11-4-2006

Exploring Characteristics of Social Classification

Xia Lin
Drexel University

Joan E. Beaudoin
Drexel University, joan.beaudoin@wayne.edu

Yen Bui
Drexel University

Xaushai Desai
Drexel University

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

EXPLORING CHARACTERISTICS OF SOCIAL CLASSIFICATION

Xia Lin <xia.lin@drexel.edu>
Joan E. Beaudoin <joan.elizabeth.beaudoin@drexel.edu>
Yen Bui <yen.bui@drexel.edu>
Kaushal Desai <kaushal.desai@drexel.edu
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104, USA

Abstract

Three empirical studies on characteristics of social classification are reported in this paper. The first study compared social tags with controlled vocabularies and title-based automatic indexing and found little overlaps among the three indexing methods. The second study investigated how well tags could be categorized to improve effectiveness of searching and browsing. The third study explored factors and radios that had the most significant impact on tag convergence. Finding of the three studies will help to identify characteristics of those tagging terms that are content-rich and that can be used to increase effectiveness of tagging, searching and browsing.

1. Introduction

In a now classic book, *Woman, Fire, and Dangerous Things: What Categories Reveal about the Mind,* George Lakoff (1987) presents a convincing path of how our understanding of human categorization and classification has developed in the past several decades. Traditionally, categorization or classification has been viewed as defining some specified bounds that can serve as abstract containers to group items with certain shared properties. This view has been challenged and expanded by new empirical-based theories in cognitive science. These theories provide a much broader and dynamic view of categories or classification. For example, boundaries of categories might be fuzzy; categorization might be done at different cognitive levels; and some members of the categories might be more "typical" than others in representing the categories, etc. These kinds of new understanding are changing not only our views of categories, but also our views of how knowledge is organized in our head and in the physical world.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Similarly, the new emerging concept of social classification has begun to challenge traditional classification schemes and controlled vocabularies-based indexing. The new approach, sometimes called social tagging or folksonomy, departs significantly from traditional knowledge organizing approaches. For example, social classification is "social," i.e., it depends on users' input or users' consensus (Hammond et al., 2005). It relies on users' freely-chosen tags as "a fundamental organizational construct." (Mathes, 2004). There is no hierarchy, no directly specified term relationships, and no pre-determined set of groups or labels used to organize user's tags. As a result, the tagging space becomes a self-defined and self-regulated network similar to those networks studied by the newly emerging science of networks (Watts, 2003).

All these suggest that the key to understanding how tagging works is to comprehend or discover hidden properties or characteristics of the tagging space. Only recently do we see some research in this direction. Golder & Huberman (2006) studied the dynamic structure of tagging and found regularities in user activity, tag frequencies, kinds of tags used, bursts of popularity and stability in tagging. Joseph et al. (2006) examined the possibility that a community vocabulary can emerge through self-organization. Voss (2006) explored the similarities and differences between Wikipedia, folksonomies, and traditional hierarchical classification. He suggests Wikipedia's category system is a thesaurus that uses a special combination of social tagging and hierarchical subject indexing.

Departing from early papers on social tagging that are mostly based on observations and discussions, these three research projects all provide concrete empirical data to reveal insights about the tagging space. It is important to continue this line of empirical work in order to develop our understanding of how social classification works and how it is different from other traditional approaches. In this paper, we report on the investigations and findings of three case studies on the characteristics of social classification.

2. Research Questions

The primary goal of this research is to provide empirical evidence to describe emerging characteristics of social classification. As a new phenomenon social classification creates a divergent and dynamic tagging space for which a complete picture would be very difficult, if not impossible, to obtain. Thus, the best way to look into this space is through snapshots taken from

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

several different points of view. In this research, three different experiments were designed to obtain snapshots of social classification in specific angles. Each experiment was designed to address a specific research question. Each is a limited case study on one or several current tagging systems. Each provides a partial view of the social classification space bound by the specific system used and the specific time period under investigation. Together, they provide some answers to the central question we are exploring: how is digital information organization different from traditional information organization?

