *The Onion* with *Patently Silly* (URL-8) according to the *Humor* topic. This preference for clustering *The Onion* together with *Patently Silly* is even more prevalent for the English *Popular Tags* condition (see Fig. 7).

But in how far are these differences in the cluster probabilities significant? In the following, we use the χ^2-Test [2, p. 199ff] for answering this question.[2] If the χ^2-Test rejects the hypothesis of equal cluster probabilities then we cannot compare $E(s_i)$-values for those experimental conditions.

In a first test, we compare the clusterings from the English variant of the experiment with those from the German variant. The test reveals that the clusterings differ significantly ($T = 161.69$, $n_1 = 1519$, $n_2 = 1002$, $p < 0.01$). Thus, we cannot use the Silhouette Coefficients for comparing the inter-resource consistency across the two language variants.

[2]For the χ^2-Test, we counted how often each of the distinct clusters has been identified by the different participants. The probability of all clusters identified by only a single user from either of the compared participant groups is combined in a single cluster "other". This is necessary for preserving the validity of the χ^2-Test (see [2, p. 201f]).

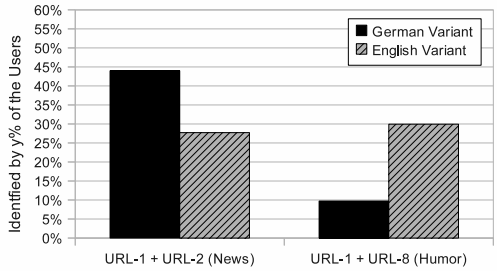

Figure 6: Differences in the clustering of URL-1 between the participants of the English and the German experiment variant. Participants of the German variant see it more as *news* related and thus cluster it with URL-2. In the English variant, the participants more emphasize its humorous aspects by clustering it with URL-8.

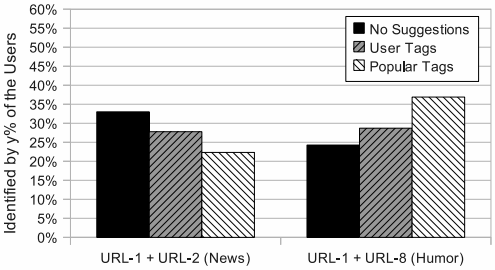

Figure 7: Differences in the clustering of URL-1 between the participants of the English experiment variant. Under the *Popular Tags* condition, more participants see URL-1 as related to *humor* and thus cluster it with URL-8.

But for evaluating our hypotheses from Section 4, it is more important whether we can use the Silhouette Coefficient for comparing the *No Suggestions* condition to the other two experimental conditions within the same language variant:

No Suggestions vs. Popular Tags Only for the German variant of the experiment the clusterings from the *No Suggestions* condition and from the *Popular Tags* condition can be considered as equal. For the English variant, the clusterings from the two conditions differ significantly. Possible explanations for the significant differences are discussed in Subsection 7.
German: $T = 39.25$, $n_1 = 339$, $n_2 = 323$, $p = 0.75$;
English: $T = 63.04$, $n_1 = 489$, $n_2 = 515$, $p = 0.06$

No Suggestions vs. User Tags For the English variant of the experiment as well as for the German variant the clusterings from the *No Suggestions* condition and from the *User Tags* condition can be considered as equal. For the German variant, the differences between the clusterings are smaller than for the English variant.
German: $T = 35.03$, $n_1 = 339$, $n_2 = 340$, $p = 0.86$;
English: $T = 51.99$, $n_1 = 489$, $n_2 = 515$, $p = 0.36$

All in all, the results in this subsection show that there are only minor differences in the cluster probabilities between the three German experimental conditions. Thus, we can compare the $E(s_i)$-values between all three German experimental conditions. In contrast, in the English experiment variant we can only compare the *No Suggestions* and the English *User Tags* condition. The English *No Suggestions*

German	$E(s_{x,i})$	$E(tr_{x,i})$
No Suggestions	0.1847	2.44
User Tags	0.2367	2.39
Popular Tags	0.1474	3.60
English		
No Suggestions	0.1713	2.76
User Tags	0.1915	2.68
Popular Tags	N/A	4.67

Table 4: Influence of the experimental conditions on the inter-resource consistency and on the inter-indexer consistency. The inter-resource consistency is measured by the average Silhouette Coefficient $E(s_{x,i})$. The inter-indexer consistency is measured by the average Tag Reuse Rate $E(tr_{x,i})$.

and the English *Popular Tags* condition cannot be compared because of the differences in the identified topical clusters.

