This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

Resource discovery through social tagging : a classification and content analytic approach

Goh, Dion Hoe-Lian; Chua, Alton Yeow Kuan; Lee, Chei Sian; Razikin, Khasfariyati

2008

Goh, D. H. L., Chua, A. Y. K., Lee, C. S., & Razikin, K. (2009). Resource discovery through social tagging: a classification and content analytic approach. Online Information Review, 33(3), 568–583.

https://hdl.handle.net/10356/91019

https://doi.org/10.1108/14684520910969961

© 2008 Emerald Group Publishing Limited. This is the author created version of a work that has been peer reviewed and accepted for publication by Online Information Review, Emerald Group Publishing Limited. It incorporates referee’s comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [http://dx.doi.org/10.1108/14684520910969961].

Downloaded on 20 Apr 2021 10:13:36 SGT

OTHER ARTICLE

Resource discovery through social tagging: a classification and content analytic approach

Dion Hoe-Lian Goh, Alton Chua, Chei Sian Lee and Khasfariyati Razikin
Nanyang Technological University, Singapore

Abstract

Purpose – Social tagging systems allow users to assign keywords (tags) to useful resources, facilitating their future access by the tag creator and possibly by other users. Social tagging has both proponents and critics, and this paper aims to investigate if tags are an effective means of resource discovery.

Design/methodology/approach – The paper adopts techniques from text categorisation in which webpages and their associated tags from del.icio.us and trained Support Vector Machine (SVM) classifiers are downloaded to determine if the documents could be assigned to their associated tags. Two text categorisation experiments were conducted. The first used only the terms from the documents as features while the second experiment included tags in addition to terms as part of its feature set. Performance metrics used were precision, recall, accuracy and F1 score. A content analysis was also conducted to uncover characteristics of effective and ineffective tags for resource discovery.

Findings – Results from the classifiers were mixed, and the inclusion of tags as part of the feature set did not result in a statistically significant improvement (or degradation) of the performance of the SVM classifiers. This suggests that not all tags can be used for resource discovery by public users, confirming earlier work that there are many dynamic reasons for tagging documents that may not be apparent to others.

Originality/value – The authors extend their understanding of social classification and its utility in sharing and accessing resources. Results of this work may be used to guide development in social tagging systems as well as social tagging practices.

Keywords Social roles, Resources, Resource management

Paper type Research paper

Introduction

The increasing popularity of social computing or Web 2.0-based applications has empowered users to create, publish and share resources on the web. Such user-generated content may include text (e.g. blogs, wikis), multimedia (e.g. YouTube) and even organisation/navigational structures providing personalised access to web content. The latter includes social bookmarking/tagging systems such as del.icio.us and Cite-U-Like.

Social tagging systems allow web users to annotate useful sites by assigning keywords (tags) and possibly other metadata, facilitating their future access by the tag

creator (Macgregor and McCulloch, 2006). These tags may be shared by other users of the social tagging system, in effect creating a community where users can create and share tags pointing to useful resources (Angus *et al.*, 2008). Put differently, the resulting user-generated tags constitute an organisational structure that supports access to resources via browsing or searching. However, tags are "flat", lacking a predefined taxonomic structure, and their use relies on shared, emergent social structures and behaviours, as well as a common conceptual and linguistic understanding within the community (Marlow *et al.*, 2006). Tags are therefore also known as "folksonomies", short for "folk taxonomies", suggesting that they are created by lay users as opposed to domain experts or information professionals such as librarians.

Proponents of social tagging have argued that it has some advantages over traditional classification. For example, hierarchical taxonomies may, in some instances, be too rigid to organise resources that contain a diversity of topics, and the non-hierarchical nature of tags might be better suited for this purpose (Morville, 2005). In addition, Bowker and Star (1999) have suggested that because traditional classification methods tend to rely on specialists such as trained cataloguers to organise and describe information, they may use terms that are specific to a specialised community, resulting in under-accessed resources. Thus, instead of relying on experts to categorise resources, tags harness the tacit knowledge of ordinary people (Lakoff, 1990), which presumably better reflects the way users want to keep track of information.

However, critics of social tagging have pointed out several disadvantages. These include the ambiguity of tags due to a lack of controlled vocabulary (Macgregor and McCulloch, 2006), and the use of subjective or ego-centric tags (e.g. "toread", "me", "todo") that have meaning only for the tag creator or a select few within a group of users (Golder and Huberman, 2006). Furthermore, the decision to tag may sometimes also be driven by the tag creator's self-serving agenda (Chua, 2003). This could lead to the problem of tag spamming, where non-related tags are indiscriminately used to draw traffic to certain websites (Koutrika *et al.*, 2007). In sum, these issues possibly hinder the use of tags for sharing, organising and navigating web resources.

Despite these shortcomings, the use of social tagging continues to grow in popularity. Concurrently there is an emerging body of research that explores their effectiveness for resource organisation and sharing. For example, from a user's perspective, work has been conducted on the motivations behind tagging (Ames and Naaman, 2007), comparing the use of tags against author-assigned index terms in academic papers (Kipp, 2006), and on tagging dynamics and usage (Farooq *et al.*, 2007). Machine learning approaches have also been used to study the ability of tags to classify blogs using text categorisation methods (Sun *et al.*, 2007), and for investigating the effectiveness of tags for classifying web resources in del.icio.us (Razikin *et al.*, 2008).

