Social tags for resource discovery : a comparison between machine learning and user-centric approaches

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

2011

Khasfariyati, R., Goh, D. H. L., Chua, A. Y. K., & Lee, C. S. (2011). Social tags for resource discovery: a comparison between machine learning and user-centric approaches. Journal of Information Science, 37(4), 391–404.

https://hdl.handle.net/10356/94238

https://doi.org/10.1177/0165551511408847

Social tags for resource discovery: a comparison between machine learning and user-centric approaches

Khasfariyati Razikin, Dion H. Goh, Alton Y. K. Chua and Chei Sian Lee
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

Abstract

The objective of this paper is to investigate the effectiveness of tags in facilitating resource discovery through machine learning and user-centric approaches. Drawing our dataset from a popular social tagging system, Delicious, we conducted six text categorization experiments using the top 100 frequently occurring tags. We also conducted a human evaluation experiment to manually evaluate the relevance of some 2000 documents related to these tags. The results from the text categorization experiments suggest that not all tags are useful for content discovery regardless of the tag weighting schemes. Moreover, there were cases where the evaluators did not perform as well as the classifiers, especially when there was a lack of cues in the documents for them to ascertain the relationship with the tag assigned. This paper discusses three implications arising from the findings and suggests a number of directions for further research.

Keywords

human evaluation; social computing; social tagging; text categorization; Web 2.0

1. Introduction

With the increasing popularity of social computing or Web 2.0-based applications, users are now empowered to create, publish and share resources on the web. The resulting information explosion demands new methods for seeking and organizing these resources. Social tagging or bookmarking is one possibility that offers a new avenue for resource discovery apart from search engines and web subject directories. It has a variety of uses and has been increasingly employed as a means to share resources among users [1, 2]. Tags function as both resource organizers and discoverers. As resource organizers, tags allow tag creators to annotate and categorize a resource that would be easily retrieved later. Tag consumers use the same tags to locate that resource. As resource discoverers, tags can be used as a means to tap into tag creators' collective intelligence to make serendipitous discoveries of additional relevant resources. Further, tag consumers can find like-minded tag creators, potentially leading to the creation of social networks [3]. Some examples of popular social tagging systems include Delicious, Flickr, YouTube, Cite-U-Like and Last.fm.

Social tagging as a means of resource organization has been compared favourably against conventional methods of categorization based on taxonomies, controlled vocabularies, faceted classification and ontology. Creating systems utilizing the latter methods requires experts with domain knowledge, and often adds to the costs of implementing them. Conventional categorization methods require rules to make their classification schemes work [4]. A complex rule system contributes to possible maintenance and accessibility issues. On the other hand, classifications done by ordinary people are driven by tacit knowledge that is dependent on a person's language and culture [5]. Based on this argument, a conventional system suffers from the lack of precision as it may not be able to supply relevant information needed by a user [2]. Social tagging systems, in contrast, make use of the knowledge from a (possibly large) community of users instead of relying on (a few) experts. These systems have a flattened hierarchy [6], doing away with the need for defining classes and subclasses. Such systems also do not have prescribed rules to define tags that should be used with a resource. Taken

Corresponding author:
Alton Y. K. Chua, 31 Nanyang Link, WKWSCI Building, Singapore 637718.
Email: altonchua@ntu.edu.sg

together, the knowledge harnessed from the community of users should be a reliable means of discovering resources when compared with the knowledge that comes from an individual [7].

However, social tagging is not without its sceptics. The lack of a controlled vocabulary [3] stemming from the freely assigned keywords contributes to the problem of vocabulary mismatch [8]. The mismatch in vocabulary between the tag creators and tag consumers is caused by the inherent polysemic and synonymous aspects of natural language. Also, the selection of tags by the tag creator might be motivated by their own self-serving agenda [9]. This may lead to spamming [10] where tag creators with malicious intent steer tag consumers to sites which have no relation with the tag assigned. Altogether, these factors may impede the value of tags as an effective means of sharing and organizing resources.

In spite of these shortcomings, social tagging continues to grow in popularity. This has given rise to an emerging body of research that explores tags' effectiveness for resource organization. Here, we define tag effectiveness as how accurately a tag is associated with the contents of a document. This definition is consistent with prior work such as Sebastiani [11] and Lewis [12]. Machine learning techniques for text categorization have been applied to study the extent to which tags can be used to classify blogs [13] and web resources in Delicious [14–16]. From a user's perspective, research has also been conducted on the motivations behind tagging [1, 17], the effect of experience on tagging behaviour [18], and on tagging dynamics and usage [19].

Extending extant literature, the objective of this work is to investigate the effectiveness of tags in facilitating resource discovery through two approaches, namely, automatic text categorization and human tagging behaviour. Machine learning techniques for document tagging rely on generalized mathematical models built iteratively from large datasets [11] while humans create tags for a variety of reasons on the basis of a complex mix of factors that includes culture, personality, and demographics [5]. Drawing our dataset from Delicious, we examined the extent to which tags serve as effective links to their associated web pages (henceforth known simply as documents). Defining an effective tag as one which has high precision and recall [20], we adopted text categorization techniques similar to other studies on tag effectiveness. We also conducted a human evaluation experiment where the relevance of the documents related to the tags was manually evaluated. The results from both the text categorization and human evaluation experiments were then compared.

Although our present work shares similar goals with existing research on investigating the effectiveness of tags for resource discovery, we complement and extend such work in the following ways. First, the data set used is more encompassing than previous works as we cover not only scholarly articles [21, 22] and blogs [23 13], but also other Web content, such as online news articles and product sites. Second, this paper represents a response to the call to investigate tag effectiveness that has been highlighted in previous studies such as Heymann et al. [15] and Morrison [24]. Studying effectiveness would determine if tags do help users find information that meets their needs since tags could be created for a multitude of reasons [1] other than for information discovery or retrieval. Finally, unlike previous work [e.g. 13, 14] that implicitly assumes that the results from the machine learning approaches are suitable estimates of users' tagging performance, we verify our results by conducting a user-centric experiment to determine if human tagging behaviour is similar to the outcomes suggested by text categorization. The rationale for comparing human tagging behaviour and automatic text categorization is that it has important implications in the use and design of robust social tagging systems, that is, systems that both help users in assigning tags meaningful to the community, and to discover new resources via tags by the community.

The remainder of this paper is organized as follows. Section 2 highlights studies related to this research while the methodology adopted for our present research is described in Section 3. Section 4 describes the results of the studies conducted. Section 5 discusses the implications of our results, and Section 6 closes with recommendations for social tagging systems and opportunities for future work.