6.2 Measuring the Inter-Resource Consistency

In this subsection, we evaluate the hypotheses from Section 4 which are related to the influence of tag suggestions on the inter-resource consistency. In the following, we use the average Silhouette Coefficient $E(s_{x,i})$ from Subsection 3.1.2 for measuring the inter-resource consistency for a tagging system X. We compare the $E(s_{ns,i})$-value for the *No Suggestions* condition with the $E(s_{pt,i})$-value for the *Popular Tags* condition and/or the $E(s_{ut,i})$-value for the *User Tags* condition. Based on our hypotheses from Section 4, we expect the following order of the $E(s_{x,i})$-values:

$$E(s_{pt,i}) < E(s_{ns,i}) < E(s_{ut,i}) \quad (3)$$

For the German experiment variant, we compute the $E(s_{x,i})$ values by comparing the tag vectors of the three experimental conditions against the union of all clusters given by participants of the German variant. This way the differences between the $E(s_{x,i})$ values are only caused by differences in the tag vectors and not also by slight differences in the cluster probabilities. For the English variant, we compare the tag vectors from the *No Suggestions* and the *User Tags* condition against the union of the clusters from the respective participants. The clusters and tag vectors from the English *Popular Tags* condition have to be excluded from the evaluation because of the significant differences in the cluster probabilities as it is shown in Subsection 6.1.

For comparing the $E(s_{x,i})$-values of two experimental conditions, we apply a two-tailed Mann-Whitney Test [2, p. 272ff]. It tests the null hypothesis whether two given $E(s_{x,i})$ values have to be considered as equal against the alternative hypothesis that they are not equal. Furthermore, we use the Hodges-Lehmann Estimator of Shift [2, p. 281f] for determining the 95% confidence interval for the difference between the two $E(s_{x,i})$-values.

A summary of the experimental results is shown in Tab. 4 and 5. For these results, we have restricted the number of users so that under each of the experimental conditions the same number of users contributed to the tag vectors. For the German variant, we restricted it to the first 74 users of each of the experimental conditions. For the English variant, we restricted it to the first 115 users. Thus, we control that different numbers of users do not cause the differences in the results. Controlling for vocabulary size and number of tag assignments only led to minor fluctuations in Tab. 4 and 5 so we omit the numbers here.

	German Variant	English Variant
$E(s_{ns,i}) - E(s_{pt,i})$	$[0.0337, 0.0582]$	N/A
$E(s_{ut,i}) - E(s_{ns,i})$	$[0.0584, 0.0955]$	$[0.0106, 0.0434]$
$E(tr_{ns,i}) - E(tr_{pt,i})$	$[-1.611, -0.691]$	$[-2.403, -1.276]$
$E(tr_{ut,i}) - E(tr_{ns,i})$	$[-0.392, 0.3437]$	$[-0.392, 0.2804]$

Table 5: 95% confidence intervals for the differences between the $E(s_{x,i})$-values and the $E(tr_{x,i})$ values under the experimental conditions.

The exact method for computing the $E(s_{x,i})$-values is as follows: Given a set of clusters which was provided by an individual user, for all ten web pages we compute the respective s_i-value. For example, in case of the German experiment variant we overall have 216 sets of clusters which have been provided by the users in this experiment variant. Thus, the $E(s_{x,i})$-values for the German experiment variant are each based on $10 \cdot 216 = 2160$ s_i-values.

When computing the $E(s_{x,i})$-values, we omit from our analysis the s_i-values of web pages which are in a cluster of size 1 in their respective set of clusters. The reason is that in such a case a_i and subsequently the s_i-value are not well defined (see [12]). For example, in case of the German experiment this excludes 334 of the 2160 s_i-values from our experiment. Nevertheless, the tag vector of the respective web page still takes part in the computation of the b_i-value for the s_i-values of the remaining web pages in the same set of clusters. Furthermore, the s_i-value of the respective web page may still be computed in the context of another set of clusters where it is not in a cluster of size 1.

6.2.1 No Suggestions vs. Popular Tags

In the following, we test the effect of suggesting popular tags on the average Silhouette Coefficient. For this purpose, we compare the $E(s_{ns,i})$-value for the *No Suggestions* condition to the $E(s_{pt,i})$-value for the *Popular Tags* condition as they are shown in Tab. 4. Because of the differences in the perception of the web pages for the English variant of the experiment (see Subsection 6.1) we can test Hypothesis 3 for the German experiment variant only.