The goal of the research reported here was to extend existing work in investigating the effectiveness of tags for resource discovery by using both a machine learning and a content analytic approach. Specifically, we obtained webpages and their associated tags from del.icio.us and studied whether the tags were effective navigational aids to those resources. We adopted techniques drawn from text categorisation (Sebastiani, 2002) and argue that an effective tag is one in which a classifier can assign documents

with high precision and recall. The rationale here is that if a classifier is able to accurately assign documents to their respective tags, then such tags are useful for organising resources, implying that users would be able to utilise them for accessing information. Further, to better understand how tags are created, we conducted a content analysis to study the relationships between the use of a tag on a document and the document's terms.

To the best of our knowledge, there are a limited number of studies that have been conducted on examining the effectiveness of tags for resource discovery using both a text categorisation and a content analytic approach. While the former provides an automated technique for investigating effectiveness and has been successfully used in a variety of domains, it does not adequately account for the performance results. Our content analysis thus complemented the machine learning approach by examining, in greater detail, the characteristics of tags that made them effective or ineffective for resource discovery. Our work can therefore be used as a basis for future research in this area as well as for designing techniques that help users in both seeking resources via tags and suggesting tags for organising resources.

Related work

The use of tagging has become a popular way of organising and accessing web resources. Sites such as del.icio.us, Flickr, YouTube and Last.fm offer this service to their users. Social tagging has correspondingly also attracted much research, concentrating on areas such as the architecture and implementation of systems (e.g. Hammond *et al.*, 2005; Puspitasari *et al.*, 2007), usage patterns in tagging systems (e.g. Angus *et al.*, 2008; Golder and Huberman, 2006), user interfaces (e.g. Farooq *et al.*, 2007; Li *et al.*, 2007), and the use of social tagging in search systems (e.g. Hotho *et al.*, 2006; Yanbe *et al.*, 2007), among others. Here we focus our review on related literature that has investigated the effectiveness of tags as a means for organising and discovering resources.

Firstly, tag effectiveness has been studied using different machine learning approaches. For example, Brooks and Montanez (2006) used 350 popular tags from Technorati and obtained 250 of the most recent blog articles from the collected tags. Clustering was conducted on these articles and the results suggested that tags were able to organise articles in a broad sense, but were not as effective in indicating the specific content of an article. Similarly, Berendt and Hanser (2007) compared the performance of blog post classification using features derived from tags, titles and article bodies, and found that tags together with article bodies yielded better classification accuracies than using any of them alone. Rather than individual blog posts, Sun *et al.* (2007) focused on classifying whole blogs with tags, and compared the classification results based on tags alone, tags together with blog descriptions (short abstract), and blog descriptions alone. It was found that tags together with descriptions had the best classification accuracy, while tags alone were more effective than using blog descriptions alone for classification.

Besides blogs, Razikin *et al.* (2008) studied the effectiveness of tags in classifying web content in del.icio.us. The corpus consisted of 100 tags and 20,210 documents. Using Support Vector Machines (SVM), experiments were run on two feature sets: document terms only and document terms plus tags. Surprisingly, the results indicated that using document terms only produced better classification results in terms of

F-measure than using terms plus tags. Nevertheless, both F-measures from the experiments were relatively low at 0.59 and 0.56, suggesting that not all tags were effective at resource discovery, and that the classifier's performance was likely to be influenced by the tag creator's motivations and his or her interpretation of the document content. Levy and Sandler (2007) investigated tags as a source of metadata to describe music. Using 236,974 tags collected for 5722 tracks from Last.fm and MyStrands, a correspondence analysis was performed to visualise a two-dimensional semantic space defined by the tags. Findings from their work suggest that tags are effective in capturing music similarity and could be used to describe mood and emotion in music.

From the perspective of a tag creator, work has been done to compare tags with controlled vocabularies to discover how they differ. For example, Lin *et al.* (2006) evaluated tags from Connotea and Medical Subject Heading (MeSH) terms and found that there was only 11 per cent similarity between MeSH terms and tags. The authors argued that this was because MeSH terms serve as descriptors while tags primarily focus on areas that are of interest to users. Likewise, Kipp (2006) compared tags with author-supplied tags from Cite-U-Like and indexing terms from INSPEC and Library Literature to determine the usage overlap. The results showed that approximately 21 per cent of the tags were the same as the indexing terms. The reason for the divergence was attributed to the different emphases placed on an article by these two groups. For example, tag creators may consider time management information (e.g. "todo", "toread", "maybe") to be important as a tag for articles to indicate a desire to read them in the future, while such information will be disregarded by expert indexers. Taken together, these findings suggest that experts who create indexing terms and tag creators employ vocabularies that have little overlap, potentially causing access problems in social tagging systems.

In sum, while our present study shared the goal of investigating tag effectiveness with the above studies, we have complemented and extended such work in the following ways:

- we focused on del.icio.us, which captures a wide spectrum of content found on the web; and
- we addressed the issue of effectiveness by adopting both a machine learning and a content analytic approach, which taken together, can better discern characteristics of tags that help users discover relevant, useful resources.

In addition, we have also extended the work of Razikin *et al.* (2008) by exploring in greater detail the relationship between tag/term use and effectiveness, as well as the possible reasons for poor classifier performance through an in-depth case study of a tag and its associated documents.