2. Related work

The growth in the usage of social tagging has contributed to the increasing amount of research done in this area from different perspectives. These include the architecture and implementation of social tagging systems [25, 26], usage patterns in these systems [6, 3, 27], visualization of tags [28–30], user interfaces in tagging systems [31, 32], and the use of tags in search systems [19, 33, 34], among others. Social tagging work also shares similarities with research on the Semantic Web, and in particular, the tags, which have attached meanings, could be used for organization and ranking [35]. Other studies have empirically examined relationships among tags using different approaches (e.g. clustering) [36] as well as identifying suitable measures such as co-occurrences, cosine similarity and Jaccard similarity [37, 38]. The diversity of literature necessitates that we focus our review on related work that investigates the effectiveness of tags as a means for organizing and discovering resources.

2.1. Tag effectiveness from the machine learning perspective

Tag effectiveness has been studied using different machine learning approaches. Pioneering work done on automatic text categorization in social tagging systems was conducted by Brooks and Montanez [23] in the blogosphere. In their study, 350 popular tags from Technorati and 250 of the most recent articles from the collected tags were used. Using TF-IDF to cluster documents and pairwise cosine similarity to measure the similarity of all articles in each cluster, they found that tags categorized articles in the broad sense, and not as effective in indicating the specific content of an article. Rather than individual postings, Sun et al. [13] concentrated their efforts on classifying whole blogs with tags. Their aim was to determine if tags were effective in classifying blogs and, at the same time, investigate the usefulness of including tags as part of the feature set used in classification. This study mined 52,709 blogs and 161 tags from BlogFlux and adopted automatic text categorization using Support Vector Machines (SVM). They compared the classification results of blogs based on tags only, tags and the description of blogs, and descriptions only. It was found that tags and descriptions yielded the best classification results in terms of precision, recall and *F*-measure values, and tags alone were a more effective classifying feature than blog descriptions alone. In short, the results suggest that tags can help users find relevant information.

Apart from blogs, a study of tag effectiveness for content discovery was conducted on a dataset drawn from Delicious [14]. Here, the corpus consisted of 100 tags with 20,210 documents. A text categorization approach using SVM was adopted to determine the tag effectiveness. The study examines the effect of different tag weighting scheme to the classifiers' performance through precision, recall and *F*-measure. Their results indicate that only some tags were useful for content discovery. Next, Heymann et al. [15] employed a machine learning approach to predict tags' association with a resource. The study used 100 tags from Delicious with 60,000 pages using features such as page text (terms), anchor text and text surrounding the anchor. Their findings suggest that the page text provides more informative content compared with other features. Finally, Zubiaga et al. [16] investigated the use of social annotations for classifying web documents. The authors compared the classification performance between tags and comments. Of the two approaches, tags gave the best performance, which suggests the effectiveness of tags over other annotation types. These studies have highlighted the following findings that have laid the foundations of our present study: (1) the usefulness of tags in resource discovery, (2) the effectiveness of tags with respect with the features of the documents, and (3) the measures that illustrate the effectiveness of tags.

2.2. Tag effectiveness from the user-centric perspective

From a user's perspective, research has also been conducted to compare tags with controlled vocabularies to determine their potential for describing and discovering relevant content. For example, Lin et al. [21] compared tags from Connotea and Medical Subject Heading (MeSH) terms and found that there was only 11% similarity between MeSH terms and tags. This was because MeSH terms functioned as resource descriptors while tags focused on areas that are of interest to users. Related to the previous study, Kipp [22] compared author supplied tags from Cite-U-Like and indexing terms from INSPEC and Library Literature to determine the usage overlap. Results showed that approximately 21% of the tags were the same as the indexing terms. The reason for the deviation was attributed to the different emphases placed on an article by these tag creators and expert indexers. For example, tag creators may consider time management information (e.g. 'to do', 'to read', 'maybe') to be important tags for articles to indicate a desire to read them in the future. However, such information will be disregarded by expert indexers. Similarly, Lu et al. [39] compared user-created tags from LibraryThing with Library of Congress Subject Headings (LCSH) terms assigned by experts. Their study showed that 50.1% of the terms in LCSH were found in LibraryThing tags. Conversely, the overlapping terms constitute only 2.2% of the tag vocabulary of LibraryThing users. Taken together, these findings suggest that expert indexers' terms and tag creators employ different vocabularies, but both could complement in the search process.

In another user-centric approach, participants were asked to evaluate the relevance of the search results of their queries [24]. This shootout style study compared the performance of search engines, web directories and social tagging systems. Additionally, the overlap of the results was examined in order to illustrate a better picture on the similarity of the results. Search engines were the best of the three methods. Interestingly, social tagging outperforms Web directories. In particular, social tagging systems performed better in cases such as searching for news. In contrast, it performed poorly when one is looking for an exact site, and for a short and factual answer. This investigation reveals the tags were effective in discovering content depending on the information need of the tag consumer.

Ding et al. [40] have attempted to bridge users' social tagging behaviour with ontology. The goal of their study was to model users' tagging behaviour such that it could be leveraged with data from other social computing applications. The study had used ontology, named Upper Tag Ontology, as a model for social tagging data. The upper level ontology was induced by analysing data from three different social tagging networks. The authors found that the tagging behaviour of users and the types of tags used in the different social tagging networks are dependent on the roles of tags in the system. For instance, tags used in Delicious were related to the content of the resource, in contrast to those used in Flickr where tags describe the features of the photos. The roles of tags in YouTube were found to be a combination of both Delicious and Flickr tag types, in addition to tags being used to describe the genre of the resource. The above studies have shed light on the effectiveness of tags in resource discovery from the users' perspective in the following ways: (1) the vocabularies used by taggers are different from experts, (2) tags are effective in finding certain information, and (3) users employ different types of tags for different types of social tagging systems, and this contributes to the diversity of tags.

3. Methodology

The dataset for the present study was obtained from Delicious. Adopting a similar approach to Brooks and Montanez [23] and Heymann et al. [15], we harvested tags from the Delicious' list of popular tags so that we were assured of a sufficiently large amount of web documents available for analysis. In line with our objectives, we conducted two sets of experiments involving machine learning categorization and human evaluation. First, we conducted six machine learning experiments to determine the effectiveness of tags using text categorization techniques. As they may not reflect actual human tagging behaviours, which are in reality more complex than can be modelled by algorithms, a second experiment involving actual users tagging documents was conducted to corroborate results from the text categorization experiments. The comparison between the machine learning experiments and the human evaluation experiments will uncover the differences between text categorization techniques and the user-centric approach to tagging. The methods adopted for the different approaches are elaborated in the following sections.

Using a customized web crawler which identified the top 100 tags on Delicious, we downloaded the tags and their associated documents. All the tags that were harvested consisted of single terms (e.g. 'java', 'economics' and 'interesting'). Each tag had an average of 1331 documents, while each document had an average of 6.66 tags. In the dataset, one tag was assigned to 3617 documents, while 65 tags were the largest number of tags assigned to a document. Figure 1 shows the distribution of the tags for the number of documents and it clearly demonstrates the power law distribution of tags.

3.1. Text categorization experiments

Six text categorization experiments were conducted using the SVMlight [41] package, with each experiment employing the top 100 frequently-occurring tags associated to some 20,000 documents from the dataset. Such an approach is similar to those adopted by other related work [e.g. 15]. As binary classifiers were used, 100 classifiers were trained, each corresponding to a tag. For instance, one of the tags harvested was 'java' and a binary classifier was trained for it.