For the German experiment variant, we can confirm that $E(s_{pt,i}) < E(s_{ns,i})$. A two-tailed Mann-Whitney Test shows that the difference between the $E(s_{x,i})$-values is significant ($T_1 = 7.4157$, $n = m = 1798$, $p < 0.01$). According to the Hodges-Lehmann Estimator of Shift, suggesting popular tags decreases the average Silhouette Coefficient by 0.0472 with a 95% confidence interval of $[0.0337, 0.0582]$.

Thus, our experimental results show that recommending the seven most popular tags of a resource has a significant influence on the indexing quality. The results support Hypothesis 3 that recommending the popular tags decreases the inter-resource consistency in tagging systems.

6.2.2 No Suggestions vs. User Tags

Now, we test the effect of suggesting the user his/her own previously used tags on the average Silhouette Coefficient. We compare the $E(s_{ns,i})$-value for the *No Suggestions* condition to the $E(s_{ut,i})$-value for the *User Tags* condition as they are shown in Tab. 4. For both language variants of the experiment, we can confirm that $E(s_{ns,i}) < E(s_{ut,i})$.

For both language variants, the two-tailed Mann-Whitney Test shows that the difference between the $E(s_{x,i})$-values is significant (German: $T_1 = -8.11$, $n = m = 1796$, $p < 0.01$;

English: $T_1 = -3.0563$, $n = m = 1721$, $p < 0.01$). For the German variant, suggesting the user his/her own previously used tags increases the average Silhouette Coefficient by 0.0775 with a 95% confidence interval of $[0.0584, 0.0955]$. For the English experiment variant, the average Silhouette Coefficient increases by 0.0306 with a 95% confidence interval of $[0.0106, 0.0434]$.

Thus, our experimental results show that suggesting the user his/her own previously used tags has a significant influence on the indexing quality. The results support Hypothesis 1 that recommending the user's tags increases the inter-resource consistency in tagging systems.

6.3 Measuring the Inter-Indexer Consistency

In this subsection, we evaluate the hypotheses from Section 4 which are related to the influence of tag suggestions on the inter-indexer consistency, as it might be measured by the average tag reuse rate $E(tr_{x,i})$ from Subsection 3.2. We argue that in the presence of tag suggestions one cannot automatically assume a positive correlation between the inter-resource and the inter-indexer consistency. Thus, the inter-indexer consistency is not suitable for measuring the indexing quality in tagging systems if the users have been influenced by tag suggestions. According to our hypotheses from Section 4, we expect that the *Popular Tags* recommender helps to increase the inter-indexer consistency. In contrast, we expect that the *User Tags* recommender either leads to a decreased or unchanged inter-indexer consistency. Overall, we expect the following order of the $E(tr_{x,i})$-values:

$$E(tr_{ut,i}) \leq E(tr_{ns,i}) < E(tr_{pt,i}) \quad (4)$$

6.3.1 No Suggestions vs. Popular Tags

In the following, we test the effect of suggesting popular tags on the average Tag Reuse Rate. For this purpose, we compare the $E(tr_{ns,i})$-value of the *No Suggestions* condition to the $E(tr_{pt,i})$-value of the *Popular Tags* condition as they are shown in Tab. 4. For both language variants of the experiment, we can confirm that $E(tr_{pt,i}) > E(tr_{ns,i})$.

For both language variants, the two-tailed Mann-Whitney Test shows that the difference between the $E(tr_{x,i})$-values is significant (German: $T = 58$, $n = m = 10$, $p < 0.01$; English: $T = 55$, $n = m = 10$, $p < 0.01$). For the German variant, suggesting popular tags increases the Tag Reuse Rate by 1.2274 with a 95% confidence interval of $[0.6912, 1.6111]$. For the English variant, the average Tag Reuse Rate increases by 1.7955 with a 95% confidence interval of $[1.2760, 2.4027]$.

Thus, our experimental results show that the suggestion of the seven most popular tags of a resource has a significant influence on the inter-indexer consistency. The results support Hypothesis 4 that suggesting the popular tags increases the inter-indexer consistency.