The three research questions explored are:

1. How different, in terms of content representation, are social classification and traditional content indexing?
2. How well can tagging be categorized in any meaningful way – i.e. do tags fall into predictable, conceptual groups?
3. Do tags converge, namely do certain tags become much more popular than others? If so, to what degree do they converge and what factors likely contribute to the convergence?

All the three questions started from a basic question of social classification: what are the characteristics of those tagging terms that are content-rich and that can be used to increase effectiveness of searching and browsing. Clearly not all tagging terms will fall into this category. It is a question of how to tell apart those “more useful tags” from those “less useful tags.” We take different approaches to address this question through the three research questions. For research question 1, we attempt to do a direct comparison between social tagging and traditional indexing methods to get an estimate of overlaps among different indexing methods. For research question 2, we explore whether tagging terms can be categorized into groups to make tagging, searching and browsing more effective. For research question 3, we investigate whether we can use convergence to select those popular tags so that, potentially, we can rely on those popular tags for content indexing.

3. Three Case Studies on Social Classification

In this section, we present the data, methodologies and results of three case studies on social tagging.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

3.1 The study on a comparison of indexing methods

In this study, we compared the overlaps among social tagging, thesaurus-based indexing, and title-based keyword indexing. The comparison is interesting for multiple reasons. First, the differences between various indexing approaches have often been discussed but rarely evaluated on empirical data from collaborative tagging platforms. An estimation of the similarity between classification methods is necessary to evaluate whether social classification is significantly different from traditional indexing results. Once the differences are quantified, it may be interesting to explore the factors that contribute to these differences. Second, the differences between traditional indexing and tagging have often been applied with the potential of achieving a qualitative improvement in IR systems and development of recommender systems (Xu & Zhang, 2005, 2006). Finally, the long-term goal of such studies would be to derive a hybrid or composite indexing schema using the strengths of each approach.

3.1.1 Data and Methodology

Key for this study is the identification of those documents that are indexed by controlled vocabularies and that have been tagged by enough users. Connotea was chosen as our platform for data selection in this study. Connotea (http://www.connotea.org/) is a free online reference management service as well as a social bookmarking tool. The tool, owned by the Nature Publishing Group, allows users to annotate research articles with single-word tags as well as phrase tags. Since Connotea also automatically populates research articles with bibliographic information from known scientific databases like PubMed, it is possible to access Medical Subject Heading terms (MeSH terms) associated with articles bookmarked in Connotea.

Data Selection. Data used in the experiment are selected based on the following steps:

- Search for publications that have a PubMed identifier associated with them.
- Select publications for which the corresponding MEDLINE records include MeSH terms.
- Select publications that are tagged by the most number of users.
- Select publications that are associated with the most number of tags.

Title-based Automated Indexing. In addition to tags and MeSH terms collected for each publication, a third set of terms was generated from an automated indexing process using the GATE text-processing engine (Gaizaukas et al., 1996). For each selected publication, the title

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

was processed using the GATE platform and the noun-phrase chunker (Ramshaw and Marcus BaseNP algorithm). The resulting sets of noun-phrases for each publication were included as automated indexing terms.

Computation of similarity score between indexing methods. Comparison of different indexing methods was carried out using simple pre-defined heuristics for estimation of a similarity score. The similarity score was quantified using the following set of rules:

- Comparison was carried out on a term-by-term basis, with a score of 1 attributed for every match between two indexing methods.
- No interpretation of meaning ("internet" is not the same as "web") or modification of terms was carried out before or during the comparison process.
- Stop words (e.g. and, a, the etc.) were ignored during comparison.
- No resolution of biological name variants was carried out prior to comparison.
- Plurals and singulars were not considered the same term.
- Conjunction (e.g. "&"), preposition and symbol (e.g. ";") tags were ignored during the comparison process.

