Social Tags: Meaning and Suggestions

Fabian M. Suchanek*
Max-Planck-Institut
Saarbrücken, Germany
suchanek@ mpii.mpg.de

Milan Vojnović
Microsoft Research
Cambridge, UK
milanv@ microsoft.com

Dinan Gunawardena
Microsoft Research
Cambridge, UK
dinang@ microsoft.com

ABSTRACT

This paper aims to quantify two common assumptions about social tagging: (1) that tags are "meaningful" and (2) that the tagging process is influenced by tag suggestions. For (1), we analyze the semantic properties of tags and the relationship between the tags and the content of the tagged page. Our analysis is based on a corpus of search keywords, contents, titles, and tags applied to several thousand popular Web pages. Among other results, we find that the more popular tags of a page tend to be the more meaningful ones. For (2), we develop a model of how the influence of tag suggestions can be measured. From a user study with over 4,000 participants, we conclude that roughly one third of the tag applications may be induced by the suggestions. Our results would be of interest for designers of social tagging systems and are a step towards understanding how to best leverage social tags for applications such as search and information extraction.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation, Theory

Keywords

Social tagging, search

1. INTRODUCTION

1.1 Motivation

Social tagging has recently received a wide adoption by various Web 2.0 services such as social book-marking and the tagging of blogs, photos, music, and videos. While in many of these applications the primary goal is to serve the needs of individual users (e.g. the organization of personal bookmark collections and their later retrieval) the idea is that the tags should also help other users to browse, categorize and find items. Furthermore, tags are used for information discovery, sharing, and community ranking. Going beyond that, tags could be useful for tasks such as search, navigation or even information extraction.

However, it is not obvious how tags can be exploited best for these tasks because the *nature of tags* is not entirely clear. Wikipedia, for example, broadly relates tags to the *description*, *classification*, and *search* of information[1]. Other descriptions stress the use for personal *organization*[2] or for *re-finding*[3] items. Thus, it is not clear what purpose the tags actually serve. Given this freedom, users tag according to their own gusto. As a result, the choice of tags varies widely: Sometimes the tags identify an item (like "Madonna"), sometimes they identify the owner of the item, and sometimes they give subjective assessments (like "funny") [5]. They may also be purely organizational (like "toread") or completely unintelligible to another person (like "#####"). Such organizational and personal tag creations would be of less use for semantic applications. These applications would rather require tags that carry "meaning" in some sense. However, it is not clear what proportion of tags actually falls in this category. In other words, the question is

(RQ 1): How "meaningful" are tags?

To some degree, social tagging environments can steer the tagging process by providing tag suggestions. The suggestions guide users based on the tags of other users. Since the suggestions act as a feedback loop, they may be powerful enough to amplify trends or possibly also to distort them. They may boost the number of meaningful tags or they may bring forward less meaningful creations. As the functionality of the suggestion mechanism is under the control of the system designer, this opens up the possibility of influencing the resulting tags and their meaningfulness. Thus, the question is

(RQ 2): In what ways are users influenced by tag suggestions?

In this paper, we aim to shed light on these issues.

*Work performed in part while an intern with Microsoft Research Cambridge.

[1] http://en.wikipedia.org/wiki/Tag_(metadata)
[2] http://del.icio.us/help/tags
[3] http://flickr.com/help/tags

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10 ...$5.00.

1.2 Contributions

We first study the semantic properties of tags in detail. Then, we develop a model of how tag suggestions influence the user. This allows us to quantify the influence of tag suggestions. It also allows us to study the influence of different interface design choices. Our analyses are based on (1) a corpus of search keywords, contents, titles, and tags applied to several thousand popular Web pages and (2) results of a Web-based user study that we conducted with 4,000+ participants.

Our main results are as follows:

• We found that up to 50% of the tag applications may be "not meaningful". This contrasts with much lower proportions of non-meaningful terms in document content and queries.

• Our analysis shows that the more popular a tag is, the more likely it is to be meaningful. In other words, aggregating the top tags of a document biases to filtering out the meaningful tags. This is not be a priori clear as some non-meaningful words can be rather common (such as "toread" or "todo").

• Our analysis also validates that the more users tagged a document, the more meaningful the top popular tags are. However, we show that the meaningfulness increases significantly only if the document is tagged by more than 100 people. This may have consequences for small-scale tagging scenarios (such as enterprize environments).

• Our analysis indicates that tags applied to a document typically intersect more with the queries and the title than with the content. This suggests that social tags could prove useful for search applications.

• We propose a novel metric (the *imitation rate*) to estimate how much the user was influenced by the tag suggestions. Note that it is non trivial to estimate this parameter, because the user may apply tags that have been suggested without actually paying attention to the suggestions.

• Our metric allows us to conclude that up to 1 in 3 tags may be induced purely by presence of tag suggestions. This result implies that the popularity of tags observed in existing systems with tag suggestions may be skewed. On the positive side, it implies that tag suggestions could provide an effective control over the tags.

• We further evaluated various other factors that affect the tagging process, such as the ordering of the suggested tags, the suggestion set size, and the user interface design.

The rest of the paper is structured as follows: We first give an overview of related work in the domain of social tagging. In Section 2, we present our data sets and methodology. Section 3 characterizes the semantic properties of tags. Section 4 introduces our probabilistic model of tagging and finally, Section 5 presents the results of our user study.

1.3 Related Work

Social tagging has attracted much attention lately. Golder and Huberman [5] see tags as a flat organizational structure, as opposed to a tree-like categorization structure. They propose categories for the purposes that tags serve and analyze the emergence of consensus among taggers. Chi [2] et al. analyze three wisdom-of-crowds areas, one of which are the tags used in del.icio.us. They analyze the entropy of tags and the relationship between rare and frequent tags. Sen et al. [11] study the tagging process from a user's perspective. They propose a model of components that have an influence on the user and analyze how users react to different types of tag suggestions. However, their analysis remains mainly on the level of 3 classes of tags. Xu et al. [16] introduce the concept of tags as *facets*, i.e. values for attributes such as `title`, `composer` and `artist` for a music piece. They propose a taxonomy of tags and derive desiderata for "good" tags. Finally, they propose a tag suggestion algorithm that aims at maximizing the proportion of such tags. Santos-Neto et al. [9] study the tagging systems CiteULike and Bibsonomy. They analyze the tagging activity and vocabulary size of users and define a neighborhood of users. They show that the neighborhood is a powerful tool for predicting tag applications. Mika [7] proposes to model a tagging system as a tripartite graph of users, tags and items. He proposes methods to extract ontological knowledge from tags. Sen et al. [10] propose feedback methods by which users can rate the quality of tags. Their user study allows them to derive desiderata for tagging interfaces. In summary, even though many interesting aspects of tagging have been studied, to our knowledge, the semantic properties of tags and the influence of tag suggestions have not been *quantified* so far.