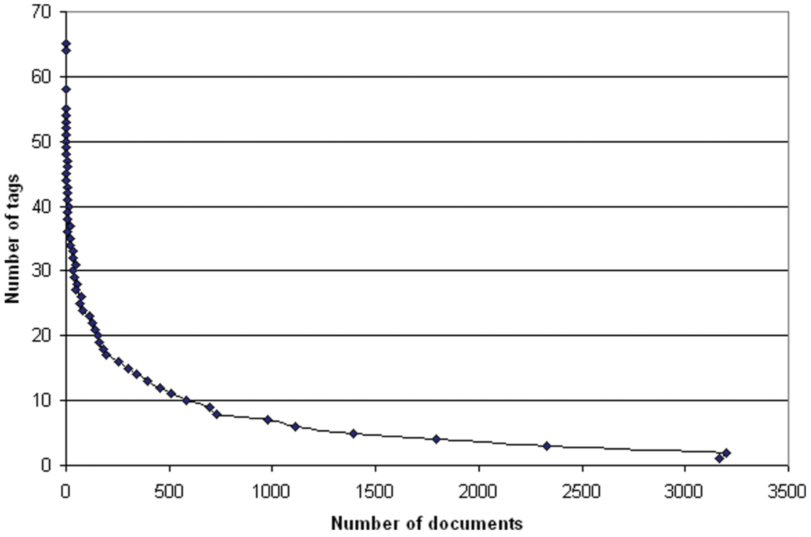


Figure 1. The distribution of tags over the number of documents.

Terms found in the content of the documents were used as features in these experiments. The feature vector provides the similarity of the documents annotated with the tag. The collected documents were first processed by removing the HTML, JavaScript and Cascading Style Sheets elements. Stopwords, such as 'is', 'them' and 'if', were then removed using the list provided as part of the Terrier IR package (http://terrier.org). The remaining words were stemmed using Porter's stemmer [42] that was also provided as part of Terrier IR package. TF-IDF values of the terms, computed using the standard formula proposed by Salton and Buckley [43], were used as feature vector for the SVM classifiers.

For each tag, we selected all the documents that were annotated with it, and these were grouped as the positive samples for the particular tag. Put differently, this meant that, on average, there were 1331 positive samples assigned to each tag in our dataset. Next, the negative samples were made up of documents which were not annotated with the tag. Here, we set the number of negative samples to be a third of the total number of positive samples. This meant that there was an average of 665 negative samples per tag that were used for the experiments. Following this, from this set of positive and negative samples, two-thirds of the documents were used for training while the rest were part of the test set, similar to the steps taken in Razikin et al. [14]. Note that the proportions of positive and negative samples were maintained. Using the same example as before, the documents that were tagged with 'java' became the positive samples for the classifier, while other documents that not tagged with 'java' were used as negative samples.

The first experiment used only the terms appearing in the document as the feature vectors. The results obtained from this experiment provided the baseline results for this remainder of this work. The five subsequent experiments were similar to the first except that they included tags, in addition to terms, as part of the feature vectors. Additionally, the TF-IDF values of the tags for the five experiments increase by one, three, five, seven and 10 times. The TF-IDF values are increased with the intention of emphasizing the importance of the tags over the other features [43, 44]. The output from the classifiers, namely precision, recall and *F*-measure values, were used to determine the effectiveness of tags in our study.

3.2. Human evaluation experiment

For this experiment, we adopted the same approach commonly used in creating large text collection activities in text categorization tasks [e.g. 45]. Five expert evaluators took part in an experiment that was conducted with the aim of comparing the results of the text categorization experiments. The evaluators held at least Master's degrees, were trained in the field of library and information science, and were familiar with the concept of tagging. They were instructed to categorize documents that had been drawn from the same dataset as the text categorization experiments. To do so, they were asked to judge the document's relevance with respect to the tag. Instructions given to the evaluators required them to apprise if the documents were good examples of the tags. That is, the goal was to ascertain if users in general could retrieve information using the tags. Providing such instructions helped the evaluators frame their judgments as tags could have been created for multiple purposes. The same 100 tags that were used in the earlier text categorization experiments were used in this experiment (for example, the 'java' tag as mentioned in the previous section).

For each of the 100 tags, 10 documents that had been assigned to the tag by the original tag creators were drawn. An equal number of documents that had not been annotated with the tag were drawn as well. In this way, our dataset comprised both positive and negative samples associated to each tag. Thus, each of the evaluators was assigned with 2000 documents from 20 tags, which they had to judge the documents' relevance to the respective tags. As human judgment is subjective to each individual, having another evaluator to assess the documents' relevance would reduce the variance in results [46]. Thus, the documents in this experiment were judge by two evaluators. For example, given the tag 'java', two evaluators had to determine which of the 20 documents that were provided were actually associated with it. The documents were judged with respect to the tag's relevance on a scale of between 1 and 4, where 1 indicated that the document was not relevant at all, and 4 indicated that it was very relevant. The relevance scores were then collapsed to binary values in order to draw comparisons with the machine learning experiments. The reliability of the evaluation was checked using Cohen's kappa, and the value obtained was 0.94, which indicated a good agreement between the evaluators [47].

The outcome of their evaluation was then compared with the tags assignments by the original tag creators'. That is the assignments made by the actual Delicious users. For our analysis, we computed the precision, recall and *F*-measure. This was obtained by comparing the relevance scores produced by the evaluators and the original tag creators' assignment. Precision was based on the number of correct assignments divided by the number of assignments by the evaluator [Equation (1)]. Here, correct assignment is defined as the documents that had been assigned with the tag by both tag creator and evaluator. Recall was based on the number of correct assignments divided by the number of assignments by the tag creator [Equation (2)].

$$\text{Precision} = \frac{\text{No. of correct assignments}}{\text{No. of assignments by evaluator}} \tag{1}$$

$$\text{Recall} = \frac{\text{No. of correct assignments}}{\text{No. of assignments by tag creator}} \tag{2}$$

For both text categorization and human evaluation experiments, the same dataset of tags and documents were used. This is done so as to ensure consistency and make for a fair comparison between both experiments. At the same time, the approaches used by the two experiments were similar. First, the multiple-level relevance scores from the human evaluators were collapsed to a binary relevance score to make them compatible with the machine learning experiments [45]. Second, the evaluation measures for the human evaluation experiment were the same as that for the machine learning experiment [e.g. 48]. The only difference between the two was the number of documents that were evaluated. Specifically, the machine learning algorithm made use of approximately 20,000 documents. However, this number made it infeasible for the human evaluators to evaluate the dataset in its entirety. Thus, each evaluator was assigned a subset of 2000 documents drawn from the 20,000 documents, with the proportions of positive and negative samples preserved. Such an approach is consistent with prior work [45]. Finally, in order to mitigate the idiosyncrasies and variances associated with human judgment, Cohen's kappa was used to assess the evaluators' judgment. This value is an indicator for reliability of the evaluators' judgment.

