Understanding del.icio.us Tag Choice Using Simulations

Rick Wash
School of Information
University of Michigan
rwash@umich.edu

Emilee Rader
School of Information
University of Michigan
ejrader@umich.edu

ABSTRACT

Understanding how users choose tags can help researchers better understand how tagging systems can be used and how to design better tagging systems for the future. We developed a simulation of del.icio.us, a popular social bookmarking tool, that allowed us to simulate users choosing tags using one of four possible strategies for tag choice found in the literature. We then compared the resulting tag choices with empirical data retrieved from del.icio.us to determine which tag choice strategies would result in choices most similar to those seen in the real world. We were able to rule out three of the strategies as unlikely to be the primary means by which tags are chosen on del.icio.us.

Topics

Information Organization

Keywords

tagging, del.icio.us, social bookmarking, simulation

1. INTRODUCTION

User-contributed metadata, also known as *tagging*, is increasingly receiving attention as a tool for digital information management. Collaborative tagging systems such as del.icio.us and citeulike.org publicly expose individual users' associations between content items and tags, thereby providing visibility into words others have used to tag similar items. Grudin [8] suggests that collaborative tagging can be a low-effort solution for shared or group information management, because it does not require that users try to conform to a controlled vocabulary or organization scheme.

In this paper, we focus on the social bookmarking website del.icio.us as a case study of a collaborative tagging system supporting both personal and shared information management. del.icio.us is an online application that allows users to save and tag their own web bookmarks so they are accessible from any networked computer. del.icio.us has recently received attention in the research literature as the canonical example of a collaborative tagging system for information management [7, 9]. We are studying how users choose tags, and hope to apply this knowledge to improve information management interfaces.

Wash and Rader [12] argue that the usefulness of del.icio.us depends critically on how users choose tags. Golder and Huberman [7] argue that users' tag choices are not random; instead, consensus seems to emerge for which tags best represent a given web page. They show that web pages bookmarked in del.icio.us demonstrate a stable frequency distribution following a power-law pattern in which the same few tags are chosen by many users, while most other tags are selected by only a few users. The "long tail" of the power-law distribution includes hundreds more tags used by only one or two people. Golder and Huberman speculate that this pattern might be due to users imitating each others' tag choices; in other words, when a user bookmarks a web page in del.icio.us, their tag choices might be influenced by tags that had been previously applied to that web page by other users.

However, it is reasonable to assume that there might be other sources of influence on users' tag choices having to do with personal information management goals. For example, a user interested in del.icio.us only for organizing and re-finding their own bookmarks might strive for consistency within their own "controlled vocabulary", to maintain a shorter list of tags [12]. Or, users might desire to expend as little effort as possible when choosing tags,

and simply select the tags the system recommends when they create a new bookmark.

The research literature contains multiple competing explanations for the patterns of tags that appear on del.icio.us. All of these explanations initially seem like reasonable explanations of the way that del.icio.us users choose tags. But they cannot all be right simultaneously. It would be nice to actually have collections of real people specifically use one or more of the strategies suggested by the literature on delicious. We could then see if those tag choices resulted in patterns of tags similar to those found on del.icio.us. Unfortunately, this technique would be prohibitively costly. Instead, we developed a computational simulation of users doing this. This simulation allows us to control exactly what strategy the simulated users used to choose tags. Such a simulation cannot tell us which strategy the real users of del.icio.us used; it can only tell us which strategies are likely to result in patterns of tags that are similar to those observed on del.icio.us. In other words, the simulation cannot confidently identify how users chose tags, but it can be used to rule out explanations that are unlikely to generate the observed patterns of tags.

We ran the simulation for four different tag selection strategies and compared the results with the tag frequency distribution observed on del.icio.us. We then used statistical techniques to determine whether the simulated tag selection strategies were consistent with the empirically observed data. We tested the following strategies:

Zipf's Law: Zipf's law states that word frequency in most written works follows a powerlaw distribution. Therefore, del.icio.us users might naturally choose their words from a powerlaw distribution.[10]

Imitation: Users might imitate previous users' tag choices. This was described by Golder and Huberman [7]

User-based: Users might favor tags that they had used previously. This was described by Wash and Rader [12]

Recommended: Users might prefer to click on the tags that are programatically recommended by del.icio.us.