Dataset and methodology

The dataset for the present study was obtained in late 2007 from del.icio.us, a popular social tagging service. Similar to the work of Brooks and Montanez (2006), we mined tags from the popular tags page, and as such the tags would be biased towards to the more commonly used ones. Nevertheless, by using popular tags, we were assured that there would be a sufficiently large amount of documents available for our work. From the list of popular tags in del.icio.us, we randomly sampled 150 tags and up to 150

English-language webpages/documents associated with each tag for a total of 22,500 web documents. Documents that were primarily non-textual (e.g. images and video) were discarded. In addition, HTML, style sheets and other scripting elements were removed. Further, after stopword removal and stemming, we applied the commonly used TF-IDF weighting scheme to form the final feature set of documents for our classifier. In our dataset, each web document had an average of 6.22 tags. There were 1352 web documents with one tag each while only one document (the Technorati webpage) had the largest number of tags, which was 100.

In this study, two text categorisation experiments were conducted. SVM was the machine learning classifier selected as it is commonly used with good performance in web-based text categorisation studies. Specifically, we used the SVM^{light} package (Joachims, 1998). The first experiment used only the terms from the documents as features and served as a baseline for the second experiment, which included tags in addition to terms as its feature set. Here, each tag was given equal weight. Since the SVM implementation was a binary classifier, we created one classifier for each tag. The training samples for each classifier consisted of both positive examples (web documents associated with the tag) and negative examples (documents not associated with the tag). In total, 150 classifiers were trained with the default options of the SVM^{light} package. Of the entire dataset, the web documents associated with each tag were further divided into two subsets: two-thirds were used for training the classifier while the other third was used for testing. Macro-averaged accuracy, precision, recall and F1 (the measure of a test's accuracy) were used as measures to determine the effectiveness of tags in helping users in accessing their associated web documents.

Results and analyses

Classifier performance

A summary of the mean accuracy, precision, recall and F1 scores for the 150 tags used in the two experiments is shown in Table I. Surprisingly, the inclusion of tags in the feature set only marginally improved precision and F1, but caused a slight degradation in accuracy and recall. However, *t*-tests to compare the differences between the means of these measures showed that none of the differences were statistically significant even at the 0.1 level, therefore suggesting that the addition of tags does not cause a change in the performance of the SVM classifiers.

Taking the results of the two experiments together, the performance measures of the 150 tags on average were approximately 80 per cent for accuracy, 90 per cent for precision and 46 per cent for recall. When considering the F1 measure (about 59 per cent), the results suggest that tags are reasonably able to assist users in information access, but users should not entirely rely on them to obtain resources to meet their information needs. Put differently, the accuracy metric suggests that the classifier could not determine if a web document should be associated or not associated with a

Table I.
Tag statistics for accuracy, precision, recall and F1 scores

	Accuracy (%)		Precision (%)		Recall (%)		F1 (%)	
	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2
Mean	80.24	79.55	89.64	92.96	46.11	45.36	59.43	59.38

Notes: Experiment 1 – terms only; Experiment 2 – terms and tags

tag about 20 per cent of the time. The precision metric indicates that of all documents classified as being associated with a tag, only approximately 10 per cent were incorrect. Recall suggests that on average, only approximately 46 per cent of all documents were correctly classified as being associated with their respective tags, implying misclassification at around 54 per cent.

The mixed performance of the measures in Table I indicates that the SVM classifiers performed significantly better for some tags than for others. We therefore sought to investigate reasons for this by looking at the properties of the tags themselves and the documents associated with them. In particular, we adapted Golder and Huberman's (2006) broad classification of tags into extrinsic and intrinsic categories. According to their definition, extrinsic tags are those that identify or describe a resource, and whose meanings are non-personal and are understood among the community of tag users. In contrast, intrinsic tags are those that have subjective meanings and are personal or only relevant to a particular tag user. One would expect that extrinsic tags (e.g. article, food, etc.), being those that characterise a resource more objectively, would perform better than intrinsic tags (e.g. cool, best, etc.), which tend to have meaning only to the creator of the tag.

Table II shows the ten best performing tags in terms of F1 scores and also reveals that these were all extrinsic tags. Interestingly, the top five were food-related (e.g. "cooking", "baking" and "foodblog"), which was probably due to the fact that the vocabulary is well-defined and understood by users. For example, an examination of a sample of webpages for "recipe" suggests that the majority contained recipes and included the term within the content. There was a minority of pages that were not food-related, but nevertheless were recipes applied to a different context, and were the likely causes of misclassification. Examples included a site containing programming tips (e.g. "python cookbook") and an article on assembling an in-car computer ("recipe for building an in-car PC"). Thus even with tags that had seemingly well-understood meanings and usage, this example illustrates that it can be expected that some tag creators will adopt alternative definitions, resulting in potential access problems by other users. However, this is partially mitigated by the fact that web resources are typically tagged with multiple terms, in effect, creating multiple paths to a resource. Still the effect of such "dead-ends" could lead to inefficiencies in a user's search session. It was also interesting to note that of the five food-related tags, the more specific tags "dessert", "cooking", "baking" and "recipe" had better accuracy, precision and recall

Tag	Accuracy (%)	Precision (%)	Recall (%)	F1 score
itunes	85.33	76.92	80.00	0.78
food	84.67	84.67	74.00	0.79
podcast	89.33	90.47	76.00	0.83
government	90.67	87.50	84.00	0.86
comics	91.33	80.33	98.00	0.88
dessert	93.33	90.00	90.00	0.90
cooking	93.33	91.67	88.00	0.90
baking	93.33	93.48	86.00	0.90
recipe	94.00	91.83	90.00	0.91
foodblog	96.00	92.31	96.00	0.94

Table II.
Ten best performing tags in terms of F1 scores

scores than the more general "food" tag. This reflects an implicit hierarchy in which more specific tags result in better access performance than less specific ones.