3.3. Content analysis

In order to delve deeper into the comparison between the text categorization experiments and the human evaluation experiment, content analysis was done on a set of selected tags. The purpose of this analysis was to shed light on and compare the process adopted by the two approaches which will lead to a better understanding of human tagging behaviour. The tags were selected based on the defined dimensions of subjective-objective and high-low calibre.

The subjective-object dimension is employed here as it is used to characterize tags, and therefore it would be instructive to ascertain its influence on the outcome of our study. Specifically, subjective tags refer to terms which could be adjectives or verbs and have meanings in relation to the documents that only the tag creator can understand [6, 49]. These tags could describe features of the resource based on the tag creator's intent, make reference to the tag creator or deal with a future action to be taken with the resource. Conversely, objective tags refer to terms which are nouns and have well-understood meanings accepted by others [20, 49]. These tags describe the content of the resource, specify the context of the resource, state the owner of the resource and/or improve upon other tags that are associated with the resource. The subjective–objective dimension is used as it is an important characteristic of social tags. Prior studies [e.g. 15] have found objective tags to be of more value than subjective tags. However, we argue that subjective tags are equally as important as these tags provide socially constructed meanings [6, 22] for other users in resource discovery. These tags express the sentiments of the tag creator and are included here because the tags are assigned to a variety of documents in Delicious.

The high calibre–low calibre dimension reflects the performance of the tags in the terms-only text categorization experiment. This experiment is selected as a basis for comparison as it had achieved the best performance among all the text categorization experiments (to be described in the following section). The high calibre tags are those which had obtained an *F*-measure value that is above the median split of the results. In contrast, low calibre tags had *F*-measure values lesser than the median split from the same set of results. *F*-measure values were used for comparison as it is the harmonic mean of precision and recall values. In other words, *F*-measure provides an overall view of the effectiveness of the tags for facilitating resource discovery.

4. Results

4.1. Text categorization experiment

The average values for precision, recall and *F*-measure of all the categorization experiments are shown in Table 1. It is surprising to note that the experiment whose feature vector consisted only of terms obtained the highest scores for the majority of the measures. A general downward trend is observed for the majority of the values as the weight of the tags increased. The precision values would be the only exception. It appears that the addition of tags as feature vectors and increasing their weights do not help in the improvement for identifying documents associated with the tags. Figure 2 illustrates the trends for the values among all experiments respectively.

Table 1. Values for precision, recall and *F*-measure from the text categorization experiments with the bold values indicating the highest values obtained

Experiments	Precision (SD)	Recall (SD)	*F*-measure (SD)
Terms only	**52.66** (4.21)	**54.86** (19.05)	**52.05** (10.99)
Tag weight 1	52.56 (6.06)	51.60 (20.75)	50.10 (13.21)
Tag weight 3	52.67 (7.38)	44.04 (2.94)	44.96 (15.47)
Tag weight 5	52.01 (10.15)	37.93 (24.90)	39.53 (18.83)
Tag weight 7	51.13 (13.72)	34.76 (6.26)	36.23 (21.14)
Tag weight 10	51.23 (16.27)	32.38 (27.08)	33.60 (22.69)

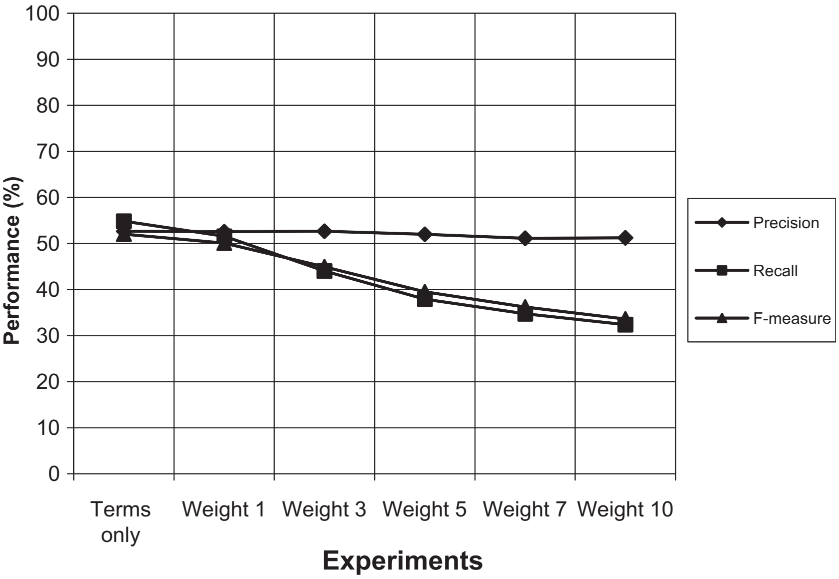


Figure 2. Performance for the text categorization experiments.

The precision values obtained were surprisingly almost constant despite increasing the tags' weights. The increase in precision for the tag weights of three and 10 are relatively insignificant with respect to the rest of the precision values obtained, implying that increasing tags' weights does not seem to help improve the classifier's precision. The recall metric values show that increasing the weight of the tags does not improve the number of relevant documents that are correctly classified. On the contrary, documents that are associated with the tag are observed to be increasingly misclassified. A similar trend is observed for *F*-measure values.

4.2. Human evaluation experiment

The average values for precision ($M = 69.79$, SD = 34.26) was above 50%. In contrast, the average value for recall was 41.60% (SD = 33.45) and the *F*-measure was at 46.12% (SD = 29.73). Precision values suggest that tags do help users in resource discovery. That is, the precision value indicates that users would be able to find relevant documents associated with the tag about 70% of the time. On the other hand, the recall value show that the evaluators felt only about 42% of the documents that had been tagged constituted relevant documents. *F*-measure, which is the harmonic mean between precision and recall, illustrates the overall effectiveness of the tags. This implies that the tags would return relevant documents 46% of the time.

We next compared the values obtained from the machine learning and human evaluation experiments (Figure 3). To reiterate, the terms-only experiment was selected as the machine learning representative as it garnered the highest values for all three measures from all the six machine learning experiments. *t*-Tests were conducted to compare the differences between the two approaches. Precision values for human evaluators ($M = 69.79$, SD = 34.26) were larger than that

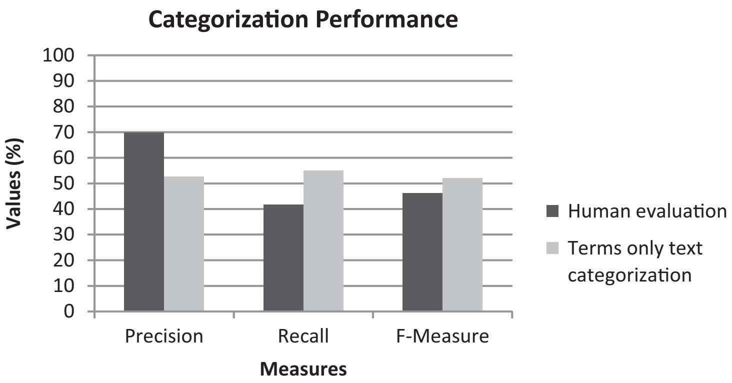


Figure 3. The performance of the human evaluation and terms-only text categorization experiments.