2. DATA AND METHODOLOGY

2.1 Data

2.1.1 The Dictionary

To assess the semantic properties of tags, we use the semantic databases YAGO [13] and WordNet [4]. For our purpose, we consider them dictionaries that associate to a word one or multiple meanings. For example, the word "Java" can mean the programming language, the coffee or the island. Beyond that, WordNet also assigns to each word one or multiple word classes. For example, the word "house" belongs to the class of nouns (when used in the sense of "building") and to the class of verbs (when used in the sense of "to house somebody"). WordNet contains 155,287 words, while YAGO, partly intersecting with WordNet, contains 2.8 million words. We have combined YAGO and WordNet and refer to this combination as *the dictionary.* WordNet is designed to cover the lexical words of the English language. YAGO builds on Wikipedia and is designed to cover well-known proper names (such as names of cities, famous people and organizations). Hence, we assume that our dictionary covers the most common English words.

2.1.2 Data Sets

We obtained a sample of the logs of popular queries to an Internet search engine [15]. The logs associate to each query the first result page that the user clicked on. Our sample contains 1,600 Web pages with their associated queries (data set QUERY).

del.icio.us[4] is a popular online service, which allows users to bookmark Web pages and tag them. We use the tagging histories of about 65,000 Web pages in del.icio.us (data set TAG). For these pages, we also downloaded their titles and HTML contents from the Web, where this was possible (data sets TITLE and CONTENT).

DMOZ[5] is a Web page directory that is edited by volunteers. It classifies Web pages into a tree of categories (with category names like "arts", "sports" or "computers"). We

[4] `http://del.icio.us`

[5] `http://dmoz.org`

collected for each page the category names on its path to the tree root. Furthermore, DMOZ gives a description for each page. We downloaded the whole directory, yielding categories and descriptions for 3,900,000 Web pages (data sets DMOZCAT and DMOZDES).

We preprocessed all data sets by joining compound words (such as "car race") by help of the dictionary. After that, we eliminated stop-words. Note that our data sets do not necessarily cover the same Web pages.

2.1.3 User Study

To study the tagging behavior in a controlled environment, we conducted a Web-based user study. We asked employees of Microsoft Research to participate and we put advertisements for the study on the Microsoft Research Web page. Overall, 4,000 Internet users participated, roughly half of them Microsoft employees. For the study, we prepared a pool of 20 predefined Web pages, chosen from lists of popular Web pages. Furthermore, we designed different settings under which a page can be tagged. For example, in one setting the system provides tag suggestions while in another one it does not. The settings differ in the layout of the tagging interface, in the number of suggested tags and in the methods used to suggest tags. When a new participant registers for the study, we generate a sequence of random pages from the pool. We also generate a sequence of random settings. Then, the participant is asked to tag each page in the sequence under its corresponding setting. We also collected general information on tagging habits. We are well aware that tagging in the context of a user study may differ from tagging in a social system. However, we believe that the insights that we gained from our user study give a valuable hint on the situation in real systems.

2.2 Methodology

2.2.1 Framework

Our work analyzes Web documents and associated meta data. These meta data may be tags, but we can also see the categories of DMOZ, the descriptions of DMOZ and the title of the document as meta data. Following [3], even the content of the document can be considered meta data. We also see search keywords that are used in a search engine to find a particular page as meta data attached to the document. The constituents of these meta data (e.g. the tags) will be called *terms*. A single occurrence of a term in the meta-data (e.g. one tag applied to one Web document) will be referred to as an *application*. A set of terms that are applied together (e.g. the tags applied by one user to one document) will be called an *event*. We represent the meta data of a document as a set of terms, where we associate with each term the number of times it has been applied to the document (its *frequency*).

2.2.2 Metrics

For a given document, the meta data acts as a function from terms to frequencies, a *frequency vector*. If f is a frequency vector, $f(t) \in \mathbb{R}$ denotes the frequency for term t and $f(t) \geq 0$ for all t. For a set T of terms, we define the shorthand notation $f(T) = \sum_{t \in T} f(t)$. For a non-trivial frequency vector f for which $f(t) > 0$ for some t, we use the notation $f^*(t) = f(t)/(\max_{t'} f(t'))$. We call the set of terms that are mapped to non-zero frequencies the *support* of f, denoted $supp(f) = \{t | f(t) \neq 0\}$. If $f(t) \in \{0,1\}$ for all t, we call f a *set*. We call the sum of all frequencies the *mass* of f, $mass(f) = f(supp(f))$. Let n_f denote the number of terms in the support of f. A frequency vector implicitly defines a *ranking* on its support, i.e. a sequence of terms $< t_1, ...t_{n_f} >$ such that $f(t_i) \geq f(t_j)$ for all $i \leq j$ (where ties are broken arbitrarily). We denote with $t_{f,i}$ the i^{th} term in the ranking of f. Often, a frequency vector $f(t)$ has to be compared to a "reference frequency vector" $g(t)$. We use the following metrics: To compare the **supports of the vectors** irrespective of their frequencies, we use the standard measures recall and precision,

$$rec(f,g) = |supp(f) \cap supp(g)| \cdot n_g^{-1}$$

$$prec(f,g) = rec(g,f)$$

These measures are useful for comparing two sets. To compare the **ranking of** f to the **support of** g (e.g.when g is a set), we use the *precision at* k:

$$prec@k(f,g) = \frac{|\{t_{f,1}, \ldots, t_{f,\min(k,n_f)}\} \cap supp(g)|}{\min(k,n_f)}.$$

The precision at k measures what proportion of the top k terms in the ranking of f is in the support of g. To compare the **ranking of** f (irrespective of the frequencies) to the **frequencies** of g, we use the normalized discounted cumulative gain (NDCG)[6]:

$$ndcg(f,g) = \frac{\sum_{i=1}^{n_f} \frac{g^*(t_{f,i})}{\log(i+1)}}{\sum_{i=1}^{n_g} \frac{g^*(t_{g,i})}{\log(i+1)}}.$$

The NDCG assumes that g gives a "gain" to each term and then measures how much "gain" the support of f delivers, giving higher weight to higher ranked terms. Furthermore, we define the *weighted recall at* k:

$$wrec@k(f,g) = \frac{\sum_{i=1}^{\min(k,n_f)} [f(t_{f,i}) > 0] \cdot g(t_{f,i})}{\sum_t g(t)}$$

Here, $[\cdot]$ denotes the Iverson bracket, which evaluates to 1 if the enclosed condition is true and to 0 else. The weighted recall at k measures what proportion of the mass of g is covered by the top k terms in the ranking of f.