Table III shows the ten worst performing tags in terms of F1 scores. Surprisingly, only one intrinsic tag appears in the list ("fun"), while the rest are extrinsic tags that tend to have broad or ambiguous meanings such as "service", "photography" and "utility". For example, in examining a sample of webpages associated with the intrinsic tag "fun", as expected, it appeared to comprise content that users think are fun. However, because what constitutes "fun" varies between users, the result was a long, diverse, subjective list consisting of cartoons, jokes, games, recreation ideas, holiday photos and programming hacks, among other topics. The tag "service" scored the worst in F1, precision and recall, and like "fun", referred to a broad range of topics including service computing, web services, email services, commercial services, and so on. It appears that one of the reasons for the low scores for "service" is because of its generality, resulting in almost any webpage being able to be included or excluded from this tag. Similarly, "photography" suffered due to the wide range of content, including cameras, photographers, images, Photoshop tips and tricks, and studios.

Tag/term analysis

To obtain a better understanding of the relationship between the application of a tag to a document and the document's terms, we identified the five most commonly used terms (apart from stopwords) in the documents of the ten tags with the highest and lowest F1 scores. These are presented in Tables IV and V respectively. In the tables, "Tag occurrences" refers to the number of times the tag itself appeared in the documents associated with the tag, while the frequency of the other terms were obtained by counting the number of times each term appeared within the documents of a particular tag. For example, in Table IV, the number of times "itunes" appeared in the 150 documents tagged with that term was 451. At the same time, the terms "song" and "music" were two of the top five terms in the subset of documents of the ten tags with the highest F1 scores, and they appeared 1,105 and 1,017 times respectively among the 150 documents tagged with "itunes".

The results in Table IV show that the high F1 scores were associated with high tag occurrences (e.g. "recipe" and "baking"). In addition, the high F1 scores were also mostly associated with high occurrences of related terms as well as low occurrences of unrelated terms. For example, in tags such as "recipe" and "itunes", there were high

Table III.
Ten worst performing tags in terms of F1 scores

Tag	Accuracy (%)	Precision (%)	Recall (%)	F1 score
service	56.67	5.88	2.00	0.03
photography	60.00	14.29	4.00	0.06
utility	61.33	16.67	4.00	0.06
fun	64.67	28.57	4.00	0.07
software	66.67	50.00	4.00	0.07
art	58.00	15.79	6.00	0.09
imported	62.67	25.00	6.00	0.10
list	62.67	25.00	6.00	0.10
article	57.33	18.18	8.00	0.11
resource	58.00	19.04	8.00	0.11

Tag	F1 score	Tag occurrences	Commonly used terms				
			food	cakes	cooking	song	music
itunes	0.78	451	6	2	1	1,105	1,017
food	0.79	609	609	233	237	4	26
podcast	0.83	486	15	2	6	247	507
government	0.86	623	51	4	4	10	30
comics	0.88	275	7	2	1	42	57
baking	0.90	899	608	370	214	3	13
cooking	0.90	310	739	255	310	5	22
dessert	0.90	190	565	572	161	3	8
recipe	0.91	1922	694	352	277	3	11
foodblog	0.94	15	2,001	1,177	490	20	32
Term frequency in top ten tag dataset			5,295	2,969	1,701	1,442	1,723
Term frequency in entire dataset			10,264	3,076	3,030	12,362	23,962

Table IV. Frequency of commonly used terms in the ten tags with the highest F1 scores

Tag	F1 score	Tag occurrences	Commonly used terms				
			picture	photo	food	apple	web
service	0.03	134	20	44	12	12	181
photography	0.06	255	244	1,220	11	9	256
utility	0.06	23	26	103	1	21	307
fun	0.07	557	110	403	81	35	300
software	0.07	249	44	253	6	20	431
art	0.09	1,905	120	579	17	7	413
imported	0.10	2	98	559	33	33	540
list	0.10	974	102	166	83	16	645
article	0.11	659	101	292	45	15	809
resource	0.11	103	28	255	15	9	524
Term frequency in bottom ten tag dataset			893	3,874	304	177	4,406
Term frequency in entire dataset			10,603	31,427	10,264	4,766	5,990

Table V. Frequency of commonly used terms in the ten tags with the lowest F1 scores

occurrences of semantically related terms such as "food", "cakes" and "cooking", and "song" and "music", respectively. Conversely, "recipe" and "itunes" had low occurrences of the non-related tags "song" and "music", and "food", "cakes" and "cooking", respectively. For the food-related tags, an additional characteristic was that the occurrence of the related terms accounted for a large proportion of the occurrences in the entire dataset. For example, the term "food" appeared 5,295 times in the subset of documents of the ten tags with the highest F1 scores, and this accounted for more than 50 per cent of all occurrences (10,264) in the entire dataset of documents. A more striking example is the term "cakes", which appeared 2,969 times in the document subset, accounting for almost 97 per cent of all occurrences of the term. In both these examples, these terms mostly appeared in documents tagged as food-related.