3.1.2 Results

Based on the Data selection procedure described above, a total of 45 documents were selected out of approximately 19000 Connotea bookmarks having PubMed identifiers. These 45 documents were tagged by 264 users with 540 tags in Connotea, and they are indexed by a total of 1034 MeSH terms in PubMed. A total of 286 title-based automated indexing terms were extracted from these documents. Out of these, 102 tags terms were found to match the results of title-indexing, whereas only 59 tag terms match MeSH terms. A total of 31 terms match both indexing terms and MeSH terms (Figure 1). Percentages of matches among the three are summarized in table 1.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

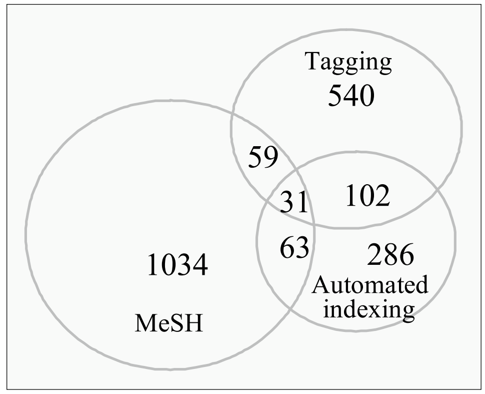

Figure 1. Overlaps among the three indexing methods (Numbers indicate the total number of terms used in each method and in the overlap areas)

% Similarity	All Tags (total 540)	Tags used by at least 2 users (Total 123)	Tags used by at least 3 users (Total 57)	Tags used by at least 4 users (Total 27)	Tags used by at least 5 users (Total 15)
Between Tagging and MeSH	11 % (59/540)	24 % (30/123)	19 % (11/57)	15% (4/27)	7% (1/15)
Between Tagging and Automated indexing	19 % (102/540)	42 % (52/123)	60 % (34/57)	56% (15/27)	53% (8/15)

Table 1. Similarities between tagging and MeSH and automatic indexing terms.

3.1.3 Discussion

The data show a low percentage of overlap (11%) between tagging and MeSH terms. This is hardly a surprise. As discussed in the literature (for example, Mathes, 2004; Hammond et al., 2005; Guy & Tonkin, 2006), precision is a general concern of tagging. While MeSH terms are mostly used for content description, tagging terms are employed by individual users for content as well as for other reasons. It appears as if users do not attempt to tag all content of a document but instead they highlight specific content or facts most interesting to them.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Two other observations are perhaps more interesting. First, the overlap between tagging terms and MeSH terms are very similar to the overlap between title-based indexing terms and MeSH terms. Second, the overlap between tagging terms and title-based indexing terms is much greater than that between tagging and MeSH terms. These, on one hand, suggest that users more likely select terms in titles for their tagging (following the principle of least effort as the title is visible when users do the tagging). On the other hand, the results also indicate that the user's tagging language is more close to natural language than to controlled vocabularies. It is interesting to note that as the number of users increases the percentage of matches between tagging terms and title-based indexing terms increases significantly, but that the percentage of matches between tagging and MeSH terms does not.

Clearly, the observations from this case study cannot be generalized without repeating the study for other similar datasets or testing with other methods of similarity score. The criteria for comparison of classification results also need to be evaluated.

3.2 The study on categorization of tags

As the above study suggested, many tags are used for different reasons or proposes other than as subject indicators. What are the reasons or purposes? Do they fall into predictable, conceptual groups? In this study, we investigate how well a set of categories can be applied to tags used for images.

In order to develop a set of categories we consulted literature on social tagging and categories of tags. Mathes (2004) identified tag categories including technical (rss, java, python), genre (photography, comics), self-organization (toread, todo), place names, years, colors, photographic terms, and ego (me). Partington's (2004) analysis suggested categories such as medium, subject, genre, name and location. Dubrinko's (2006) categories consist of events, personalities and social media tagging (tags developed by the flickr community, such as whatisinyourfridge, Circleinsquare). Schmitz (2006) developed common facets of tags: location, activity, event, people and things. Golder and Huberman's (2006) tag categories include topics (what or who it is about), proper nouns, genre (kind of thing), ownership (who owns it), refinements (terms which refine or qualify existing categories), adjectives, self reference (typically terms with "my") and task organization (items tagged for a specific task, such as tofind, indexingpaper). In all, several categories, such as location, people, subjects and events, are nearly universally recognized among

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

these authors as being highly used categories for image tags. Other categories are identified by one or two authors only.