To compare the **frequencies** of the two vectors, we use the well-known cosine similarity:

$$cos(f,g) = \left(\sum_t f(t) \cdot g(t)\right) \cdot \left(\sum_t f(t)^2\right)^{-\frac{1}{2}} \cdot \left(\sum_t g(t)^2\right)^{-\frac{1}{2}}$$

Furthermore, we introduce the *fuzzy recall*:

$$frec(f,g) = 1 - \frac{\sum_t \max(g^*(t) - f^*(t), 0)}{\sum_t g^*(t)}$$

The fuzzy recall punishes terms to which f assigns less frequency than g. If f assigns its maximum value to all terms, the fuzzy recall is 1. If the frequency vectors are sets, the fuzzy recall is identical to the standard recall[6].

All of the above metrics deliver values between 0 and 1. We decided not to use the KL-divergence and the Jenson-Shannon-divergence because they deliver unbounded values that are difficult to interpret. Other metrics that are often

[6] Precision measures can be defined analogously to all recall measures defined here, but are not necessary for our analyses.

Table 1: Data Sets

	TAG	CONTENT	TITLE	QUERY	DMOZCAT	DMOZDES
Applications	27659051	17609669	196769	732837	20892365	49450506
Pages	65211	48156	46494	1627	3896745	3914147
Terms	1137848	1392897	37866	53050	140070	1320781
Events	14196449	-	-	498739	4282334	4310575
Applications/Page	424.15	365.68	4.23	450.42	5.36	12.63
Avg Terms/Page	111.86	222.51	4.00	65.95	5.16	11.39
Applications/Event	1.95	-	-	1.47	4.88	11.47
Events/Page	217.70	-	-	306.54	1.10	1.10

used to compare two rankings are Kendall's Tau, the Kendall tau rank correlation coefficient, the footrule distance and the Spearman coefficient. However, these metrics pay equal attention to high-ranked discordant pairs and to low-ranked discordant pairs, which makes them less useful for our analyses.

3. THE SEMANTICS OF TAGS

3.1 Overview

Table 1 gives an overview of our data sets. For TAG, we observe that each page has been tagged on average by 217 users and that each user applied on average 1.95 tags per event (implying that many users tagged the page with just one term). The CONTENT contains on average 366 words per document (of which 223 are distinct). The TITLE has on average 4 words per document. QUERY shows us that each page was searched on average 306 times and that queries were on average 1.47 words long. There are different numbers of pages in DMOZCAT and DMOZDES and 10% of the DMOZ pages have been categorized multiple times.

3.2 Semantic Analysis

3.2.1 Meaningfulness

It has been hypothesized [5, 16, 7] that tags are often non-words (like "toread") and that polysemy is highly prevalent in tags. To assess these hypotheses, we looked up each tag application in our dictionary and counted how many meanings the tag had. As a point of reference, we conducted a similar analysis for CONTENT and QUERY and restricted the analyses to those pages that appear in all three data sets (1521).

Table 2: Applications per # of Meanings

# Meanings	TAG'	CONTENT'	QUERY'
unknown	54.62 %	19.45%	18.22 %
	Out of the applications known to the dictionary:		
1	13.35 %	13.25%	21.50 %
2	12.83 %	8.69%	13.28 %
3	7.45 %	6.80%	6.29 %
4	5.20 %	6.32%	4.90 %
5	3.79 %	5.69%	4.13 %
6	4.52 %	5.13%	3.66 %
7	2.34 %	3.48%	3.07 %
8	1.65 %	3.00%	2.07 %
9	3.86 %	3.10%	2.81 %
10 or more	45.02%	44.54%	38.30 %
# Applications	271,357	464,100	345,510

As Table 2 shows, more than half of the tag applications use words that are not known to the dictionary. This confirms that the proportion of personal tag creations (such as "toread"), misspellings, proper names and possibly also foreign words is indeed high in tags and may even be larger than 50%. Of course, our dictionary can only be a proxy for proper words. However, the comparison with the other data sets shows that the proportion of non-words is significantly higher in tags than in queries or the Web pages.

As a side-result, our analysis shows that among the words known to the dictionary, most have more than one meaning. Family names, e.g, can refer to dozens of well-known people in our dictionary. Given that a word may have even more meanings than our dictionary knows, this constitutes the proof that polysemy is indeed highly prevalent in del.icio.us. The high proportion of non-meaningful terms and the high polysemy may seem discouraging results for semantic applications. In the following, we show how the meaningful tags can be filtered out.

3.2.2 Meaningfulness Tracked Down

We considered the top k tags for each page and computed the precision with respect to our dictionary. The average of these values across pages (weighted by their popularity) is shown in Figure 1: The highest precision is found for small k. This shows that the top popular tags for a page are indeed the "meaningful" ones. Note that it is not a priori clear that this would necessarily hold, as some non-meaningful words may be in rather common use (such as personal organization tags like ”todo” and ”toread”). A possible reason might be that different taggers use different non-words for their personal organization, but use the same meaningful tags for a general description of the page. This observation is supported by the proportion of DMOZ category terms, which are also proportionally more prevalent in the popular tags. Since our dictionary gives a lower bound on the proportion of English words, we obtain as a side-result that, on average, at least 80% of the top 7 tags are proper English words. This implies that, on average, already the top 7 tags can be of use for semantic applications.