In contrast, Table V shows that low F1 scores were associated with comparatively lower tag occurrences (e.g. "utility" and "imported"). In addition, the proportion of occurrences of commonly used terms against the entire document dataset was mostly much lower. For example, the term "picture" occurred 893 times in the documents associated with the bottom ten tag dataset and this accounted for about 8 per cent of all

occurrences (10,603) in the entire dataset of documents. In addition, although the term "photo" had the highest occurrence (1,220) in documents tagged as "photography", this accounted for only approximately 12 per cent of all occurrences (31,427) in the entire dataset. Further, in both these examples, these terms seemed to appear across most of the documents in the bottom ten tag dataset. This suggests that the terms in this collection of documents have less discriminating power, accounting for the poor performance of the SVM classifier.

Tables VI and VII offer a different perspective by showing the TF-IDF values of the commonly used terms found in Tables IV and V. Here, TF-IDF values indicate the weights or importance of a term in our entire dataset of documents by taking into account a term's occurrence both within a document and across the entire dataset. In Table VI, terms semantically associated with their respective tags had higher TF-IDF values than terms that were not. For example, the terms "song" and "music" had comparatively higher TF-IDF values for the tags "itunes" and "podcast" than food-related terms such as "cakes" and "cooking". In contrast, there was mostly no discernable pattern for the distribution of TF-IDF values in Table VII with the exception of the terms "picture" and "photo" associated with the tag "photography". Taking Tables IV to VII together, our findings suggest that tags whose semantic meanings are more specific will result in better classification performance than those that are more general. This appears to be independent of whether a tag is extrinsic or

Table VI.
TF-IDF values of commonly used terms in the ten tags with the highest F1 scores

	Commonly used terms				
Tag	food	cakes	cooking	song	music
itunes	0.00019	0.00062	0.00080	0.02106	0.02319
food	0.04970	0.02544	0.01890	0.00024	0.01136
podcast	0.00842	0.00120	0.01538	0.01568	0.02053
government	0.01080	0.00108	0.00592	0.00022	0.00340
comics	0.00364	0.00282	0.00055	0.00229	0.01021
baking	0.04971	0.02544	0.02147	0.00018	0.00125
cooking	0.04450	0.02249	0.02347	0.00018	0.00108
dessert	0.02290	0.02739	0.02819	0.01488	0.01337
recipe	0.02162	0.00956	0.01830	0.00019	0.00019
foodblog	0.04971	0.01034	0.01538	0.00019	0.00013

Table VII.
TF-IDF values of commonly used terms in the ten tags with the lowest F1 scores

	Commonly used terms				
Tag	picture	photo	food	apple	web
service	0.00635	0.00768	0.00867	0.04016	0.01075
photography	0.01142	0.04217	0.00454	0.00480	0.01003
utility	0.00635	0.00768	0.00454	0.04016	0.00970
fun	0.00671	0.03956	0.00495	0.00892	0.01890
software	0.00384	0.00937	0.00117	0.00469	0.01075
art	0.00569	0.03956	0.00419	0.00201	0.01003
imported	0.00359	0.01075	0.00450	0.00468	0.00947
list	0.00359	0.00927	0.00398	0.04016	0.01816
article	0.00011	0.00116	0.00752	0.00010	0.00182
resource	0.01075	0.00359	0.00398	0.00464	0.01890

intrinsic. Put differently, a more crucial determinant in resource discovery is that the vocabulary behind the tag is well-defined, meaning that there is a set of commonly used terms associated with the tagged documents.

The exception was the tag "foodblog", which did not occur frequently (15 times) in the documents but which obtained the best F1 score. This is likely due to the fact that documents associated with "foodblog" had high occurrences of food-related terms, and also that documents not associated with the tag did not contain these terms as part of their content. In addition, this example suggests that tags are not merely metadata but are more "content-associated" with documents (Berendt and Hanser, 2007). Furthermore, effective tags may encompass those that describe a resource (e.g. "dessert") as well as those that describe a category to which this resource belongs (e.g. "foodblog") (Golder and Huberman, 2006).

"Photography": an analysis of an ineffective tag

Of the ten tags with the lowest F1 score, the SVM classifier performed rather poorly on certain tags that were expected to yield good results. For example, the tag "photography" appeared to have a specific meaning and yet it obtained the second lowest F1 score in the entire tag dataset. In this section, we attempt to uncover the reasons behind this by conducting an analysis of the URLs associated with this tag.

The output of the SVM classifier showed that there were 12 false positives, meaning that the classifier incorrectly tagged 12 documents as "photography" when they should have been associated with something else. In addition, there were 48 false negatives, meaning that these documents were to be tagged as "photography" but were instead associated with other tags.

An examination of the content of the URLs of the 12 false positives attributed to the SVM classifier reveals that five of the documents could have been tagged with "photography" but for some reason the tag creator did not do so. Put differently, the number of false positives would have been lower had the tag creator associated "photography" with these documents. For example, the URL http://digital-photography-school.com/blog/blur-movement/was tagged "design", "toread", "technique", "cool", "website", "tutorial", "interesting", "tricks", and "photo", among other keywords. The use of intrinsic tags (e.g. "cool", "toread") indicates that these were created for personal use, and that other users would have difficulty associating these tags with a website on digital photography techniques. Although the extrinsic tags "photo", "tutorial" and "tricks" do provide a possible navigation path to the website, it is also interesting to note that the tag "photography" was not used despite the term's appearance in the URL and document content. Again, from the perspective of the tag creator, however, this is understandable because these tags could have been created for personal access. In addition, because there are variations for "photography" (e.g. "photo" and "photos"), the lack of established guidelines for tag creation means that certain word forms could have inadvertently been overlooked despite their obvious usefulness. This finding illustrates the vocabulary mismatch problem, arising from the lack of a controlled vocabulary in social tagging systems (Macgregor and McCulloch, 2006).