The set of categories (and their brief definitions) developed for this study is shown in Figure 2. It was believed that these categories best reflect the tags employed on the flickr.com site for images. In this study we test how consistently these categories can be applied to the tags, what categories of tags occur most often, and what categories are often used together.

Figure 2. Categories of tags

Categories of Tags for Images

Compound – terms with two or more words combined. e.g.: newyorkcity, mysisteranna, christmascookies

3.2.1 Data and Methodology

The data used for this study was gathered through flickr.com's open APIs. The dataset, called *User Tags* in the subsequent discussion, was created by randomly choosing 14 flickr users and gathering their top ten most frequently used tags (providing a total of 140 tags). The API method utilized here returns a list of the tags alongside the number of times the user has employed the tag. After extracting the data from flickr.com, the dataset was processed using simple functions in Excel. For the file *User Tags,* category title(s) were entered in the empty column adjacent to each tag. Tags could be assigned to several categories. Thus, a tag such as "cross" could simultaneously be considered a verb, a thing, and an adjective.

In order to evaluate the model, five participants categorized the tags in both datasets. Each participant received the dataset alongside a document with the categorization rules, given above. The tag categorizations chosen by each participant were identified by participant and then all were combined into a single master file. The number of occurrences were then tallied and used to present percentages of usage and agreement among participants.

3.2.2 Results

The categorization of the 140 tags of the *User Tags* dataset resulted in the application of 897 category choices across the 5 participants. There were many instances of tags receiving two (occurring 143 times in the dataset) and a smaller number to receive three (occurring 27 times)

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

category choices. Table 2 presents the percentages of category assignment. On average, the most frequently used categories for the tags in this dataset were Place-name (28%), Compound (14%) and Thing (11%). The high incidence of tags utilizing two or more words is seen in the high percentage of the category choice Compound. The least frequently occurring categories for this dataset were Humor, Poetic, Number and Emotion all falling at below 1% usage.

	A		**B**		**C**		**D**		**E**		**Totals**	
Category	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Place-name	55	30.90	46	27.22	55	31.25	46	27.22	51	24.88	253	28.21
Compound	23	12.92	22	13.02	30	17.05	22	13.02	29	14.15	126	14.05
Thing	22	12.36	22	13.02	13	7.39	22	13.02	23	11.22	102	11.37
Person	17	9.55	16	9.47	14	7.95	16	9.47	16	7.80	79	8.81
Event	13	7.30	7	4.14	11	6.25	7	4.14	13	6.34	51	5.69
Unknown	5	2.81	11	6.51	5	2.84	11	6.51	11	5.37	43	4.79
Photographic	2	1.12	5	2.96	18	10.23	5	2.96	11	5.37	41	4.57
Time	9	5.06	7	4.14	8	4.55	7	4.14	9	4.39	40	4.46
Adjective	7	3.93	6	3.55	7	3.98	6	3.55	9	4.39	35	3.90
Verb	5	2.81	10	5.92	5	2.84	10	5.92	4	1.95	34	3.79
Place-general	4	2.25	6	3.55	1	0.57	6	3.55	8	3.90	25	2.79
Rating	6	3.37	4	2.37	1	0.57	4	2.37	5	2.44	20	2.23
Language	4	2.25	1	0.59	1	0.57	1	0.59	7	3.41	14	1.56
Living Thing	2	1.12	2	1.18	6	3.41	2	1.18	2	0.98	14	1.56
Humor	2	1.12	1	0.59	0	0.00	1	0.59	2	0.98	6	0.67
Poetic	0	0.00	1	0.59	0	0.00	1	0.59	4	1.95	6	0.67
Number	2	1.12	1	0.59	0	0.00	1	0.59	1	0.49	5	0.56
Emotion	0	0.00	1	0.59	1	0.57	1	0.59	0	0.00	3	0.33
Totals	178	100.0	169	100.0	176	100.0	169	100.0	205	100.0	897	100.0

Table 2. Percentages of category assignment for the *User Tags* dataset.