The latter three are the same strategies studied in a large-scale statistical analysis by Rader and Wash [11].

The goal of the research described in this paper was to detect similarities between the tag frequency distributions produced by our four simulated strategies, and the actual frequency distribution of tags on a random sample of websites bookmarked in del.icio.us, henceforth called the *empirical* distribution. In this way, we hoped to identify which strategies were capable of producing distributions similar to the empirical distribution and which were not. Strategies producing similar distributions can be considered plausible, in the sense that they could possibly have given rise to the empirical distribution. Strategies producing distributions that are dramatically different from the empirical distribution can be ruled out as unlikely to have been used widely on del.icio.us

We began by downloading and parsing the entire bookmark and tag histories for approximately 12,000 different websites in del.icio.us. We generated the tag frequency distribution for a randomly selected subset of those websites, and compared it against seven known distribution types (powerlaw, log-normal, geometric, etc.), to determine the best fit. Knowing the family of distributions to which the del.icio.us tag frequency distribution belongs allowed us to statistically determine whether the distributions produced by our simulation were of the same type.

Next, we developed a simulation of users' tag choices, details of which will be provided later in the paper. The simulation was instructed to follow each one of the four strategies listed above, in turn. For each strategy, 30 different websites were simulated; for each website users' simulated tag selections were recorded and statistically compared with the empirical distribution.

We were primarily interested in looking at two specific patterns of tags. First, as many people have pointed out, the tags associated with a specific website in del.icio.us tend to follow a powerlaw distribution [7]. Tag choice strategies that do not result in a powerlaw-shaped distribution of tags are unlikely to have been used by the users of del.icio.us. This powerlaw distribution seems to be an important property of tags on del.icio.us [7]. Second, we have looked at the average inter-user agreement between users who bookmark the same website. Furnas et al. [6] have found that typically two individuals will choose the same word to describe an object less than 20% of the time. On del.icio.us, this number is closer to 15% of the time. Tag choice strategies that result in many users agreeing with each other are also unlikely to have been used on del.icio.us. Inter-user agreement is a metric that is sufficiently different than the powerlaw distribution of tags, and is a good complementary metric for characterizing a set of tag choices.

In this paper, we first report the results from our analysis of the empirical data downloaded from del.icio.us. The

results from this analysis were used in the simulation to make the simulation more realistic. We then describe the simulation in detail, and report the results of the analysis comparing the empirical and simulated data. We conclude with a discussion of the implications of this work.

2. ANALYSIS OF EMPIRICAL DATA

Over two weeks in January 2007, we downloaded the entire bookmark and tag history for approximately 20,000 different webpages in del.icio.us. The webpages were chosen by periodically sampling the "recently posted" and "popular" del.icio.us pages. We randomly chose 30 webpages from our sample that had been bookmarked by at least 100 users. Then, in June 2007 we downloaded the complete public bookmark histories for all of the approximately 12,000 users who had ever bookmarked any of these 30 webpages. In other words, our dataset contains the complete tag histories for 30 webpages bookmarked in del.icio.us, as well as tag histories for all users who ever bookmarked any of those 30 webpages.

We used this data to estimate two distributions: 1) The distribution of tag frequency — for a given website bookmarked, how frequently was each tag applied to it? and 2) The number of tags chosen by a user — when a user is bookmarking a website, how many tags will he or she apply?

We fit the data to multiple families of distributions and see which distribution fits "best." "Best" here is a statistical determination [4]. We fit the data from each site to seven different discrete probability distribution families (discrete powerlaw, negative binomial, binomial, discrete lognormal, discrete exponential, poisson, and geometric), estimating parameters with maximum likelihood estimation. We then used a non-nested Kolmogorov-Smirnov test to conduct pairwise comparisons between these distributions.