Our analysis indicates that of the 48 URLs identified as false negatives, 35 had no relation to photography but were tagged as "photography" by users and thus were incorrectly classified by the SVM classifier. In other words, they were falsely classified

as false negatives. A total of 13 URLs appeared to be related to photography and were therefore actual false negatives. Of the 35 other URLs mentioned earlier, they consisted of a varied collection of search engines, news sites, personal pages, shop fronts and so on. We note that some of these could have a tangential relation to photography such as Google Earth (http://earth.google.com), which contains some photographs uploaded by users but photography is not the main focus of the site, or the Daily Color Scheme (http://beta.dailycolorscheme.com), which suggests colour schemes that could be used for digital images and art. However, our analysis also reveals that there were sites that had no association with photography. At best, our analysis leads us to the conclusion that the use of the tag "photography" is subjective in that it has meaning only to the tag creator or a selected group of users. Alternatively, these findings may suggest a case of inaccurate assignment of tags and/or an example of the vocabulary mismatch problem. At worst, our findings illustrate an example of tag spamming (Koutrika *et al.*, 2007) in which tag creators mislead users into visiting certain websites by using a variety of popular but unrelated terms. It is interesting to note however that the SVM classifier was able to identify a number of such sites.

Discussion

In summary, three main findings have emerged from our study. First, our experiments revealed that the inclusion of tags as part of the feature set did not result in a statistically significant improvement (or degradation) of the performance of the SVM classifiers, which suggests that tags vary in their effectiveness as navigational aids to resources. This can be attributed to the lack of a controlled vocabulary in social tagging, resulting in a proliferation of tags of varying quality (Macgregor and McCulloch, 2006), and because tags may be created for a variety of reasons, of which providing public access to resources is but one (Ames and Naaman, 2007).

Second, among the ten worst performing tags in terms of F1 scores, nine had broad or ambiguous meanings such as "service", "fun" and "utility". Conversely, among the ten best performing tags in terms of F1 scores, the top five were specifically related to food (e.g. "cooking", "baking" and "foodblog"). Here, it seems that among tags created for the purposes of sharing resources by a community of users in del.icio.us, those with broader meanings tend to be less effective than those with more precise or well-understood definitions. Therefore, if the purpose is to share content among users, tag creators would do well to not only employ tags that come from a shared vocabulary among the users of the social tagging system, but also to pick tags with more specific meanings (Sen *et al.*, 2006).

Third, tags with high F1 scores appeared to be associated with documents that had high tag occurrences within their content (e.g. "recipe" and "baking" tags). In addition, the high tag scores were also mostly associated with high occurrences of related terms as well as low occurrences of unrelated tags. Conversely, tags with low F1 scores appeared to be associated with comparatively lower tag occurrences in the document content (e.g. "utility" and "imported" tags). In addition, our study also found that the proportion of occurrences of semantically related terms (see Tables VI and VII) was mostly higher for tags with high F1 scores than for those with low F1 scores. Taken together, our results suggest that the effectiveness of resource sharing among a community of users can be enhanced if tag creators select tags whose meanings are closely associated with the terms found in the document content.

To understand why certain tags that were expected to perform well yielded poor results instead, we undertook an analysis of "photography" – a tag with a seemingly well-understood definition but which obtained the second lowest F1 score in the entire dataset. Of the 12 false positive documents, five could have been tagged with "photography" but were not. Instead, tags such as "design", "toread", "technique", "cool", "website" and "tutorial" were used. It appears that these tags were meant more for personal use than to be shared with other users (Golder and Huberman, 2006). Of the 48 false negative documents, 35 had no relation to photography but were tagged as "photography". These included search engines, news sites, personal pages and shop fronts. This finding suggests either evidence of tag spamming (Koutrika *et al.*, 2007) or that tags have a variety of uses known only to the tag creators (e.g. Ames and Naaman, 2007). Users therefore cannot naively assume that tags have been created to facilitate navigation to web resources. Instead, as with other user-generated content, the onus is on the user to understand and accept the strengths and limitations of social tagging.

Conclusion

Social tagging is an increasingly popular means of organising content in websites. In this paper, we have investigated if tags can help users to access relevant web resources effectively. We randomly sampled 150 popular tags from del.icio.us and up to 150 English-language webpages associated with each tag. We then trained SVM classifiers to determine if our dataset of documents could be associated accurately with their corresponding tags. As discussed, we obtained mixed results (see Table I) and this can be explained by the fact that tags can be employed for a variety of uses, and that tag creators have many reasons for tagging documents that may not be apparent to others.

However, the fact that some tags do have high F1 scores indicates that there are benefits in allowing users to create and share such organisational/navigational structures to access resources. Here, we provide some recommendations for effective use of social tagging for resource discovery based on our findings.

First, tag creators should make a better distinction between tags meant for personal use and those for sharing (i.e. individual consumption versus public consumption). For example, in our analyses, poorly performing tags from the perspective of the SVM classifier seemed to be those created for personal use such as "fun" (which is subjective) and "list" (the contents of which have meaning only to the list creator). Restricting access to personal tags to the individual tag creator or only to selected users would be a good first step to increasing the utility of tags for resource access among public users. Another possibility is to organise and display tags into those that are unique to a particular creator and those that have been created by multiple users. This approach should give tag consumers an indication of the purpose of a given tag.