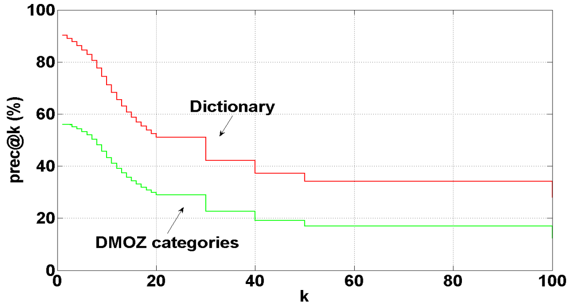

Figure 1: Precision@k for Tags (page average)

The above results suggest that a simple aggregation of the top tags of a document may be effective for extracting the meaningful tags. While this may be effective on average across all the pages, it may fail for pages that were tagged by only few users. We computed for each page the proportion of its top 7 tags that were meaningful in the sense of our dictionary. Figure 2 shows the result for the pages, grouped by the total number of tag applications that a page received. Our results suggests that the more popular a page is, the higher its proportion of known words is in the top 7 tags. (We obtain similar results when considering more top tags). Again, the observation is confirmed by the proportion of DMOZ category terms, which are also proportionally more prevalent in popular pages. This means that the top tags of popular pages are even more likely to be meaningful than the top tags of less popular pages. Interestingly, the precision increases only slowly for the first 100 tag applications. This entails that for less popular content or small scale systems such as in enterprise scenarios, simple aggregation of the top tags may not be sufficient to filter out the meaningful tags.

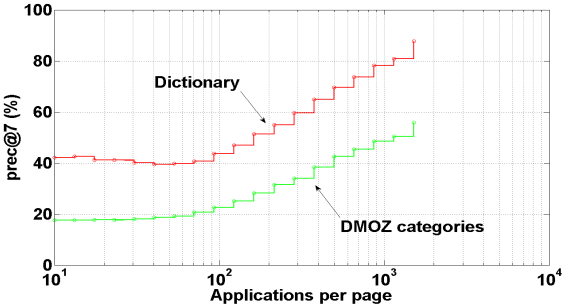

Figure 2: Precision@7 for tags, by page popularity

3.2.3 Word Classes

We were interested in the word classes that tags belong to. We distinguish the lexical classes nouns, verbs, adjectives and adverbs. We consider a term a plural noun if it can be stemmed by the PlingStemmer [12] so that the result is a known noun. Furthermore, we consider the class of URLs (as determined by an appropriate regular expression match). A term belongs to the class of categories, if it is used in the DMOZ category system. We also consider the class of proper names. Table 3 gives the percentage of tag applications that fall into these classes (one term can belong to multiple classes).

Table 3: Applications per Word Class

Word Class	TAG'	CONTENT'	QUERY'
URLs	2.29 %	2.81 %	6.58%
Categories	25.36 %	18.21 %	20.77 %
	Out of the applications known to the dictionary:		
Common Nouns	75.63 %	68.31%	55.00%
Pl. nouns	16.20 %	14.66%	15.66%
Sg. nouns	57.21 %	53.64 %	39.34%
Verbs	23.62 %	25.11%	18.42%
Adjectives	7.60 %	11.58%	9.12%
Adverbs	3.00 %	4.82 %	5.29%
Proper Names	70.10 %	69.91 %	74.06 %
# Applications	271,357	464,100	345,510

Out of the applications known to our dictionary, the applications roughly follow the same distribution as HTML content and queries: Common nouns are most prevalent, followed by proper names. Interestingly, a quarter of tag applications are terms that are also used as categories in DMOZ. This seems to suggest that users often think in terms of categories when they apply their tags.

3.3 Tags Cross-Compared

We wanted to know whether the distribution of tags resembles more the distribution of terms in the content or in the queries for a given page. To this end, we computed for CONTENT and QUERY the pages it had in common with the TAG data set (shown in the first line in Table 4). For each of the common pages, we computed the similarity of the frequency vector in the TAG data set to the frequency vector in the other data set, where the latter serves as the reference vector in the sense of Section 2.2.2. If we may see the frequencies as an indicator of how important the term is[7] to summarize the document (CONTENT) or to query for that document (QUERY), then the cosine similarity and the NDCG tell us that tags rather resemble the queries for that document than a summarization. The fuzzy recall shows that the terms that are frequent in the content are rather infrequent as tags, whereas the terms that are frequent in the queries are better covered by the frequent tags.

Table 4: TAG in Comparison

	CONTENT	QUERY
Common pages	47577	1535
Average cos	16.76 %	21.61 %
Average frec	4.94 %	19.64 %
Average NDCG	12.64 %	25.07 %

We were interested in the number of tags that are needed for a document in order to cover a substantial portion of the other meta data. For this purpose, we analyzed for each page the weighted recall of its top k with respect to the other data sets. The weighted recall at k is high if the top k tags cover frequent and thus "important" terms in the reference set. It is low if the tags cover only terms that are rare in the reference set. This measure is meaningful for both sets and distributions. The average of the weighted recall over pages (weighted by their popularity) is shown in Figure 3.

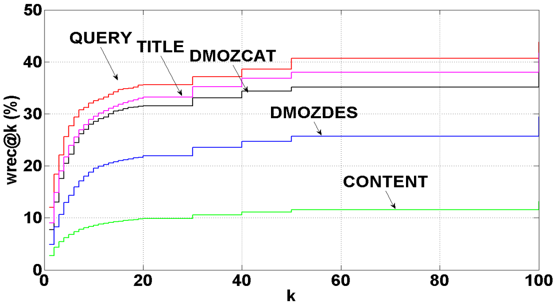

Figure 3: Weighted Recall@k of Tags

First, the plot confirms that tags cover better the queries for the document than its content. The recall for titles, descriptions and categories cannot be compared easily in

[7] Remember that stop-words have been removed.

this way, because these meta data are sets of different sizes. However, the analysis shows that in all of these cases the proportion of covered terms does not increase much beyond $k = 20$. The top 20 tags cover one third of the terms in the title and the categories and one fifth of the terms used by DMOZ to describe the document. Again, this confirms that most of the "useful" terms (as identified by the other data sets) appear in the top popular tags of a page.

In summary, our results suggest that tags provide useful summary information for documents that intersect well with the search keywords and titles and thus could be leveraged, e.g., to enhance search applications.

4. TAG GENERATION

4.1 Tag Suggestions

4.1.1 The Effects of Suggesting Tags

Many existing systems provide tag suggestions. del.icio.us, e.g., suggests the tags that have been most popular so far. On the one hand, tag suggestions may reduce tagging effort (cognitive or typing), thus serving as a participation incentive. They may also elicit conformance in vocabulary usage. On the other hand, suggestions may sway a user's tagging behavior and obscure his true preferences. Suggesting tags may also make users more passive and as a result make it harder for him to recall which tags he applied to a previously tagged item.