6.3.2 No Suggestions vs. User Tags

Now, we test the effect of suggesting the user his/her own previously used tags on the average Tag Reuse Rate. We compare the $E(tr_{ns,i})$-value for the *No Suggestions* condition to the $E(tr_{ut,i})$-value for the *User Tags* condition (see Tab. 4). For both language variants of the experiment, we can confirm that $E(tr_{ut,i}) \leq E(tr_{ns,i})$.

For the German variant, we cannot reject the hypothesis of observing equal $E(tr_{x,i})$-values for the two conditions (Mann-Whitney, $T = 104$, $n = m = 10$, $p = 0.97$ two-tailed). Accordingly, the 95% confidence interval for $E(tr_{ut,i}) - E(tr_{ns,i})$ is $[-0.392, 0.3437]$. Also for the English variant, we cannot reject the hypothesis of observing equal $E(tr_{x,i})$-values for the two conditions (Mann-Whitney, $T = 113$, $n = m = 10$, $p = 0.57$ two-tailed). In this case, the 95% confidence interval for $E(tr_{ut,i}) - E(tr_{ns,i})$ is $[-0.392, 0.2804]$.

Thus, our experimental results support Hypothesis 2 that $E(tr_{ut,i}) \leq E(tr_{ns,i})$. Given the current data set, $E(tr_{ut,i}) = E(tr_{ns,i})$ has to be favored over $E(tr_{ut,i}) < E(tr_{ns,i})$ because the difference in the $E(tr_{x,i})$-values is not significant.

7. DISCUSSION

In Section 6, we have presented the results how different kinds of tag suggestions influence the inter-resource and the inter-indexer consistency in tagging systems. But in our experiment, we also discovered effects of tag suggestions which cannot be measured by our proposed methodology because in case of the English *Popular Tags* condition the suggestions not only influenced the tag vectors but also the topical clusters of the participants. It has been shown in Fig. 7 that the participants of the *Popular Tags* condition have a significantly higher probability to cluster *The Onion* according to its humorous aspects than the other participants.

It seems plausible that these differences are due to the influence of the tag suggestions. Indeed, under the English *Popular Tags* condition, for 107 participants the list of suggested popular tags contained the tag "satire". Additionally, the tag "fun" was contained 104 times in the list and "humor" 89 times. We assume that these tags helped to increase the likelihood of recognizing the humorous aspects of *The Onion* and of clustering it with *Patently Silly*. It seems that seeing the tags not only changed the users' vocabulary for describing the resource but also how they understood it.

But why didn't we observe a similar effect for the German *Popular Tags* condition? It seems that in the German experiment variant not enough participants recognized the humorous aspects of *The Onion* in order to push such tags into the list of popular tags. A possible reason is the overall lower probability of describing the humorous aspects of *The Onion* in the German experiment variant (see Fig. 6). Indeed, in the German *Popular Tags* condition the list of popular tags contains only for one participant a tag related to the humorous aspects, namely the tag "lustig" (=funny). Consequently, we do not observe an increased probability of clustering *The Onion* with *Patently Silly* when compared to the other German experimental conditions. Quite contrary: The dominance of news related tags in the list of popular tags for the German *Popular Tags* even decreases the probability of clustering *The Onion* with *Patently Silly* from 13% for the other two German experimental conditions to 3%.

All in all, it thus seems that suggesting popular tags has the potential to not only influence the tag vectors but also how users understand web pages, i. e. tag suggestions may lead to permanent learning effects as they are also discussed in [4]. This may potentially have a positive effect on the indexing quality but it is not measurable with our proposed methodology of measuring the inter-resource consistency. But our experiment also suggests that certain preconditions have to be fulfilled for learning effects to occur: We only observed it for a single web page and only in the English experiment variant. It would be subject to further research to identify these preconditions and to study in how far it is a regular effect or rather an exception.

8. CONCLUSIONS

In this paper, we have discussed how to measure the influence of tag recommenders on the indexing quality of tagging systems. We have proposed to use the inter-resource consistency as the main target parameter to be optimized by tag recommenders because it influences the precision and recall of queries in a tagging system [19]. Improving the inter-indexer consistency should only be a secondary target of tag recommenders. We have applied our methodology for measuring the inter-resource and inter-indexer consistency for two exemplary baseline recommenders: (1) The *Popular Tags* recommender which recommends the seven most popular tags of a resource, and (2) the *User Tags* recommender which recommends a user his/her previously used tags.