- number of times the category was assigned in the dataset.

% - percentage of category assignment.

In order to further evaluate the usefulness of the categorization model, the percentage of overall agreement of category assignment was calculated. The results of these calculations are presented in Table 3. For the *User Tags* dataset just over 34% of all category assignments to the tags were agreed upon by all five participants. The second highest percentage was that of the application of a category to a tag by a single individual, at roughly 29%. Approximately 57% percent of the category assignments for the tags were agreed upon by the majority of participants (3 of 5). Interestingly, it was discovered that the lowest percentage (approximately 11%) of agreement was that found for 4 participants.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Category Agreement	Instances of Agreement	Percentage
All Agreed	100	34.25
4 Agreed	32	10.96
3 Agreed	34	11.64
2 Agreed	41	14.04
1 Agreed	85	29.11
Totals	292	100.00

Table 3. Percentages of agreement of category assignment across 5 participants for the User Tags dataset.

In order to discover if the percentages of greater and lesser performing categories in terms of category agreement across the participants, further calculations were undertaken. The figures for these calculations are provided in Table 4. Instances of Category calculations were determined by the number of times the category was assigned to a tag by at least one participant. Greater that half (11) of the total number of categories utilized (18) were agreed upon by all participants greater than 50% of the time. The highest performance for this dataset was that of Place-name at roughly 85% agreement. The lowest agreement is seen in the category Poetic with 30%. The difference in the percentages seen in Tables 2 and 3 is the result of the calculations being based on the application of multiple category assignments for a single tag. Thus, calculations are based on the total number of category assignments in Table 3, while those in Table 4 are based on the total number of category assignments for a singular category.

3.2.3 Discussion

The data revealed that named geographical places (e.g. losangeles, amsterdam), things (e.g. water, boat) and people (e.g. steve, meghan) were the most frequently used tags by these flickr users. Compound tags (e.g. sanfrancisco, rockstarglasses) are the second highest occurring category of tags to be employed by flickr users. Unlike the other categories which describe concepts, the Compound category is applied for reasons of form alone. Compound tags are the result of the assignment of multiple terms for a single tag. Their high occurrence illustrates that flickr users prefer to use composite expressions rather than single terms to describe their images.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Category	Instances of Category	Total Possible Category Choices	Actual Category Choices	Percentage
Place-name	59	295	253	85.76
Time	10	50	40	80.00
Person	21	105	79	75.24
Event	14	70	51	72.86
Verb	10	50	34	68.00
Emotion	1	5	3	60.00
Compound	43	215	126	58.60
Adjective	12	60	35	58.33
Place-general	9	45	25	55.56
Thing	38	190	102	53.68
Unknown	17	85	43	50.59
Living Thing	6	30	14	46.67
Photographic	20	100	41	41.00
Humor	3	15	6	40.00
Number	3	15	5	33.33
Language	9	45	14	31.11
Rating	13	65	20	30.77
Poetic	4	20	6	30.00
Totals	292	1460	897	61.44

Table 4. Percentages of agreement of category assignment across 5 participants for the *User Tags dataset*.

There are several categories appearing with modest frequency in the dataset which deserve further discussion. Event tags (e.g. party, wedding) are used by most individuals in this dataset at least once in their top ten most frequently used tags. Tags with photographic or imaging categorization (e.g. cameraphone, nikon) are also used by a majority of flickr users at least once

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

in their top ten tag list, and for one individual this type of tag accounted for half of the top ten tags. This suggests that most individuals using flickr are interested in the processes and devices used in the creation of their images. Time (e.g. may, 2005) was an unusual category in terms of its frequency of use since it varied so widely among flickr users. While eleven of the fourteen users did not have a single instance of Time among their top ten tags, three users found it to be a useful categorization. For one of these flickr users it accounted for six of the ten top tags. The varied use of Time by flickr users is an interesting discovery. Flickr provides users with a means of organizing images into folders automatically based on the date recorded by the camera and saved into each image file. The reason why some users continue to tag their images by date needs further exploration.