4.1.2 Suggestion Methods

We consider methods that suggest a set of tags of size k to each user, where k is a system configuration parameter. k is typically much smaller than the total number of distinct tags applicable to an object. We focus on methods that suggest tags for an object based solely on the tags applied by previous users to this document.

In our study, we consider four different suggestion methods, designed to span a broad range in their biasing to suggesting popular tags. First, we consider a method that makes no suggestions at all (NONE), which we use as a reference. Under this method, the tags applied by the users are not biased at all by tag suggestions. Second, we consider the standard *Top Popular* method (TOP), which to any user suggests the set of the k most popular tags. This method aims at best "guessing" what tags would be applied by the next user. The issue with this scheme, however, is that it may bias the tag popularity ranking if users tend to apply tags from the suggestion set. This implies that the resulting tag frequencies may become distorted relative to the frequencies that would hold if no suggestions were made (a phenomenon called the *popularity bias* [14]). This complicates the learning of a user's true preference of tags for an object.

Following [14], we consider two additional methods, which are designed to mitigate the tag popularity bias by suggesting some less popular tags than the top k tags. *Frequency Move-to-Set* (FMTS) is a method that behaves similarly to TOP, but tends to draw its suggestions from a larger set of popular tags than just the top k. It works as follows: For each document, it maintains a frequency vector of tags. The method will always suggest the top k tags in that vector for the document. After a tagging event, the method increments the frequencies of those tags in the vector that were applied by the user, but did not appear in the suggestion set. This way, the method facilitates the rise of previously unpopular tags while still biasing towards the popular tags in general.

The second method, *Move-to-Set* (MTS), allows even completely unpopular tags to appear in the suggestion set. Like FMTS, it maintains a frequency vector of tags per document. However, it maintains a threshold $\Theta \in \mathbb{N}$ and never increases the frequency beyond Θ. The suggestion set is exactly the set of tags with frequency Θ. If a tag is applied that has a frequency smaller than Θ, its frequency is incremented. If, this way, its frequency becomes Θ, the frequency of a random tag in the suggestion set is decremented, so that the overall number of tags in the suggestion set is always k. This way, any tag is catapulted to the suggestion set once it has been applied Θ times. This induces a high rate of perturbance in the suggestion set, while popular tags will still be a little bit more prevalent than less popular ones.

The theoretical properties of these methods are not in the focus of this paper. We refer the reader to [14]. Here, we only constate that TOP, FMTS and MTS cover a broad range in their biasing toward popular tags.

4.2 Measuring the Effect of Suggestions

Suggested tags may influence the user to a certain degree. This section establishes two measurements for this influence. The first one, the *matching rate*, serves to prove that users tag differently when suggestions are present. The second measurement, the *imitation rate*, aims to quantify the extend to which users are influenced by the suggestions.

For the following definitions, we consider a single fixed document D. We will denote a suggestion method by X (with, e.g., $X = \text{FMTS}$). We consider multiple tagging events $i = 1, ..., n$ by different users on D under the suggestion method X. Let $T(X, i)$ and $S(X, i)$ be the set of applied tags and the set of suggested tags at the ith event, respectively.

4.2.1 Matching Rate

The *matching rate* under the method X is the proportion of the applied tags that appear in the suggested tags, i.e.

$$mr(X) = \frac{\sum_{i=1}^{n} |T(X,i) \cap S(X,i)|}{\sum_{i=1}^{n} |T(X,i)|}$$

If all applied tags appear in the suggestion set, the matching rate is 1. The matching rate alone does not tell whether the tag generation was influenced by suggestions; in any case, some portion of the applied tags would appear in the suggestions and we expect schemes such as TOP to have a higher matching rate than schemes such as MTS that suggest also some less popular tags. In order to show that the tag generation was influenced by the suggestions of X, we consider a paired setting, in which users tag the document D in n tagging events under the method NONE. The *control matching rate* is the proportion of the applied tags under the method NONE that appear in the suggested tags of X[8]:

$$cmr(X) = \frac{\sum_{i=1}^{n} |T(NONE,i) \cap S(X,i)|}{\sum_{i=1}^{n} |T(NONE,i)|}$$

The matching rates can be averaged over tagging events and over documents. If the averaged matching rate and the

[8] Since these definitions are intended for large n and since the tagging events under NONE are independent, the order of tagging events for NONE does not matter.

averaged control matching rate differ in a statistically significant way, this means that users are more likely to choose tags from the suggestion set if the suggestion set is actually displayed. This, in turn, proves that users are influenced by the suggestions.

4.2.2 The Imitation Rate for a Suggestion Set

Note that the matching rate does not tell us what portion of applied tags is a result of the presence of the suggestions – even an uninfluenced user may happen to select a tag that was suggested. In order to quantify the extent of this influence, we introduce the *the imitation rate* for a single set of suggestions, S. Given a method X and a sequence of tagging events $i = 1, ..., n$, we consider only the tagging events in which S was suggested. We let $prec_n(X, S)$ be the portion of applied tags in these events that are in S:

$$prec_n(X, S) = \frac{\sum_{i=1}^{n} |T(X, i) \cap S|[S(X, i) = S]}{\sum_{i=1}^{n} |T(X, i)|[S(X, i) = S]}$$

We further consider the portion of applied tags that are in S under the suggestion method NONE:

$$prec_n(NONE, S) = \frac{\sum_{i=1}^{n} |T(NONE, i) \cap S|}{\sum_{i=1}^{n} |T(NONE, i)|}$$

Both $prec_n(X, S)$ and $prec_n(NONE, S)$ are standard precision metrics where the suggested tags are considered the “relevant items” and the applied tags are the “retrieved items”. Equivalently, $prec_n(X, S)$ and $prec_n(NONE, S)$ can be seen as recall metrics. In this case, the applied tags are the relevant items (selected by the user) and the suggested tags are the retrieved items (suggested by the system).

The imitation rate is defined by:

$$\alpha_n(S) = \frac{prec_n(X, S) - prec_n(NONE, S)}{1 - prec_n(NONE, S)} \tag{1}$$

We first briefly discuss the imitation rate, and then provide its justification. If the tags that the user applied and the suggested tags in the system X are statistically independent, then $prec_n(X, S)$ and $prec_n(NONE, S)$ will converge to the same value as n grows large, and we then have that $\alpha_n(S)$ goes to 0, indicating no imitation. If, on the contrary, tag applications in the system X are biased towards the suggested tags, then $prec_n(X, S)$ will be larger than $prec_n(NONE, S)$, for sufficiently large n, and then a positive $\alpha_n(S)$ will indicate imitation. In the extreme case, $prec_n(X, S)$ will tend to 1 and this will result in $\alpha_n(S)$ tending to 1, indicating full imitation.