Table 2. Tags selected for content analysis based on their characteristics

Performance	Subjective	Objective
High calibre	interesting	library
	funny	3d
Low calibre	free	java
	re	economics

Table 3. Tags with their respective *F*-measure values from the user study and from the terms-only text categorization experiment

Performance	Subjective			Objective		
	Tag	Terms-only text categorization	Human evaluators	Tag	Terms-only text categorization	Human evaluators
High calibre	interesting	59.11	42.86	library	54.39	73.68
	funny	62.70	–	3d	63.06	16.67
Low calibre	free	53.41	60.00	java	49.72	33.33
	re	40.00	–	economics	36.54	46.15

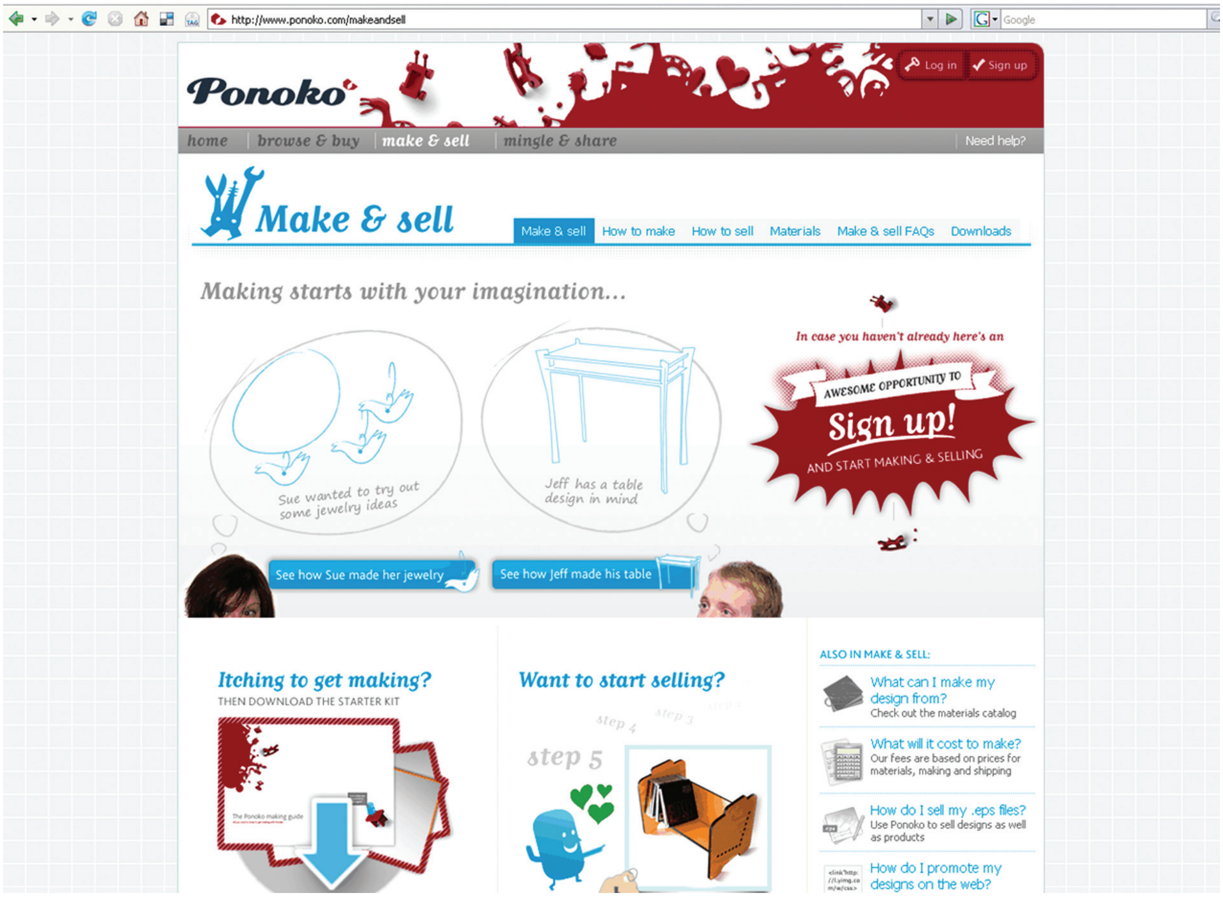
obtained for machine learning (M = 52.66, SD = 4.21), and this difference was significant [t(99) = 4.87, $p \leq 0.001$]. The difference between the recall values for machine learning (M = 54.86, SD = 19.04) was larger than those for the human evaluators (M = 41.60, SD = 33.45), [$t(99) = 3.35$, $p \leq 0.01$]. However, there were no significant differences between the F-measure values for both experiments [$t(99) = 1.81$, $p = 0.07$]. In order to uncover the reasons behind the performance in the user study, content analysis of documents associated with selected tags was performed.

4.3. Content analysis

Based on the characteristics highlighted earlier, eight tags from the terms-only experiment were randomly drawn to fill the four quadrants equally in Table 2. The table organizes the tags along two dimensions, namely subjective–objective and high calibre–low calibre. The high–low calibre characteristics are drawn based on the F-measure values the tags obtained from the terms-only text categorization experiment.

Table 3 shows the F-measure values of the tags obtained from both experiments. The empty values for 'funny' and 're' indicate that the human evaluators found none of the documents to be associated with the tags. Tags 'free', 'library' and 'economics' scored better with human evaluators. In contrast, the remaining tags, 'interesting', 'funny', '3d', 're' and 'java', did not perform well with human evaluators.

Given the variations between the values, we further analysed the tags together with the documents used in the human evaluation experiment to uncover discernible patterns. Our findings are elaborated in the following paragraphs.



Making starts with your imagination...

Sue wanted to try out some jewelry ideas

Jeff has a table design in mind

In case you haven't already here's an AWESOME OPPORTUNITY TO Sign up! AND START MAKING & SELLING

Itching to get making? THEN DOWNLOAD THE STARTER KIT

Want to start selling?

ALSO IN MAKE & SELL:

What can I make my design from? Check out the materials catalog

What will it cost to make? Our fees are based on prices for materials, making and shipping

How do I sell my .eps files? Use Ponoko to sell designs as well as products

How do I promote my designs on the web?

Figure 4. A document which has been tagged with ‘3d’ by the tag creator.