The discrete powerlaw distribution fit the empirically observed tag distributions better than any of the other 6 distributions we tested. The fitted distribution had an average exponent (α) of 1.92 ± 0.40. This is a rather low exponent for a powerlaw distribution, and indicates that the "long tail" of tags is very long and heavy. This low exponent also has another important implication. Newman [10] explains that powerlaw distributions with an exponent less than 2 have an infinite (or undefined) mean. Therefore, estimates of a "mean" or average tag are undefined, and any inferential statistics based on the mean of the tag distribution cannot be used.

The number of tags chosen by a user fit a discrete log normal distribution better than the other 6 distributions we tested. The fitted distributions had a mean log value of 0.82 ± 0.45 and a standard deviation log value of 0.73 ± 0.19. On average a user of del.icio.us will choose 2.51 tags (with a standard deviation of 1.42) when bookmarking a website.

Inter-user Agreement It has long been accepted that people use language imprecisely, and meaning is negotiated on the fly during conversation [3]. This imprecision is evident not only in communication, but also when people are asked to create keywords for recipes and names for common editing operations [6] and when user-generated index terms are compared with Library of Congress subject headings [5]. In fact, the probability that two people will generate the same label for the same object, called the "vocabulary problem," is widely held to be less than 20% [6, 2]. When a user wants to take advantage of the collective properties of social bookmarking by browsing or searching on tags, the vocabulary problem becomes apparent. If users are unlikely to choose the same tags to represent the same topics, such diversity would decrease search precision. When a given tag is applied to bookmarks in an inconsistent manner by many users, more variability exists in the content returned when a user searches with that tag. The desired bookmark may be returned, but there would be too much other "noise" in the results for it to be noticed.

To measure the extent of the vocabulary problem on del.icio.us, we calculated the average inter-user agreement for a sample of 200 users from each of the 30 websites that we had full user data for. On average, users who bookmarked these websites chose the same tag for the website only $14\% \pm 5\%$ of the time. This percentage is low, indicating a fair bit of disagreement between users, though it is higher than the 8% reported by [6] for their text-editing operations dataset. Figure 1 shows a histogram of the inter-user agreement values for the 30 websites.

3. SIMULATING TAG CHOICES

To compare the four tag choice strategies, we simulated 120 websites for each of the four strategies. Each of the 120 websites was paired with one of 30 real websites randomly selected from our sample downloaded from del.icio.us, and the number of users for each simulated website was chosen to match the real website.

Each simulated website was assigned to have the same number of users as its matched real website, and each simulated user was matched with a real user who bookmarked that website. In essence, we are simulating what would happen if the same set of users bookmarked the real

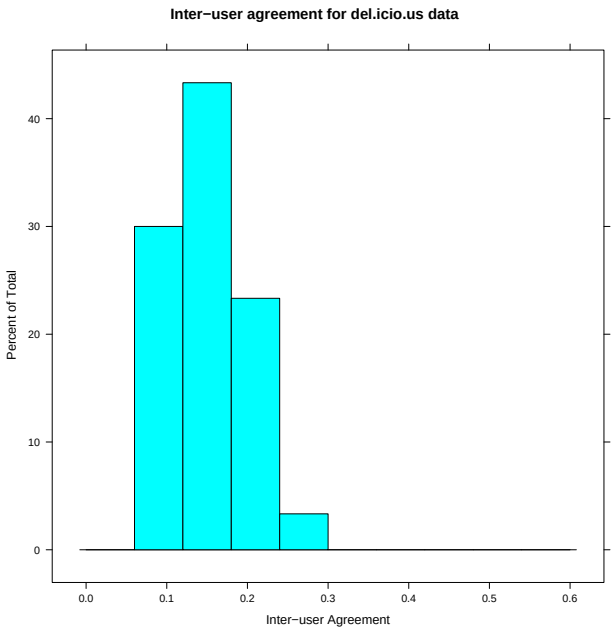

Figure 1: Histogram of the Average Inter-user Agreement for 30 websites bookmarked on del.icio.us

website, but chose their tags according to one of our four hypothesized strategies (and bookmarked it in a random order). To simulate a user choosing tags for that website, two choices have to be made. First, the simulator chooses how many tags that user will apply to the website. Second, the simulator chooses which specific tags would be applied.