Illustrating how difficult it was to discern the meaning for some of the tags was the high percentage (nearly 5%) of tags that could not be categorized by the participants. In some cases these Unknown tags (e.g. psfg) had highly personal meanings for the flickr users who used them and so they were not readily understood by the participants performing the categorization. In other cases the Unknown tags illustrate how important contextual knowledge is to the categorization of the tags. This was seen in the number of tags relating to flickr groups or photographic devices which were unfamiliar to several of the participants and therefore received the categorization Unknown. This situation is more appropriately part of a discussion concerning category agreement among participants and so we turn to this topic next.

The percentage of category agreement among the participants suggests that the overall categorization model is modestly effective in describing the concepts of the tags assigned by flickr users. Taken as a whole, there was a 34% agreement among the five participants on the categorization of all tags using this model (Table 2). Studies conducted by Markey (1984) and Wells-Angerer (2005) have shown percentages of agreement in image indexing to fall between 5% and 21%. While these studies are slightly different in that they were investigating the co-occurrence of indexing terms applied to images of artwork, the figures do offer some standard by which to judge the effectiveness of the categorization model. The performance of the model could be improved through the modification of those categories with low percentages of agreement across participants.

When the agreement results are viewed by category (Table 4) the percentages show where the model is the least and most effective. Eleven of the eighteen categories were successful in

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

producing agreement of all five participants more than half the time, while four categories (Place-name, Time, Person, and Event) were agreed upon at a rate greater than 70%. The categories to fall in the lower ranges of agreement seem to share several common characteristics. As was discussed above with the tags receiving the Unknown categorization, the lack of agreement across participants for the Language and Photographic categories was likely due to issues surrounding the specialized knowledge needed to assign them. It appears to have been difficult to identify a word as belonging to a foreign language unless the participant had some familiarity with it. Similarly, photographic or imaging processes and devices have unique vocabularies and so these tags caused some confusion for participants.

3.4 The study on convergent nature of tagging

In traditional controlled vocabulary-based indexing, all terms assigned to a document carry more or less equal weight. In social tagging, this is certainly not the case. Certain tags will be much more popular than others. In this study, we examine how tags converge and under what conditions and ratios convergence occurs. Convergence usually refers to the pattern of tag distribution where certain tags become much more popular than others over the entire dataset. In this project, tag convergence is considered in a slightly different sense where the URL is the unit of study. For a URL certain tags may be chosen much more often by users and will "float to the top" of the tag list for that URL. In other word, tags converge towards a few popular tags used in describing a document represented by the URL. This behavior implies a degree of consensus among users regarding what they think a document is about, albeit without any coordinating effort by any user in particular or by any third party.

Tag convergence needs to meet both individual and collective requirements. Individually, for a given URL or document, certain tag frequencies need to be much higher than others and therefore account for a high percentage of the total tag count fitting some preset ratio (e.g. 30/70). Collectively, certain percentages of URLs or documents in the selected sample pool (or the whole dataset) need to be converged (when judged individually). Since no standards or criteria currently exist to determine what ratio constitutes convergence, we explore various ratios and discuss possible criteria. In addition, the study examines results to see if certain factors are likely to affect convergence – namely, what factors can predict whether a URL's tags will converge.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

3.4.1 Data and Methodology

A semi-random sample of tagging data was retrieved from the website del.icio.us. First, all the tags in the "Popular tags" list were selected . Then for each tag on this list, two more tags from the "Related Tags" list were randomly selected. After excluding duplicate tags, the total number of tags was 93. For each of these 93 tags, up to 20 URLs were retrieved. If a tag had more than 20 URLs, only 20 were picked up. If a tag had less than 20 URLs, all available URLs were collected. The total number of URLs retrieved was 708. The dataset consisted of the URLs, the total number of tags per URL, the number of different tag terms associated with each URL, the frequency of each tag, and the number of users who have tagged that URL. The data was retrieved using a PHP program run at different times from 5/25/2006 to 5/28/2006.