4.3.1. High-calibre, subjective tags. Both ‘interesting’ and ‘funny’ did not perform well with the human evaluators as they did not seem to agree with the association of the tag with the documents. It appears that both tags have a highly subjective interpretation with respect to the nature of the documents’ content. The *F*-measure values obtained illustrate that the human evaluators did not share the same views as the tag creators for the documents that had been annotated with these tags. For instance, the evaluators did not find any of the documents that had been associated with ‘funny’ by the tag creators to be humorous. Perhaps they were unable to appreciate the underlying humour presented in the related documents as it did not fit into their idea of the context of being funny.

However, these tags performed better than the median *F*-measure value in the text categorization experiments. It was found that the documents associated with these tags were diverse. It is found that the large number of terms found in the documents contributed to diverse terms. The heterogeneity of the terms enabled the classifier to elucidate a document’s possible association to the tag based on the terms and its TF-IDF values [14].

4.3.2. High-calibre, objective tags. ‘Library’ performed better than ‘3d’ in the human evaluation experiment as the documents were found to be relevant to the tag. The differences in the performance between the human evaluators and the text categorization algorithm could be attributed by the different approaches to categorization. As humans categorize things by implicit knowledge [5], factors like experience and intelligence could affect how one classifies an object. On the other hand, text categorization algorithms using weights (e.g. TF-IDF) examine documents in terms of their similarity to other documents in the collection. Additionally, the TF-IDF value is dependent on the occurrence of the terms in the documents.

The poor performance of ‘3d’ in the human evaluation experiment could be attributed to the fact that the majority of the evaluators found the concept to be abstract. For instance, the document shown on creating your own products (Figure 4) was tagged with ‘3d’ and the appropriateness of the tag’s relation to the document was found to be debatable by the evaluators. We can only surmise that the document was tagged with ‘3d’ for reasons only known to the actual tag creator. As the figure shows a site that allows consumers to create their own products, perhaps the page was tagged ‘3d’ as the web application enables users to create their products with a three-dimensional view.

4.3.3. Low-calibre, subjective tags. ‘Free’ performed well with the human evaluators while ‘re’ performed poorly in the same experiment. For ‘free’, despite being a subjective tag, evaluators agreed to the tags’ association with their respective documents. A possible reason for the good performance of ‘free’ is that the associated documents offered contextual cues that the evaluators could relate to the tag.

On the other hand, ‘re’ did not perform as well because its meaning is not discernable and therefore is unlikely to find common agreement in terms of usage by everyone. For example, a Web page tagged with ‘re’ might be due to its content being related to religious education. At the same time, a communication research organization page was also tagged with ‘re’. We can only guess what the tag creator’s intention was when ‘re’ was used to associate the latter document with. Here, ‘re’ could possibly mean recommendation as the document had no discernable relation to religious education, real estate or others. The evaluators perhaps needed to be anchored on a certain frame of similarity based on past knowledge. However, in cases such as these, they were not able to find such a reference point and, thus, they were unable to discern the intention of the tag creator. In contrast, the text categorization algorithm performed better when classifying these documents despite the generic tags used because the algorithm looks at the common terms among the documents in the collection.

4.3.4. Low-calibre, objective tags. Here, ‘java’ did not perform as well as ‘economics’ for the human evaluation experiment. Perhaps the poor performance of ‘java’ is attributed to its polysemic nature and evaluators were unaware of the tag’s different senses. For instance, a layperson might not know that ‘java’ could be associated with a programming language whereas a technology-oriented person would. Put differently, the technology-oriented person is able to discern the relevance of the tag’s association with the documents by relying on their experience.

‘Economics’ performed better with the human evaluators as they were aware of its meaning. Like documents associated with ‘free’, the documents associated with ‘economics’ consisted of documents that dealt with inheritance and taxes, and thus had cues that were contextually related to economics. This contributed to the better performance as the evaluators had a frame of reference for the tag, and were able to relate that reference to the documents.

5. Discussion

To reiterate, our research seeks to shed light on the effectiveness of social tags for resource discovery. Three main findings emerge from our text categorization experiments and human evaluation. First, there are important differences between the outcomes of the text categorization and human evaluation experiments. Our analysis suggests that the process adopted by the former was more consistent than that adopted by the latter. Automated text categorization largely employs mathematical processes for decision-making. In particular, decisions are made based on the mathematical properties of the features extracted from the dataset and no semantics are attached to them [11]. In contrast, the human decision-making process is significantly more complex [50] as it is dependent on many factors such as background knowledge, culture and imagination [5]. However, this complexity is a double-edged sword. Our evaluators were likely to agree on terms that have commonly agreed upon meanings. On the other hand, terms which were abstract (e.g.’3d’) or affective (e.g. ‘funny’) required the evaluators to reflect. The resolution to the reflection process is dependent on a variety of factors unique to each of the evaluators. This puts into perspective the complexity of the human cognitive process when compared with the automated categorization process. Specifically, human decision-making is subjected to intuition and biases [51], and is in contrast to the consistency of mathematical processes, where the outcome is predictable at all times.

Second, cues in documents appear to help human evaluators make better judgments than automated categorization processes. The tags that performed better in the human evaluation experiment were ‘java’, ‘economics’, ‘library’ and ‘free’. A common pattern seen between all the documents selected for this experiment was that there were cues found in the documents [52], either visually or in textual format, or both. This helped the evaluators to establish the relationship between the document and the tag. Additionally, tag consumers might also use such tags to locate the documents, signifying that these tags act as information scent [53, 54], providing signals to the tag consumers that such tags may lead them to information which meets their needs. In contrast, the documents that obtained low performance for the human evaluation experiment did not have cues that highlighted any connection between the document and the tag. For example, as discussed previously, the tag ‘re’ referred to a diversity of topics resulting in a lack of discernible patterns of relationships between documents and the tag itself. It is thus not surprising that the absence of cues for ‘re’ resulted in its poor performance in the user study.

Next, the performance of the text categorization experiments did not fare better when tags were included as part of the feature set. Coupled with the relatively large standard deviations of the performance metrics, our findings imply that tags vary in their effectiveness for discovering relevant content. This outcome can be attributed to a variety of reasons. The lack of a controlled vocabulary in social tagging could result in a proliferation of tags of varying quality [2]. Further, tag creators may have different levels of experience within the social tagging system and therefore could use a vocabulary that may not be shared by other members of the community [49]. In addition, the implicit assumption in the text categorization experiments is that tags are created for public access to content, but in reality, there are a variety of other motivations for creating tags including for self-retrieval and memory [1]. Finally, the results could also be due to overfitting of the classification model, and this could possibly be contributed by noise in the dataset.

In summary, our study has uncovered differences between machine learning and human categorization process for tagging. The consistency of machine learning process is unmarred by the variety of factors that could influence human judgments. In addition, the cues found in documents assist humans with their assessment. Finally, our findings indicate that tag creators may not be familiar with effective tagging strategies, that tag consumers did not share the same tag vocabularies with the tag creators, or simply that the tags were meant for uses other than retrieval. The differences between the machine learning and user-centric approaches call for a set of techniques that could help users in discovering new resources via social tagging. Three implications from our findings on using tags for resource discovery are therefore as follows.