As we found above for del.icio.us, the number of tags applied by a user tends to be distributed according to a discrete log normal distribution. For each simulated website, the simulation randomly chose a set of parameters (mean log, and standard deviation of the log) for a discrete log normal probability distribution to match the distribution of parameter values we found empirically on del.icio.us. Once a set of parameters was chosen, this specifies an exact probability distribution. For each simulated user, the simulator would then choose a random number from this probability distribution, and that number would be the number of tags that simulated user would choose.

We found that that tags applied to a given website on del.icio.us are distributed according to a discrete powerlaw distribution. In a manner similar to that used for the number of tags, the simulator chooses a parameter (alpha) for a discrete powerlaw distribution. This distribution then will serve as the base distribution that specific tags are drawn from. The alpha parameter is chosen not to directly match the value observed above (1.92), but to average 2.5. This was done because any single user's tags cannot repeat, and this lack of repetition (i.e., sampling without replacement) forces the user to choose more tags in the 'tail' of the distribution than they would if tag choices were truly independent. If I choose three tags, they cannot all be the most popular tag. Through experimentation, we found that having users choose tags from a powerlaw distribution with an alpha about 2.5 tends to yield a site-wide powerlaw distribution for tags that approximately match the observed distributions.

Each number from this distribution is then mapped onto a specific tag, according to its rank in the frequency distribution. The tags from the matched real site are ordered from most-frequently used to least-frequently used, with ties being broken randomly. A 1 from the random number generator is then mapped onto the most-frequently-used tag, 2 onto the second most frequently used tag, and so on. Any numbers larger than the number of tags on the matched site are left as numbers.

For each user, the specific tags that they choose will depend on which of the four strategies they are assigned to use. The only difference between these four strategies is in specific tag choice; all other decisions (number of users, number of tags per user, etc.) are identical.

Zipf's Law The simplest strategy is to follow Zipf's law, and choose tags directly from the base powerlaw distribution. The simulator continually chooses random numbers from the base powerlaw distribution until it has the required number of unique numbers. These numbers are then mapped onto tags as described above.

Imitation For users to imitate previous users' tag choices, it is necessary for those previous users to exist. The first few users who bookmark a website will have no one to imitate. To handle this, the first 20 users will draw as described above for Zipf's law and serve as 'seeders.' All users after the first 20 who use this strategy will choose a tag from the current empirical distribution of tags for his simulated website. This means that if there are two tags, 'A' and 'B', and 'A' has been used twice previously and 'B' only once, then tag 'A' is chosen with probability $\frac{2}{3}$ and tag 'B' is chosen with probability $\frac{1}{3}$. However, to ensure growth of the vocabulary beyond that used by the initial 20 seeders, each tag choice has a 10% probability of choosing a new, previously unused tag. This probability was chosen to match the average empirically observed probability from the del.icio.us data. The average website in our sample from del.icio.us has a new tag probability of $10.5\% \pm 8.3\%$.

User-based Not all users will want to apply tags that they have used before. As such, when simulating the **user-based** strategy, the simulated users had a 50/50 chance of choosing tags according to Zipf's law, and of choosing tags they had used before. When choosing tags they had used before, the simulator computes the overlap (set intersection) between the tags the user had ever used and the tags that were ever applied to the matched site. It then randomly chooses among the tags in this overlap set. If that is not enough tags, then additional tags are chosen randomly from the base powerlaw distribution.

Recommended Del.icio.us does not make their algorithm for choosing which tags to recommend public. As such, we could not directly simulate users choosing from del.icio.us's recommended tags. We did, however, create a simple approximation. We proposed that the tagging system could simply recommend the N most popular tags for that website. Then users could randomly choose among those N tags.