The data was used to determine if tags followed a power law distribution. What this means is that for a URL certain tags would be chosen much more often by users than other tags. These tags would create a core while a large number of other tags would contribute less significantly and so establish the long tail seen in a power law distribution. Since this is an exploratory study, different distribution ratios were tested. The ratios tested were 20/80 (aggregated frequency of the top 20% tags account for 80% of the total tag count), 30/80, 30/70, and 40/60. The results of these tests tell us how concentrated the core was, and how narrow and long the tail was.

The statistical method Logistic Regression was used to determine what factors play a role in predicting whether a URL's tags would converge or not. The criterion variable is convergence, the predictor variables are total tag count (Total_Tag), number of distinct tags (Unique_Terms), and number of users (Num_of_Users).

3.4.2 Results

Table 5 gives the results of how tags behaved in relation to the various ratios tested.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Ratio	Total URLs	# converged	% converged
20/80	708	5	.07%
30/80	708	111	15.7%
30/70	708	296	41.8%
40/60	708	528	74.6%

Table 5. Convergence results for the 4 different ratios

Very few URLs met the criteria at the 20/80 ratio. The percentage of URLs to meet this criteria is less than .1%. The convergence rate improved for the 30/70 ratio, but it remained below 50%. At the 40/60 ratio many more URLs met the criteria.

To determine which factors affect whether tags on an URL meet the 30/70 convergence level, a Logistic Regression was run on the tagging data of the URLs in the sample. The criterion (or grouping) variable was "converge" (1 if the URL met the 30/70 ratio, and 0 if it did not). The predictor variables are: Total_Tags, Unique_Tags, and Num_of_Users. The results are given in Table 6.

	B	S.E.	Wald	Sig.	Exp(B)
Total_Tags	.002	.001	3.953	.047	1.002
Unique_Tags	.183	.020	82.690	.000	1.201
Num_of_Users	-.002	.002	.600	.438	.998
Constant	-3.707	.296	156.673	.000	.025

Table 6. Logistic Regression Results

The -2 Log Likelihood was 434.73. The goodness of fit Hosmer and Lemeshow test was significant with the Chi-square value of 31.72. The model was 90.5% correct in predicting cases that did not converge and 84.1% correct in predicting cases that converged, for a total correct percentage of 87.9%.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

3.4.3 Discussion

The results indicate the 20/80 ratio is a too stringent criterion for tag convergence. The 40/60 ratio shows a markedly better rate of convergence among the URLs. However, since almost half of the tag terms only account for 60% of the total tag count this ratio shows a still too scattered tag distribution. It seems the 30/70 ratio would be a good criterion for URL tag convergence. The ratio should also be applied to the whole sample set; namely, 70% of the 780 URLs in the sample need to meet the 30/70 tag convergence ratio.

It is interesting to note that while Unique_Tags is significant at the .05 level, Total_Tags is barely significant, and Num_of_Users is not a significant factor. What the model predicts is that the number of different terms used as tags is the most significant contributing factor, while the number of users (which has traditionally been thought of as a potential factor) does not make it as a predictor for a URL's converging potential.

4. Conclusions

Data from three case studies on social tagging reveal some interesting characteristics of social classification. First, there is little overlap among tags, automated indexing terms, and controlled vocabularies. While these have been predicted, our data suggest that tags are more similar to automatic indexing than to controlled vocabulary indexing, particularly when the number of users increases. Second, categorization of tags into some meaningful and stable groups seems to be feasible; this has implication when designing systems or interfaces for tag-based searching and browsing. Third, the number of unique tags was found to be the most significant factor to predict the convergence of tags and this occurs at 30/70 ratio rather than the 20/80 ratio many would have expected.

The results of these studies demonstrate the usefulness of studying characteristics of social classification. However, it should be emphasized that these findings are "snapshots" of dynamic tagging spaces that are still evolving. As such, the findings call for confirmation through large-scale studies over a longer period of time.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Lund, B., Hammond, T., Flack, M., and Hannay, T. (2005). Social bookmarking tools (II): A case study – Connotea. *D-Lib Magazine* 11(4). Retrieved from http://www.dlib.org/dlib/april05/lund/04lund.html

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research* 6: 155-177.