First, owing to the variety of reasons that tags could be created, users need to be made aware of effective tagging techniques [16] to select tags which have relevance to the document, especially for the purposes of sharing and facilitating retrieval. For example, such users could be introduced to the different types of tags (e.g. hypernyms, synonyms, etc.) that could be used for organizing and sharing.

Next, new tagging methods could be implemented to help tag creators with the selection of their tags. The following are some possibilities:

- As most of the objective tags in our study were selected based on cues found in the documents, tools could be implemented to help tag creators select tags based on terms found in the document [e.g. 55]. This would reduce the cognitive effort needed to select appropriate tags for the document, and simultaneously ensuring quality tags that enhance resource discovery. Subjective tags, in contrast, would be recommended to be kept private as these might not meet the information needs of the tagging community.
- A tag recommendation system could include suggestions that encompass hypernyms or hyponyms that could be found by deducing semantic relationships between the tags [e.g. 36–38]. Such a system could also suggest tags which are preferred by the community of users [e.g. 49, 19]. Tag consumers will also benefit as they are likely to obtain more relevant documents that match their information needs. Of course, in the spirit of social tagging, tag creators would have the choice of overriding these recommendations and using tags that they see fit.
- In order to harness cues found in documents that are predominantly made up of non-textual multimedia content, automated content analysis should be implemented for tag suggestion. For instance at a basic level, links to media files with extensions like mp3, avi, jpeg and so on, could lead to automatic suggestions describe the media such as ‘music’, ‘video’ and ‘image’. More sophisticated processing techniques could be used to identify objects and themes in such media as well. Further, a social tagging system could harness other social computing services by requesting the tags used by the documents’ creator. For example, tags that had been used by the document creator in Flickr could be imported and recommended to a user that is tagging the same document in Delicious.
- Measures addressing the ambiguity of tags (e.g. polysemic tags) could be put in place when using such tags to search for relevant documents. For example, a social tagging system could prompt a user to select the precise context of a tag in order to retrieve documents that meet his/her needs.

Finally, tags could be rated by users in terms of usefulness in meeting information needs, potentially ‘smoothing’ out the subjectivity inherent in human decision-making. For instance, a user could rate a tag that was employed for searching based on the relevancy of the documents returned by the social tagging system. Based on these ratings, the usefulness of the tags as resource descriptors could be determined over time. Such ratings will also help in the ranking of tags for recommendations during tag creation.

6. Conclusion

Social tagging provides a new avenue for resource discovery. In this paper, we have investigated the effectiveness of tags in assisting tag consumers to discover relevant content. Two different techniques, namely text categorization and user-centric approaches were adopted in the course of our investigation. For the first technique, six text categorization experiments were conducted. The first experiment made use only of terms as feature vectors. The subsequent experiments added tags in addition to the terms as part of its feature vectors. Additionally, the weights of the tags were increased by three, five, seven and 10. The terms-only experiment yielded the best results out of the six experiments.

In addition, a human evaluation experiment was conducted to compare the performance of human tagging against machine learning categorization techniques. One of our findings showed that evaluators performed better than the automated classifiers when tags possessed objective characteristics. In contrast, the evaluators did not perform as well as the classifiers when the tags were subjective. In addition to the subjective nature, this could be due to the lack of cues, which did not help the evaluators to ascertain the relationship between the tag and the document.

The findings in the present study represent ongoing work that provides opportunities for future research. First, work could be done on establishing other dimensions of tags in addition to objective/subjective or high/low calibre characteristics [e.g. 49]. This would enable a deeper understanding on the nature of tags which could help improve tag creation and use in social tagging systems. Next, this study assumes that the harvested tags were meant for future public retrieval. However, we acknowledge that tags could be created for a multitude of reasons [1] which remain unknown to us. Future work could thus conduct experiments to uncover the actual intent of social tags from tag creators. Further, the tags that were selected for this study were the 100 most popular tags. Future work could investigate the effectiveness of the less popular tags. In addition, it would also be worthwhile to perform similar analyses on other domains such as Cite-U-Like or Amazon, where the intent [e.g. 1, 56] of the tag creators could be different from that of Delicious, as well as extending the analyses to multimedia content such as those found in YouTube and Flickr. Additionally, similar analysis could take place in social tagging systems with well-defined domains (e.g. TechNet) or those whose users are professionally homogenous (e.g. Cite-U-Like or Dogear). Extending the investigation into these areas would help with understanding the use and effectiveness of tags in relation to resource discovery. Next, the ranking of tags by performance is an important area to examine as, by doing so, we would be able to identify their characteristics that could be effective for resource discovery. Hence, another future area of work is to compare the rankings of the tags from the results of the both machine learning and human evaluation experiment. Finally, the results of this study were based on the perspective of popular tags, which are often generalized and well understood by many users. However, less popular tags could be more specialized and lead to different experimental outcomes. Hence, a study of both types of tags would be a useful future endeavour.

Acknowledgements

This work was partly funded by A*STAR grant 062 130 0057.

References

[1] Ames M, Naaman M. Why we tag: motivations for annotation in mobile and online media. In: *SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 2007, p. 971–80.

[2] Macgregor G, McCulloch E. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 2006; 55: 291–300.

[3] Marlow C, Naaman M, Boyd D, Davis M. HT06, tagging paper, taxonomy, Flickr, academic article, to read. *Seventeenth conference on hypertext and hypermedia*. New York: ACM Press, 2006, pp. 31–40.

[4] Morville P. *Ambient findability*. Beijing: O'Reilly, 2005.

[5] Lakoff G. *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago, IL: University of Chicago Press, 1987.

[6] Golder SA, Huberman BA. Usage patterns of collaborative tagging systems. *Journal of Information Science* 2006; 32: 198–208.

[7] Surowiecki J. *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday, 2004.

[8] Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human-system communication. *Communications of ACM* 1987; 30: 964–971.

[9] Chua AYK. Knowledge sharing: a game people play. *Aslib Proceedings* 2003; 55: 117–129.

[10] Koutrika G, Effendi FA, Gyöngyi Z, Heymann P, Garcia-Molina H. Combating spam in tagging systems. In: *Third international workshop on adversarial information retrieval on the web*. New York: ACM, 2007, pp. 57–64.

[11] Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys* 2002; 34: 1–47.

[12] Lewis DD. Evaluating text categorization. In: *Workshop on speech and natural language*. Association for Computational Linguistics, 1991, pp. 312–318.

[13] Sun A, Suryanto MA, Liu Y. Blog classification using tags: an empirical study. In: *International conference on Asian digital libraries 2007*. Berlin,: Springer-Verlag, 2007, pp. 307–316.

[14] Razikin K, Goh DH-L, Chua AYK, Lee CS. Can social tags help you find what you want? In: *12th European conference on research and advanced technology for digital libraries*. Berlin: Springer-Verlag, 2008, pp. 50–61.