If del.icio.us's recommendation algorithm is solely based on the frequency of applying tags to that website, then this approximation is a reasonable proxy for del.icio.us's recommendation algorithm. If the real algorithm includes other data (such as the tags a user has used, or the popularity of tags across the site) then this simulated strategy will not be a good proxy for the del.icio.us algorithm.

To simulate users choosing tags with the **recommended** strategy, we first create 20 'seeders' in the same way we did for the **imitation** strategy simulation. All of the remaining users then are simulated to have been presented with $N = 5$ 'recommended' tags (the 5 most popular tags at that point) and then randomly choose between these recommended tags. If they need to apply more than 5 tags, then the remaining tags are chosen randomly from the base powerlaw distribution.

4. SIMULATION RESULTS

One of the benefits of simulation is that the development process forced us to be very explicit about what information users would need to follow a hypothesized strategy. Golder and Huberman [7] suggest that the powerlaw distribution of tags for a given website could arise from users intentionally imitating previous users tag choices. They cite the networks literature [1, 10] for evidence that powerlaw distributions can arise from such path-dependent choices. When trying to replicate these decisions for our simulation, we found that this only works if a user chooses tags from the empirical distribution *at the time of decision*. This means if tag 'A' has been used twice as much as tag 'B', then tag 'A' need to be chosen with twice the probability of tag 'B'. This is a very high information requirement for users – they must know the exact proportions of existing tags to choose appropriately. Del.icio.us does not make this information easily available; however we assume this knowledge for our simulations with the **imitate** strategy.

Table 1: 1) Average difference between simulated powerlaw exponent and real powerlaw exponent, and 2) Average Kolmogorov-Smirnov goodness-of-fit statistic

	Delta	*KS*
Real World	**X**	0.07
Zipf's Law	0.23	0.08
User	0.19	0.08
Imitation	0.32	0.14
Recommended	0.34	0.22

To compare the four tag choice strategies, we simulated 120 "websites" for each of the four strategies. Each of the 120 websites was paired with one of 30 real websites, and the number of users was chosen to match the real website. Figure 2 shows the tag distribution (on a log-log plot) for all four of the strategies on one of these websites. Also on the graph is the empirically observed tag distribution from the paired website. The non-powerlaw nature of the **recommended** strategy stands out, with a small number of roughly equally likely tags (the recommended tags) and then a sharp drop in probability for the other tags.

For each simulated website, we fit the simulated tag distribution to a discrete powerlaw distribution. We used this fit for two comparisons. First we computed the difference between the powerlaw exponent in the simulated distribution and the exponent in the real distribution. The first column (Delta) of Table 1 shows the average difference in exponents. Second, we conducted a Kolmogorov-Smirnov goodness-of-fit test to see how well the simulated distribution fit a powerlaw. A KS statistic of 0 means that the distribution is identical to a powerlaw, and higher numbers indicate greater deviation from a powerlaw. The second column (KS) of Table 1 indicates the average KS statistic for each of the four strategies. As a comparison, the empirical data from del.icio.us fits a powerlaw with an average KS statistic of 0.067. Neither the **imitation** nor **recommended** strategies fit a powerlaw very well. The distributions from **Zipf's Law** and **User-based** strategies fit as well as the real data from del.icio.us.

Inter-user Agreement In addition to fitting the tag distribution, we also calculated the average inter-user agreement between users of each of the 480 simulated websites. Figure 3 shows the distribution of average inter-