We now justify the definition of the imitation rate formally. Assume that tags are applied according to the following probabilistic model: The probability that a user applies a tag t, if S is displayed, is a mixture of two distributions:

$$Pr(t|S) = \alpha_S \cdot f_S(t) + (1 - \alpha_S) \cdot g(t) \tag{2}$$

where with probability α_S, the user decides to take a tag from the suggestion set. $f_S(t)$ is the probability that the user chooses the tag t from the suggestion set. Consequently, $f_S(t) = 0$ for all tags $t \notin S$. With probability $1 - \alpha_S$, the user chooses an arbitrary tag, which may or may not be in S. $g(t)$ is the probability distribution over tags for this choice. Following [14], we assume $f_S(t) = g(t)/g(S)$, for $t \in S$, i.e. in the cases when the sampling is from the distribution f_S, the user preference over tags is proportional to g but confined to the set S.

We want to estimate the parameter α_S. It indicates the “persuasive power” of the suggestion set S, i.e. the likelihood that the user decides to make use of the suggestions. We call it the *imitation rate* for the suggestion set S. We next identify an unbiased estimator of α_S, which follows from the results of Boes [1]. Let T be the set of possible tags. Let $h_n(\Delta)$ be the portion out of n tag applications that fall in the set of tags Δ by sampling from the mixture distribution Eq. (2). From a corollary of Theorem 1 in [1], we have that the following is an unbiased, minimum-variance estimator of α_S:

$$\alpha_n(S) = \frac{g(T \setminus \Delta_1) - h_n(T \setminus \Delta_1)}{g(T \setminus \Delta_1) - f_S(T \setminus \Delta_1)}$$

under $g(T \setminus \Delta_1) > f_S(T \setminus \Delta_1)$, where Δ_1 is a subset of T that has to satisfy some factorization criteria given in Theorem 1 [1]. These factorization conditions hold if we can find two sets Δ_i, $i = 1, 2$ such that

$$\frac{f(t)}{g(t)} \text{ is constant for } t \in \Delta_i,\ i = 1, 2.$$

Now, note that this holds for the sets $\Delta_1 := S$, and $\Delta_2 := T \setminus S$. Indeed, $f_S(t)/g(t) = 1/g(S)$, for $t \in \Delta_1$ and $f_S(t)/g(t) = 0$, for $t \in \Delta_2$. In this case, we have

$$\alpha_n(S) = \frac{g(T \setminus S) - h_n(T \setminus S)}{g(T \setminus S) - f_S(T \setminus S)}.$$

This can be further simplified by noting that $p(S) = 1 - p(T \setminus S)$ for any probability distribution p on T and that $f_S(S) = 1$:

$$\alpha_n(S) = \frac{h_n(S) - g(S)}{1 - g(S)}. \tag{3}$$

It remains only to note that $prec_n(X, S)$ is an estimator of $h_n(S)$ and $prec_n(NONE, S)$ is an estimator of $g(t)$. We can thus estimate the imitation rate for S as given by Eq. (1).

4.2.3 Average Imitation Rate

The previous section established the imitation rate for a single suggestion set. Now, we want to capture the average imitation rate under a suggestion method X. To this end, it is natural to consider the imitation rate for all sets S that X suggests and to compute their weighted average:

$$\alpha_n = \sum_{S} \pi_n(S) \alpha_n(S) \tag{4}$$

Here, $\alpha_n(\cdot)$ is given by Eq. (3) and π_n is a probability distribution over suggestion sets. For example, $\pi_n(S)$ can be the proportion of tagging events that happened when S was displayed. Alternatively, $\pi_n(S)$ can be the proportion of tag applications that happened when S was displayed. This will give us a “per application” viewpoint:

$$\pi_n(S) = \frac{\sum_{i=1}^{n} |T(X, i)| \cdot [S(X, i) = S]}{\sum_{i=1}^{n} |T(X, i)|}. \tag{5}$$

Using Eq. (5) in Eq. (4), we can interpret α_n in Eq. (4) as the portion of tag applications that were a result of imitation under the method X.

5. RESULTS OF THE USER STUDY

We now present the results of our user study. Unless otherwise indicated, our results for a given tagged Web page under a given setting are based on 100 tagging events by different users for that page under that setting.

5.1 The Influence of Suggestions

In this section, we will determine the influence of the suggestions. We will first use the matching rate to validate that, indeed, a user's tag applications are influenced by the suggestions. Then, we turn to estimating the proportion of applied tags that were induced by the suggestions by help of the imitation rate.

5.1.1 *Matching Rate*

We computed the matching rate on 4 different Web pages for the methods TOP, FMTS and MTS. Fig. 4 shows the imitation rates averaged over the pages. As expected, TOP shows a slightly higher matching rate, indicating that users were more likely to chose tags from the suggestion set. To prove that users tag differently when the system provides suggestions, we computed the control matching rate. We observe that the matching rates are roughly twice as large if suggestions are shown. This shows that the tag generation in our user study was indeed influenced strongly by the suggestions.

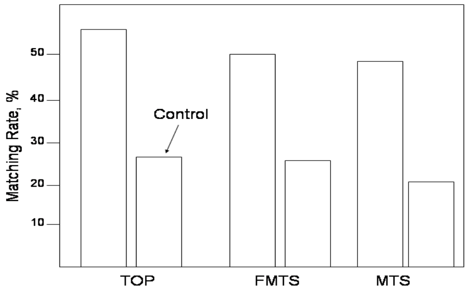

Figure 4: Matching Rates and Control Matching Rates for different methods, averaged over 4 pages

5.1.2 *The Imitation Rate*

Table 5 shows the imitation rates under different suggestion methods for 4 pages. The imitation rates are consistently between 30% and 40%. This tells us that roughly 1 out of 3 applied tags does not mirror the original user preferences, but was induced purely by the presence of the suggestions.