During our user experiment with 582 participants, we have contrasted our measure of the inter-resource consistency with a measure of the inter-indexer consistency. In the literature about tagging systems, the inter-indexer consistency is often used as a measure of indexing quality. But we have shown that the inter-indexer consistency is not positively correlated with the inter-resource consistency if users are influenced by tag recommendations. In case of the popular tags, the recommendations increased the inter-indexer consistency and decreased the inter-resource consistency. For the user tags, the recommendations had no influence on the inter-indexer consistency while they increased the inter-resource consistency.

From these results of the user experiment one can conclude that the tag vectors of related resources get more dissimilar to each other if popular tags are recommended. In contrast, the tag vectors of related resources get more similar to each other if the user tags are recommended. Thus, the user tags would not only improve the retrieval results if a user searches in his own collection, as one might expect, but also if he searches for resources tagged by other users.

The only precondition for this positive effect during retrieval is that users have a similar judgment of the relevance of resources to each other, i. e. that they form similar topical clusters. But our results have also shown that this is reasonable to assume because we couldn't measure significant differences in the topical clusters between our experimental conditions. The only exception was the English *Popular Tags* condition but there this effect is restricted to a single web page for which different ways of looking at it exist.

9. SUPPLEMENTAL MATERIAL

The data set from our user experiment can be downloaded: http://west.uni-koblenz.de/Research/DataSets/tagging-experiment/. The experiment interface is still accessible at http://userpages.uni-koblenz.de/~klaasd/experiment/.

10. ACKNOWLEDGMENTS

We thank Thomas Gottron, Isabella Peters, Julia Preusse and Ansgar Scherp for their feedback, as well as all participants of the experiment for donating their time. This work has been co-funded by the German Research Foundation (DFG) under the Multipla project (grant 38457858) and by the EU in FP7 in the ROBUST project (grant 257859).

11. REFERENCES

[1] D. Bollen and H. Halpin. An Experimental Analysis of Suggestions in Collaborative Tagging. In *International Joint Conference on Web Intelligence and Intelligent Agent Technologies*, 2009.

[2] W. Conover. *Practical Nonparameteric Statistics*. John Wiley, 3rd edition, 1999.

[3] F. Floeck, J. Putzke, S. Steinfels, and K. Fisch. Imitation and Quality of Tags in Social Bookmarking Systems – Collective Intelligence Leading to Folksonomies. In T. Bastiaens, U. Baumöl, and B. Krämer, editors, *On Collective Intelligence*. Springer, 2010.

[4] W.-T. Fu and W. Dong. From Collaborative Indexing to Knowledge Exploration: A Computational Social Learning Model. *IEEE Intelligent Systems*. In Press.

[5] S. Golder and B. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208, 2006.

[6] T. Kannampallil and W.-T. Fu. Trail Patterns in Social Tagging Systems: Role of Tags as Digital Pheromones. In *International Conference of Human-Computer Interaction*, 2009.

[7] T. Kowatsch and W. Maass. The Impact of Pre-Defined Terms on the Vocabulary of Collaborative Indexing Systems. In *European Conference on Information Systems*, 2008.

[8] M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag Sources for Recommendation in Collaborative Tagging Systems. In *ECML PKDD Discovery Challenge*, 2009.

[9] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[10] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Hypertext Conference*, 2006.

[11] A. Mathes. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Website, December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[12] P. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. of Comput. and Applied Mathem.*, 20:53 – 65, 1987.

[13] J. Saarti. Consistency of Subject Indexing of Novels by Public Library Professionals and Patrons. *Journal of Documentation*, 58:49–65, 2002.

[14] S. Sen, S. Lam, A. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Conference on Computer Supported Cooperative Work*, 2006.

[15] F. Suchanek, M. Vojnovic, and D. Gunawardena. Social Tags: Meaning and Suggestions. In *Conference on Information and Knowledge Management*, 2008.

[16] M. Tatu, M. Srikanth, and T. D'Silva. RSDC'08: Tag Recommendations using Bookmark Content. In *ECML PKDD Discovery Challenge*, 2008.

[17] M. Vojnovic, J. Cruise, D. Gunawardena, and P. Marbach. Ranking and suggesting popular items. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1133–1146, 2009.

[18] H. White and B. Griffith. Quality of Indexing in Online Data Bases. *Information Processing & Management*, 23(3):211–224, 1987.

[19] P. Zunde and M. Dexter. Indexing Consistency and Quality. *American Documentation*, 20:259–267, 1969.