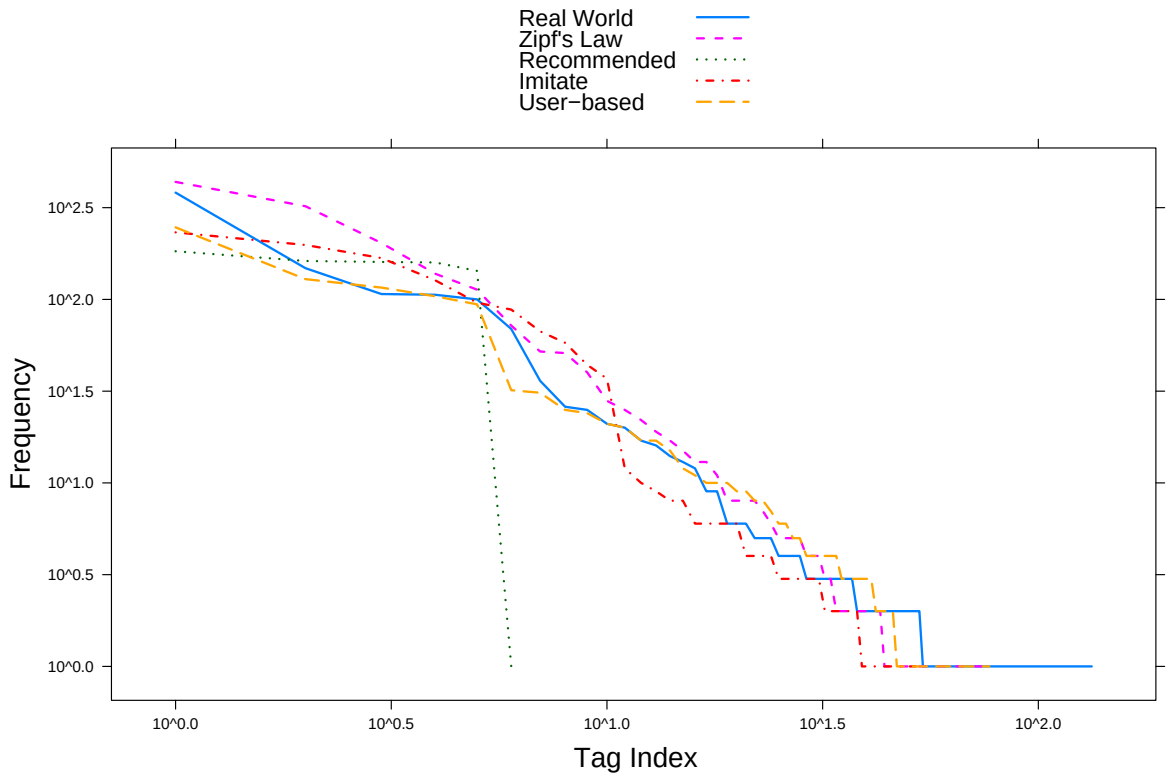

Figure 2: Tag distributions for four simulated runs, matched with a real empirically-observed tag distribution (on a log-log scale)