Table 5: Imitation Rates per Method

Method	Page 1	Page 2	Page 3	Page 4
TOP	0.3047	0.3888	0.4372	0.3760
FMTS	0.3305	0.3569	0.3488	0.3210
MTS	0.3545	0.3593	0.3603	0.3759

We further wanted to test whether suggestions that are more popular cause more users to imitate. In other words, the claim is that more plausible suggestions have a higher persuasive effect. Of course, more plausible tags are more likely to be chosen anyway, independent of the suggestions. But the imitation rate allows us to assess whether the very presence of the suggestions further boosts the plausible tags, beyond their actual true popularity. To assess this claim, we conducted a set of experiments for 10 pages with the suggestion sets fixed to either the top popular tags from del.icio.us (High), somewhat less popular tags (Middle) and even less popular tags (Low).

Fig. 5 shows the imitation rates for the three suggestion types. The plot shows that popular tag suggestions influence the user more than unpopular tag suggestions.

In summary, the results suggest that tag generation can be significantly influenced by suggestions and that this can be even further exacerbated by suggesting popular tags. This raises the issue of whether the tag applications observed in real systems that use TOP-like suggestion methods reflect the users' true preference or are an artifact of the suggestion mechanism.

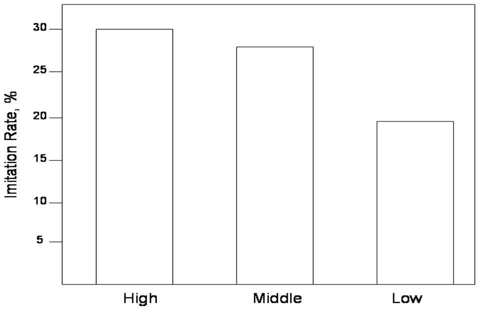

Figure 5: Imitation Rates per Tag Quality

5.1.3 *The Popularity Bias*

To further investigate how distorting the effect of a TOP-like suggestion method can be, we compared the tag frequencies under TOP with the final tag frequencies under the method NONE. We understand the final tag frequencies of NONE as the users' true preference distribution over tags. Fig. 6 shows how the aggregated tag frequencies after each event compare to the "true" tag frequencies. We show the results of the NDCG metric for two exemplary pages (the cosine similarity yields qualitatively equivalent results). For the right page, TOP deviates drastically from the true frequencies. This proves that there exist cases in which, after some number of events, the method TOP distorts the tag frequencies significantly. This is worrying, as TOP-like methods seem to be the most common suggestion schemes.

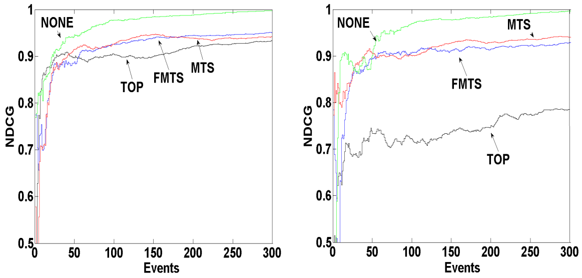

Figure 6: Comparison of the tag frequency vectors for two sample pages

5.1.4 *Usefulness of Suggestions*

After each tagging event, participants were asked to provide a feedback about the tag suggestions. We presented different predefined options (listed in Table 6) and also a freeform field "other", which was rarely used. Recall that the suggestion methods in our study tend to suggest quite different types of tags (TOP suggests the top k popular tags, FMTS tends to suggest a larger set of popular tags and MTS an even larger set; in fact, any tag would have a chance to appear in the suggestion set). Hence, one might expect that

Table 6: User Feedback on Tag Suggestions

Suggestion method	In general I pay no attention to suggested tags	I haven't noticed the suggested tags	They were confusing	They were OK, but not very relevant	They were generally helpful
TOP	63.81 ± 1.34	14.29 ± 0.98	6.37 ± 0.68	5.80 ± 0.65	**9.73 ± 0.83**
FMTS	66.52 ± 3.72	9.57 ± 2.32	7.29 ± 2.05	6.61 ± 1.96	10.02 ± 2.37
MTS	64.16 ± 3.67	9.66 ± 2.28	6.87 ± 1.93	6.01 ± 1.82	**13.31 ± 2.60**

users would prefer TOP-like suggestion methods over those that suggest less popular tags. Our results do not support this hypothesis. About 65% of users claimed to pay no attention to the suggested tags at all. The remaining feedback data suggests that the users had no particular preference for one suggestion method over another. In particular, users did not favor the method TOP over methods that suggest less popular tags. In fact, the highlighted results in Table 6 indicate that with statistical significance (at $\alpha = 5\%$) the portion of users who opted for "They were generally helpful" was larger for MTS than for TOP.

These are interesting results as they suggest that the design objective to suggest a few most popular tags (which seems to be standard practice) may not be necessarily the best design choice from the perspective of the users' judgement.

5.2 Other Factors

We conducted a number of experiments to evaluate a range of other factors that can influence the tag generation.

5.2.1 Position Bias

We wanted to understand whether users bias to applying tags from particular positions in the list of suggested tags. In our experiments, the tags in the suggestion list were presented in random order. In absence of a position bias, we should thus observe that the frequency of any tag application does not depend on the position of this tag in the list. It is of interest to understand whether the position bias exists as this can affect designs that do not randomize the order of tags (but, e.g., apply an alphabetical or popularity order instead).

The position bias is well known to feature user interaction with search engine results sets where the few top items in the list get most of the clicks. Our setting differs in that the order of tags in the list is random, the length of the list is typically small, and tags in the list are arranged in a row, not in a column.

Fig. 7 shows that users bias to selecting the leftmost tag in the list. However, this is not very pronounced (< 15% relative to the second tag from the left in the list). The bias over tags at other positions is statistically insignificant.

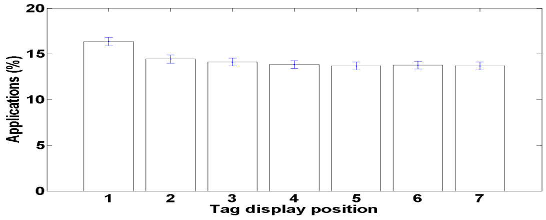

Figure 7: Tag Applications per Tag Position

In summary, we find that there is a weak position bias, which might make it reasonable to randomize the order of the suggested tags to avoid undesired tag popularity skews.