Next, the utility of tags meant for sharing (i.e. public consumption) would be maximised if the resources being tagged were associated with more specific concepts and had well-defined vocabularies. In our work, we found that the SVM classifier performed better for such tags (e.g. the food-related tags) than for others. In addition, this was more important than whether a tag was intrinsic or extrinsic – in the list of the ten worst performing tags, there was only one intrinsic tag. Related to this, better guidelines for tag creation could be provided by social tagging systems, although this appears to go against the spirit of free keyword assignment. Nevertheless, our analysis of a poorly performing tag ("photography") illustrates our reason for this

recommendation. For example, access to documents related to this concept could have been better if users did not miss out on creating obviously useful tags, such as "photography"! Here, a semi-automated tagging approach may be envisioned in which the system analyses a resource such as a webpage and suggests possible tags, but leaves the user the freedom to make his or her own selections.

Finally, our findings also suggest the likely existence of tag spamming, where tag creators deliberately assign common, popular but unrelated tags to a web resource in order to draw traffic to it. Here, spam filtering and reputation mechanisms could be incorporated into a social tagging system to combat this phenomenon.

This is ongoing work, and there were some limitations to the research reported here that may be addressed in future work. For example, the documents in our dataset were restricted to HTML content, but social tagging systems such as del.icio.us provide access to a variety of other formats such as PDF and Microsoft Word. Therefore, a logical extension would be to expand the number of document formats supported. In addition, it would also be worthwhile to perform similar analyses on other media types such images and video, given the popularity of media sharing sites such as Flickr and YouTube.

Next, in our study, one classification experiment was run using terms and tags appearing in our document dataset as features for the SVM classifier. However, further work could be conducted on an expanded feature set using other associated metadata (e.g. descriptions and comments), together with different weighting schemes for tags. For example, because the number of tags associated with a document are much fewer than the document terms, higher weights could be assigned to each tag as compared to a document term.

Further, in order to obtain sufficient documents, the present study used only popular tags, but the number of such tags is proportionately smaller than the entire collection of tags in del.icio.us. Future work could utilise a wider variety of tags to determine if performance may be affected. For example, less popular tags may be associated with more esoteric, but more specific concepts and therefore could result in better classifier performance.

Finally, our study utilised objective measures (i.e. accuracy, precision, recall and F1 scores) to determine the effectiveness of the tags. Since tagging is not an individual process of categorisation but is in effect a social process of indexing, knowledge creation and resource sharing involving many users (Sen *et al.*, 2006), it would be worthwhile to consider users' perceptions in the measurements as well. For example, future research should look into complementing the objective measures with subjective measures such as users' perceived usefulness of tags.

From a research standpoint, we have extended our understanding of social classification and its utility in sharing and accessing resources. We argue that there is a need to distinguish between the motives (i.e. personal consumption versus public consumption) behind tagging. Specifically, our findings suggest that this motivational force behind the tagging process is important and has immense impact on the utility of a tag. From a practice standpoint, the findings from this research have important implications on collaboration in the workplace, in addition to general access to web documents. Current enterprise content management tools are not effective in managing conceptual enterprise information such as those related to competitive intelligence (McGillicuddy, 2006). Social tagging, however, allows enterprises to apply metadata to

conceptual enterprise information and ultimately facilitates the managing and exchanging of conceptual information. Hence, effective tagging mechanisms are likely to benefit businesses in terms of managing and organising their web-based resources.

References

Ames, M. and Naaman, M. (2007), "Why we tag: motivations for annotation in mobile and online media", *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, New York, NY, pp. 971-80.

Angus, E., Thelwall, M. and Stuart, D. (2008), "General patterns of tag usage among university groups in Flickr", *Online Information Review*, Vol. 32 No. 2, pp. 89-101.

Berendt, B. and Hanser, C. (2007), "Tags are not metadata, but just more content – to some people", *Proceedings of the International Conference on Weblogs and Social Media*, available at: www.icwsm.org/papers/paper12.html (accessed 9 June 2008).

Bowker, G.C. and Star, S.L. (1999), *Sorting Things Out: Classification and Its Consequences*, MIT Press, Cambridge, MA.

Brooks, C.H. and Montanez, N. (2006), "Improved annotation of the blogosphere via autotagging and hierarchical clustering", *WWW2006: Proceedings of the 15th International Conference on World Wide Web*, ACM Press, New York, NY, pp. 625-32.

Chua, A. (2003), "Knowledge sharing: a game people play", *Aslib Proceedings*, Vol. 55 No. 3, pp. 117-29.

Farooq, U., Kannampallil, T.G., Song, Y., Farooq, U., Ganoe, C.H., Carroll, J.M. and Giles, L. (2007), "Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics", *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, ACM Press, New York, NY, pp. 351-60.

Golder, S.A. and Huberman, B.A. (2006), "Usage patterns of collaborative tagging systems", *Journal of Information Science*, Vol. 32 No. 2, pp. 198-208.

Hammond, T., Hannay, T., Lund, B. and Scott, J. (2005), "Social bookmarking tools (I): a general review", *D-Lib Magazine*, Vol. 11 No. 4, available at: http://dx.doi.org/10.1045/april2005-hammond (accessed 2 June 2008).