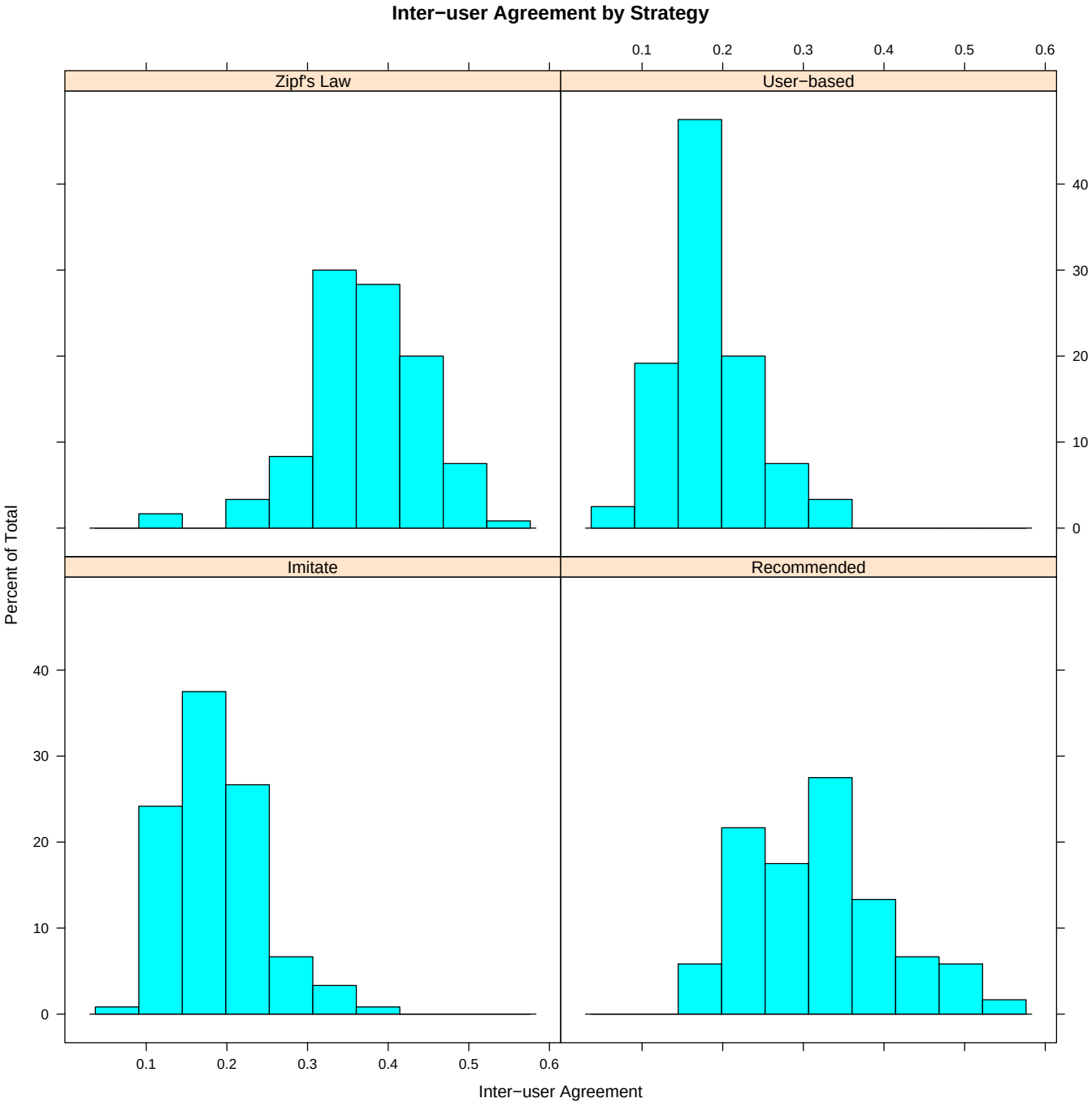

Figure 3: Histograms of the average inter-user agreement for the simulated websites. There is one histogram for each tagging strategy used. Notice that two strategies (**Recommended** and **Zipf's Law**) have a higher mean than the other two.

Table 2: Average Inter-user Agreement for each of the four simulated tagging strategies

	Mean IUA	σ^2
Real World	0.144	0.051
Zipf's Law	0.373	0.074
User	0.182	0.052
Imitation	0.184	0.056
Recommended	0.317	0.088

user agreements for each of the four strategies. Table 2 provides the mean inter-user agreement value for each of the four strategies (along with their standard deviation), and a comparison point with the real-world data from del.icio.us. The mean inter-user agreement for the **user-based** strategy and the **imitation** strategy are statistically indistinguishable. All other pairwise comparisons between inter-user agreement values are statistically significant at the 10% level.[1]

Our simulations resulted in users agreeing with each other much more often when using the **Zipf's law** strategy or the **recommended** strategy. This indicates to us that these strategies are unlikely to have been used by the users who bookmarked these 30 websites on del.icio.us. However, the inter-user agreement tests cannot rule out either the **user-based** strategy or the **imitation** strategy for tag choices, because they resulted in similar levels of inter-user agreement.