5.2.2 Suggestion Set Size

In our base experiments we fixed the suggestion set size to seven, motivated by existing systems (e.g. del.icio.us). In practice, this choice may have been motivated by earlier experiments on the human capabilities of information processing [8]. We wanted to evaluate how the tag generation process is influenced by the number of suggested tags. To this end, we considered 10 Web pages. For each of them, we created 5 experimental settings in which the suggested tags were fixed to the 0, 3, 5, 7, and 20 top popular tags (according to del.icio.us).

We measured the imitation rate for each setting. The results, averaged over the 10 pages, are shown in Fig. 8 (top). One would expect more imitation for larger suggestion sets. However, the estimated imitation rates are *non monotonic* with respect to the suggestion set size. This indicates a non-trivial relationship between the persuasive power of a suggestion set and its size.

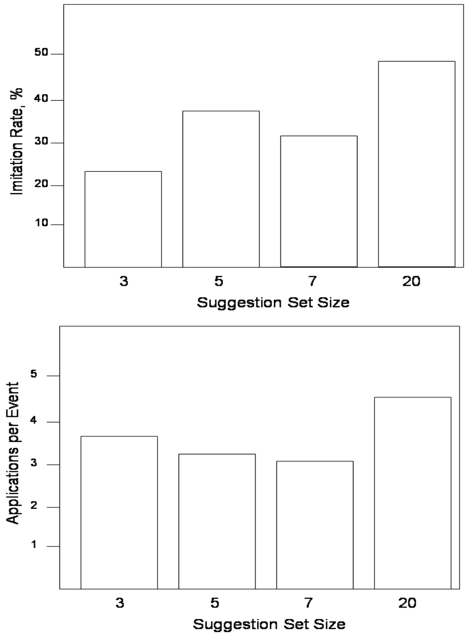

Figure 8: Effects of the Suggestion Set Size

We also analyzed the number of applied tags per event. We found that the average number of applications per event does not appear to be monotonic with respect to the suggestion set size (Fig. 8 (bottom)). In particular, we find that with statistical significance (at $\alpha = 5\%$) there exists a suggestion set size for which the average number of applications per event is larger than for any other smaller suggestion set size.

These are interesting findings as they suggest that the size of the suggestion set may influence the users' imitation rate significantly in non-trivial, non-monotonic ways.

5.2.3 Click or no Click

We wanted to evaluate whether users would bias to applying tags from the suggestion set if they can be selected by a *click*. To this end, we considered 10 Web pages and fixed the suggestion sets to the 7 most popular tags (according to del.icio.us). For each such setting, we ran two experiments: (A) with clickable suggestions and (B) with non-clickable suggestions. We found that in case (A) more than 45% of applied suggested tags were selected by clicking on a suggested tag, confirming that users were making good use of this feature. At the same time, we found that in both cases the imitation rate was the same.

This suggest that making tags clickable will not bias the tag applications, but benefits those users who prefer clicking over typing.

6. CONCLUSION

We found that user-generated tags feature substantial semantic noise, more than the terms from either page content or search queries. Yet, our analysis reveals that meaningful tags emerge in the more popular tags of a document and that this meaningfulness improves with the number of users that tag the document. We also found that popular tags for a document cover better the terms of the queries than the frequent content terms. The proportion of "useful" terms (titles, categories, search keywords and descriptions) in the tags increases rapidly among the most popular tags, grows very slowly beyond the top twenty tags and overall attains a moderate recall. Overall, the results are encouraging news, suggesting that popular tags contain useful terms, which could be leveraged for advanced applications.

Our study on the influence of tag suggestions yields that the users' tendency to bias applied tags towards the suggestions could be substantial. We also found evidence that users may tend to bias even more towards the suggestions, if the suggested tags are popular. Interestingly, we found that users did not prefer a suggestion method that suggests a few popular tags over those that tend to cover a larger set, including less popular tags. We also identified and analyzed several other factors that influence the tag generation and derived consequences for the user interface design. In summary, the results raise the question whether the tag applications observed in real systems reflect users' true preference over tags or are an artifact of suggestions. The observation that popularity of the suggested tags may even further encourage users to follow suggestions (and thus provide less information about their unbiased inner preference over tags) raises the question whether the standard design choice to suggest a few top popular tags is indeed the best practice. The user feedback indicating indifference over the suggestions of varying popularity could be leveraged in the design of advanced suggestion methods.

Future work may investigate the design of suggestion methods that would best support specific users tasks such as search and navigation.

Acknowledgement

We would like to thank Misha Bilenko, Silviu-Petru Cucerzan, and Ryen White from Microsoft Research Redmond for providing us with QUERY data set used in this paper. We thank Nick Duffield, John Mulgrew, and Andy Slowey from Microsoft Research Cambridge for providing us with technical assistance in conducting our user study. Last but not least, we thank the participants of our user study.

7. REFERENCES

[1] D. C. Boes. On the estimation of mixing distributions. *The Annals of Mathematical Statistics*, 37(1):177–188, 1966.

[2] E. H. Chi, A. Kittur, and T. Mytkowicz. Augmented social cognition: Understanding social foraging and social sensemaking. In *Proc. of the HCIC 2007 Winter Workshop*, Fraser, Colorado, Jan 31–Feb 4 2007.

[3] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proc. of the ACM SIGIR '03*, 2003.

[4] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

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

[6] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. pages 41–48. ACM Press, 2000.

[7] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *ISWC*, LNCS, pages 522–536. Springer, 2005.

[8] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

[9] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi. Tracking user attention in collaborative tagging communities. In *Workshop on Contextualized Attention Metadata*. ACM Press, 2007.

[10] S. Sen, F. M. Harper, A. LaPitz, and J. Riedl. The quest for quality tags. In *Proc. of GROUP'07*, Florida, USA, November 2007.

[11] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, and D. Frankowski. tagging, communities, vocabulary, evolution. In *Proc. of the ACM CSCW*, 2006.

[12] F. M. Suchanek, G. Ifrim, and G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In *KDD*, 2006.

[13] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A Core of Semantic Knowledge. In *WWW*, New York, NY, USA, 2007. ACM Press.

[14] M. Vojnović, J. Cruise, D. Gunawardena, and P. Marbach. Ranking and suggesting tags in collaborative tagging applications. Technical Report MSR-TR-2007-06, Microsoft Research, February 2007.

[15] R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proc. of the SIGIR 2007*, pages 159–166, 2007.

[16] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proc. of the Workshop on on Collaborative Web Tagging at WWW 2006*, 2006.