[15] Heymann P, Ramage D, Garcia-Molina H. Social tag prediction. In: *31st annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM, 2008, pp. 531–538.

[16] Zubiaga A, Martinez R, Fresno V. Getting the most out of social annotations for web page classification. In: *9th symposium on document engineering*. New York: ACM, 2009, pp. 74–83.

[17] Nov O, Ye C. Why do people tag? Motivations for photo tagging. *Communications of ACM* 2010; 53: 128–131.

[18] Lee CS, Goh DH-L, Razikin K, Chua AYK. Tagging, sharing and the influence of personal experience. *Journal of Digital Information* 2009; 10: 1–15.

[19] Farooq U, Kannampallil TG, Song Y, Ganoe CH, Carroll JM, Giles L. Evaluating tagging behaviour in social bookmarking systems: metrics and design heuristics. In: *2007 International ACM conference on supporting group work*. New York: ACM, 2007, pp. 351–360.

[20] Goh DH-L, Lee CS, Chua AYK, Razikin K. Resource discovery through social tagging: a classification and content analytic approach. *Online Information Review* 2009; 33: 568–583.

[21] Lin X, Beaudoin JE, Bui Y, Desai K. Exploring characteristics of social classification. In: *17th workshop of the American Society for Information Science and Technology special interest group in classification research*. Austin, TX, 2006.

[22] Kipp ME. Exploring the context of user, creator and intermediate tagging. In: *ASIS&T 2006 Information Architecture Summit*. Vancouver, 2006.

[23] Brooks CH, Montanez N. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: *15th international conference on world wide web*. New York: ACM, 2006, pp. 625–632.

[24] Morrison JP. Tagging and searching: search retrieval effectivenss of folksonomies on the world wide web. *Information Processing and Management* 2008; 44: 1562–1579.

[25] Chudnov D, Barnett J, Prasad R, Wilcox M. Experiments in academic social book marking with Unalog. *Library Hi Tech* 2005; 23: 469–480.

[26] Puspitasari F, Lim E-P, Goh DH-L, et al. Social navigation in digital libraries by bookmarking. *International Conference on Asian Digital Libraries 2007*. Berlin: Springer-Verlag, 2007, pp. 297–306.

[27] Santos-Neto E, Condon D, Andrade N, Iamnitchi A, Ripeanu M. Individual and social behaviour in tagging systems. In: *20th ACM conference on hypertext and hypermedia*. New York: ACM, 2009, pp. 183–192.

[28] Sinclair J, Cardew-Hall M. The folksonomy tag cloud: when is it useful? *Journal of Information Science* 2008; 34: 15–29.

[29] Dubinko M, Kumar R, Magnani J, Novak J, Raghavan P, Tomkins A. Visualizing tags over time. In: *15th International conference on world wide web*. New York: ACM, 2006, pp. 193–202.

[30] Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. In: *International conference on multidisciplinary information sciences and technologies, InSciT2006*. Mérida, Spain, 2006.

[31] Freyne J, Brusilovsky P, Smyth B, Coyle M. Collecting community wisdom: integrating social search and social navigation. In: *IUI '07: 12th international conference on Intelligent user interfaces*.New York: ACM Press, 2007, pp. 52–61.

[32] Shiri A. An examination of social tagging interface features and functionalities. *Online Information Review* 2009; 33: 901–919.

[33] Heymann P, Koutrika G, Garcia-Molina H. Can social bookmarking improve web search? In: *WSDM '08*. New York: ACM, 2008, pp. 195–205.

[34] Yanbe Y, Jatowt A, Nakamura S, Tanaka K. Can social bookmarking enhance search in the web? In: *2007 joint conference on digital libraries*. New York: ACM, 2007, pp. 107–116.

[35] Xu Z, Fu Y, Mao J, Su D. Towards the semantic web: collaborative tag suggestions. In: *Collaborative web tagging workshop*, 2006.

[36] Specia L, Motta E. Integrating folksonomies with the semantic web. In: *Extended semantic web conference 2007*. Berlin: Springer, 2007, pp. 624–639.

[37] Cattuto C, Benz D, Hotho A, Stumme G. Semantic grounding of tag relatedness in social bookmarking systems. In: *International semantic web conference 2008*. Berlin: Springer, 2008, pp. 615–631.

[38] Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Stumme G. Evaluating similarity measures for emergent semantics of social tagging. In: *18th international conference on world wide web*. New York: ACM, 2009, pp. 641–650.

[39] Lu C, Park J- r, HX. User tags versus expert-assigned subject terms: a comparison of LibraryThing tags and Library of Congress subject headings. *Journal of Information Science* 2010; 36: 763–779.

[40] Ding Y, Jacob EK, Zhang Z, et al. Perspectives on social tagging. *Journal of the American Society for Information Science and Technology* 2009; 60: 2388–2401.

[41] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: *10th European conference on machine learning*. London: Springer-Verlag, 1998, pp. 137–142.

[42] Porter M. An algorithm for suffix stripping. *Program* 1980; 40: 211–218.

[43] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 1988; 24: 513–523.

[44] Liu Y, Loh HT, Sun A. Imbalanced text classification: a term weighting approach. *Expert Systems with Applications* 2009; 36: 690–701.

[45] Voorhees EM. Overview of the thirteenth Text Retrieval Conference (TREC 2004). In: *Thirteenth Text Retrieval Conference (TREC 2004) NIST*, 2004.

[46] Harter SP. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science and Technology* 1996; 47: 37–49.

[47] Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

[48] Brill E, Ngai G. Man vs. machine: a case study in base noun phrase learning. In: *37th annual meeting of the Association of Computational Linguistics for Computational Linguistics*.Pennsylvania: Association of Computational Linguistics, 1999, pp. 65–72.

[49] Sen S, Lam SK, Rashid AM, et al. tagging, communities, vocabulary, evolution. In: *20th anniversary conference on computer supported cooperative work*. New York: ACM, 2006, pp. 181–190.

[50] Dumais S. Data-driven approaches to information access. *Cognitive Science* 2003; 27: 491–524.

[51] Betsch C, Kunz JJ. Individual strategy preferences and decisional fit. *Journal of Behavioral Decision Making* 2008; 21: 532–555.

[52] Fu W-T. The microstructures of social tagging: a rational model. In: *CSCW '08*. New York: ACM, 2008, pp. 229–238.

[53] Pirolli P, Card S. Information foraging. *Psychological Review* 1999; 106: 643–675.

[54] Pirolli P. *Information foraging theory: adaptive interaction with information*. New York: Oxford University Press, 2007.

[55] Hong L, Chi EH, Budiu R, Pirolli P, Nelson L. SparTag.us: a low cost tagging system for foraging of Web content. In: *Working conference on advanced visual interfaces*. New York: ACM, 2008, pp. 65–72.

[56] Thom-Santelli J, Muller MJ, Millen DR. Social tagging roles: publishers, evangelists, leaders. In: *2008 SIGCHI conference on human factors in computing systems*. New York: ACM, 2008, pp. 1041–1044.