5. DISCUSSION

We compared four possible strategies that users might use to choose tags on del.icio.us. Our simulations indicate that choosing the recommended tags would result in a skewed distribution that fits a powerlaw distribution less than the empirically observed distribution. Choosing the recommended tags will also cause higher levels of inter-user agreement than are empirically observed. Both of these findings are a direct result of the fact that in our simulation users choose uniformly at random among the recommended tags. They do this because they have no better way of determining which recommended tags are appropriate. On del.icio.us, it is unclear how users choose among the tags that del.icio.us recommends. If they just 'click a couple' of tags, then del.icio.us would end up with skewed tag distributions and high inter-user agreement. Since we did not find that on del.icio.us, if users do utilize the recommended tags our simulations suggest that they do so more intentionally than 'just clicking a few' of the recommended tags.

The hypothesis that users choose tags by imitating other users, with previous tag choices influencing the current choice, is also unlikely to be what del.icio.us users are really doing. The information requirements that would be necessary for users to choose tags this way are large; users need to know both which words were previously applied as tags and how frequently they were applied to use this hypothesized strategy. But even if users could easily get the proper information, the resulting tag distributions from the simulations are skewed and have a shorter tail than the tag distributions we observed on del.icio.us. This suggests that users are driven by something more than just imitating past users.

We also believe that the Zipf's law hypothesis is unlikely, as it results in dramatically higher inter-user agreement than we observed on del.icio.us. This high inter-user agreement is because every user who bookmarks a website with this strategy is choosing tags from the same power-law distribution — the tag that is most likely to be chosen is the same for all users. This suggests that users are not all choosing tags from the same distribution – at the very least they have individualized distributions of words to choose from.

We cannot rule out the user-based hypothesis based on our simulations. We found that when our simulated users choose tags using the user-based strategy, the resulting tag distribution is as close to a powerlaw distribution as our empirical data from del.icio.us, and the inter-user agreement is approximately similar to the level of inter-user agreement in our sample of del.icio.us. Choosing tags from the user's set of tags results in a similar level of inter-user agreement and a powerlaw tag distribution. This suggests that each user has his or her own way of choosing tags, and that individual and idiosyncratic way of choosing tags is a major influence on tag choices. However, this research is not able to distinguish what some of the different idiosyncratic strategies are. It is clear, though, that future systems that support collaborative tagging will need to be sufficiently flexible to allow each user to choose their own way of determining which tags to use.

Since this is a simulation study, we can only compare aggregate measures of tag choices like the shape of the resulting tag distribution. This study cannot look at individual tag decisions to determine which strategy was used for that specific tag choice, as the same strategy might not be used across all users, or even all the tag choices of an

[1]Comparisons were done with a series of t.tests, and used a Bonferroni correction to adjust the p-values for multiple tests. All remaining comparisions were statistically significant at the 0.1% level except the two: the comparison between the real world data and the **user-based** strategy, and the comparison between the real world data and the **imitation** strategy.

individual user. In another part of this project, we [11] use a logistic regression to attempt to determine which of these four strategies was in use for individual tagging decisions on del.icio.us. Fortunately, the results from that work corroborate the results here.

REFERENCES

[1] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[2] M. J. Bates. Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13):1185–1205, 1998. BatesJASIS1998.

[3] H. Clark. *Using Language*. Cambridge University Press, Cambridge, England, UK, 1996.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. Preprint, Jun 2007.

[5] L. Y. Collantes. Degree of agreement in naming objects and concepts for information retrieval. *Journal of the American Society for Information Science*, 46(2):116–132, 1995.

[6] G. Furnas, T. Landauer, L. Gomez, and S. Dumais. Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806, 1983.

[7] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[8] J. Grudin. Enterprise knowledge management and emerging technologies. In *HICSS '06*, 4-7 January 2006.

[9] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07*, 2007.

[10] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[11] E. Rader and R. Wash. Collaborative tagging and information management: Influences on tag choices in del.icio.us. Working Paper, University of Michigan., September 2007.

[12] R. Wash and E. Rader. Public bookmarks and private benefits: An analysis of incentives in social computing. In *ASIS&T Annual Meeting*, 2007.