Mathes, A. (2004). Folksonomies: Cooperative classification and communication through shared metadata. Retrieved from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Michlmayr, E. (2005). A case study on emergent semantics in communities. *Workshop on Semantic Network Analysis, International Semantic Web Conference (ISWC2005), Galway, Ireland, November 2005*. Retrieved from http://wit.tuwien.ac.at/people/michlmayr/publications/michlmayr_casestudy_on_ _final.pdf.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. *Proceedings of International Semantics Web Conference (ISWC 2005)*: 522-536.

Ohmukai, I., Hamasaki, M., and Takeda, H. (2005). A proposal of community-based folksonomy with RDF metadata. *Proceedings of the ISWC 2005 Workshop on End User Semantic Web Interaction, Galway, Ireland, November 7, 2005*. Retrieved from http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-172/.

Partington, B. (2004). Tagging it up: Some suggestions for tagging your images. Retrieved from http://www.flickr.com/groups/central/discuss/2026/.

Schmitz, P. (2006). Inducing ontology from Flickr tags. *WWW2006.*

Szekely, B., and Torres, E. (2005). Ranking bookmarks and bistros: intelligent community and folksonomy development. Retrieved from http://torrez.us/archives/2005/07/13/tagrank.pdf.

Voss, J. (2006). Collaborative thesaurus tagging the Wikipedia way. Retrieved from http://arxiv.org/abs/cs.IR/0604036.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

References

Bearman, D., and Trant, J. (2005). Social terminology enhancement through vernacular engagement: Exploring collaborative annotation to encourage interaction with museum collections. *D-Lib Magzine*, 11(9). Retrieved from http://www.dlib.org/dlib/september05/bearman/09bearman.html

Dubrinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. (2006). Visualizing tags over time. *WWW2006.*

Gaizaukas, R., Cunningham, H., Wilkis, Y., Rodgers, P., and Humphreys, K. (1996). Gate: An environment to support research and development in natural language engineering. *Proceedings of Eighth IEEE International Conference on Tools with Artificial Intelligence, Nov 1996*.

Golder, S., and Huberman, B.A. (2006). The structure of collaborative tagging systems. *Journal of Information Science* 32(2): 198-208. Retrieved from http://arxiv.org/abs/cs.DL/0508082.

Guv, M., and Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine* 12(1). Retrieved from http://www.dlib.org/dlib/january06/guy/01guy.html.

Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social bookmarking tools (I): A general review. *D-Lib Magazine* 11(4). Retrieved from http://www.dlib.org/dlib/april05/hammond/04hammond.html.

Joseph, S., Yukawa, J., Suthers, D., and Harada, V. (2006). Searching emergent vocabularies: Exploring methods to reduce cognitive load during web navigation and resource contribution. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences: Volume 07* (January 4-7, 2006). Washington, DC: IEEE Computer Society. Retrieved from http://dx.doi.org/10.1109/HICSS.2006.429.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

ISSN: 2324-9773

Lin, X., Beaudoin, J. E., Bui, Y., Desai, K. (2006). Exploring characteristics of social classification. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-19. doi:10.7152/acro.v17i1.12491

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York, NY: W. W. Norton & Company.

Weinberger , D. (2005). Tagging and why it matters. Berkman Center Research Publication No. 2005-07. Retrieved from http://cyber.law.harvard.edu/home/uploads/507/07-WhyTaggingMatters.pdf.

Wells-Angerer, T. (2005). *A study of retrieval success with original works of art comparing the subject index terms provided by experts in art museums with those provided by novice and intermediate indexers.* Unpublished Master's paper, School of Information and Library Science, University of North Carolina, Chapel Hill.

Xu, Y., and Zhang, L. (2005). Personalized information service based on social bookmarking. *Lecture Notes in Computer Science* 3815: 475-476.

Xu, Y., Zhang, L., and Liu, W. (2006). Cubic analysis of social bookmarking for personalized recommendation. *Lecture Notes in Computer Science* 3841: 733-738.

ISSN: 2324-9773