Hotho, A., Jäschke, R., Schmitz, C. and Stumme, G. (2006), "Information retrieval in folksonomies: search and ranking", *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006*, Springer, Heidelberg, pp. 411-42.

Joachims, T. (1998), "Text categorization with support vector machines: learning with many relevant features", *Proceedings of the 10th European Conference on Machine Learning*, Springer, Berlin, pp. 137-42.

Kipp, M.E. (2006), "Exploring the context of user, creator and intermediate tagging", *Proceedings of ASISandT 2006 Information Architecture Summit*, available at: www.iasummit.org/2006/files/109_Presentation_Desc.pdf (accessed 14 March 2008).

Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P. and Garcia-Molina, H. (2007), "Combating spam in tagging systems", *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, ACM Press, New York, NY, pp. 57-64.

Lakoff, G. (1990), *Women, Fire, and Dangerous Things*, University of Chicago Press, Chicago, IL.

Levy, M. and Sandler, M. (2007), "A semantic space for music derived from social tags", *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, available at: http://ismir2007.ismir.net/proceedings/ISMIR2007_p411_levy.pdf (accessed 14 May 2008).

Li, R., Bao, S., Fei, B., Su, Z. and Yu, Y. (2007), "Towards effective browsing of large scale social annotations", *Proceedings of the 16th International Conference on World Wide Web*, ACM Press, New York, NY, pp. 943-52.

Lin, X., Beaudoin, J.E., Bui, Y. and Desai, K. (2006), "Exploring characteristics of social classification", *Proceedings of the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, available at: http://dlist.sir.arizona.edu/1790/ (accessed 14 May 2008).

Macgregor, G. and McCulloch, E. (2006), "Collaborative tagging as a knowledge organisation and resource discovery tool", *Library Review*, Vol. 55 No. 5, pp. 291-300.

McGillicuddy, S. (2006), "Social bookmarking: pushing collaboration to the edge", *Tech Target*, 21 June, available at: http://searchcio.techtarget.com/news/article/0,289142,sid182_gci1195182,00.html (accessed 14 March 2008).

Marlow, C., Naaman, M., Boyd, D. and Davis, M. (2006), "HT06, tagging paper, taxonomy, Flickr, academic article, to read", *Proceedings of the 17th Conference on Hypertext and Hypermedia*, ACM Press, New York, NY, pp. 31-9.

Morville, P. (2005), *Ambient Findability*, O'Reilly Media, Sebastopol, CA.

Puspitasari, F., Lim, E-P., Goh, D.H., Chang, C-H., Zhang, J., Sun, A., Theng, Y-L., Chatterjea, K. and Li, Y. (2007), "Social navigation in digital libraries by bookmarking", in Goh, D.H., Cao, T., Sølvberg, I. and Rasmussen, E.M. (Eds), *Proceedings of the 10th International Conference on Asian Digital Libraries, Lecture Notes in Computer Science 4822*, Springer, Berlin, pp. 297-306.

Razikin, K., Goh, D.H., Chua, A.Y.K. and Lee, C.S. (2008), "Can social tags help you find what you want?", *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science 5173*, Springer, Berlin, pp. 50-61.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys*, Vol. 34 No. 1, pp. 1-47.

Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, M.F. and Riedl, J. (2006), "Tagging, communities, vocabulary, evolution", *Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work*, ACM Press, New York, NY, pp. 181-90.

Sun, A., Suryanto, M.A. and Liu, Y. (2007), "Blog classification using tags: an empirical study", in Goh, D.H., Cao, T., Sølvberg, I. and Rasmussen, E.M. (Eds), *Proceedings of the 10th International Conference on Asian Digital Libraries, Lecture Notes in Computer Science 4822*, Springer, Berlin, pp. 307-16.

Yanbe, Y., Jatowt, A., Nakamura, S. and Tanaka, K. (2007), "Can social bookmarking enhance search in the web?", *Proceedings of the 2007 Conference on Digital Libraries*, ACM Press, New York, NY, pp. 107-16.

About the authors

Dion Hoe-Lian Goh is Associate Professor at the School of Communication and Information, Nanyang Technological University. He is also Director of the MSc in Information Systems. Dion Hoe-Lian Goh's research interests are in portal and digital library applications, information retrieval, web and text mining, and the use of information technology in education. Dion Hoe-Lian Goh is the corresponding author and can be contacted at: ashlgoh@ntu.edu.sg

Alton Chua is Assistant Professor at Nanyang Technological University (NTU). He teaches in the Master of Science (Information Systems) and Master of Science (Knowledge Management) programmes. His research interests lie in information and knowledge management, and

communities of practice. Besides having published in journals such as *Journal of the American Society for Information Science and Technology*, *Journal of Information Science* and *Journal of Knowledge Management*, he is currently on the editorial board of two refereed journals, and is a member of the expert panel of the Civil Service College (Singapore).

Chei Sian Lee is Assistant Professor at the School of Communication and Information, Nanyang Technological University. Her broad research interests include computer-mediated information, the organisational and social impacts of information systems, and organisational issues of social computing. She teaches in the Information Systems and Knowledge Management programmes at NTU.

Khasfariyati Razikin is a Project Officer with Nanyang Technological University. She is also pursuing her Master of Science (Information Systems) degree in the same university. Her current research interests are in social information retrieval, usability engineering, data mining and machine learning.