This study investigates the contribution of collaborative tagging to the design of user-driven vocabularies in knowledge management systems (KMS). Three metrics, tag growth, tag reuse, and tag discrimination, were used to examine the evolution of the tagging vocabulary of the knowledge management community of interest in CiteULike over a three-year period. Results indicate a steady decrease in the number of unique tags over the four years, suggesting an increasing stability in the community vocabulary over time and the establishment of domain-specific vocabulary. Members reused each others’ tags over time and exhibited increasingly collaborative tagging behavior. Tag discrimination was high, with 4.11 distinct articles per tag. The stable and discriminatory nature of the community’s tags suggests that collaborative tagging may serve as a useful resource for vocabulary choice or maintenance by KMS managers.

Introduction

The purpose of this study is to empirically explore collaborative tagging as a promising solution to the creation of user-driven vocabulary in the form of tags that can help in designing effective user-friendly knowledge management systems (KMS). Collaborative tagging, with the resulting set of tags, also referred to as “folksonomies”, offers an interesting alternative to the creation of metadata for KMS in enterprises. In a number of enterprises today, KMS are being recognized as “strategic enablers for value creation and organizational transformation” (Benbya, 2008). The design of KMS, the type of systems that are specifically designed with the intent to manage organizational knowledge, is identified as a core driver to the efficacy of knowledge management (KM) in enterprises (Goel & Mousavidin 2008; Alavi & Leidner 2001). The success of KMS in organizations, however, has been elusive. Despite significant investments in KMS, most enterprises have reported inefficiencies in knowledge representation, discovery, retrieval, and dissemination (Goel & Mousavidin 2008; Akhhavan et al. 2005; Hammer et al. 2004). Most users of KMS in enterprises have reported difficulty in finding and sharing knowledge (Hansen, Nohria, & Tierney, 1999; Hargadon & Sutton, 2000). Furthermore, employees using enterprise KMS have reported difficulties in remembering and retrieving items that they have previously found and thought to be interesting. This difficulty might be attributed to the assumption by KMS designers that users know what they are looking for. Evidence shows that more often than not most users “are not always exactly sure of what they are looking for but when they see it, they know what it is” (Benbya

1 The authors would like to thank those who assisted in refining this paper.
Users also differ in their use of vocabulary. As a result, it is hard for users to know the right descriptor or keyword to use that will get them what they are looking for (Shirky, 2005).

In this study we investigate the contribution of tags in creating domain specific vocabulary; in this case, the vocabulary of a knowledge management community of interest. We examine the collaborative or community value of tags, i.e., whether tagging is useful in building a community’s vocabulary. Our study differs from prior tag-related research in terms of scale, a focus on community of interest created tags, and in the use of the dynamic metrics, tag growth, tag reuse, and tag discrimination, to define tag usefulness in building community-driven vocabulary. We build and extend on previous studies by integrating metrics for analyzing how tag vocabulary evolves and investigating a more comprehensive dataset, over three years worth of CiteULike data. By focusing on a specific community of interest, using the three metrics, (i.e., tag growth, tag reuse, and tag discrimination), using a substantive dataset, and exploring the data over time, we examine and evaluate the value of tags in building community vocabulary. The study’s aim is to contribute to the understanding of how tag vocabulary evolves among users with a common interest. Such an understanding is important to KMS administrators and designers to help them gauge how tagging can aid knowledge discovery, knowledge retrieval and knowledge dissemination. Our research will aid in future developments of computational frameworks, tools, techniques and methodologies for representing and accessing the unstructured and semi-structured documents in organizational KMS.

Background

At the core of most KMS or information systems is the metadata, a set of descriptive elements essential for the organization and retrieval of objects. For enterprise KMS, the use of metadata to enhance resource accessibility has become indispensable; it involves the addition of information or descriptive elements to knowledge “resources to further describe them so that anyone” is likely to discover them in a search attempt (Wolfe, 2000). Metadata facilitates resource discovery by “disclosing sufficient information about a resource to enable users or intelligent agents to discriminate between what is relevant and what is irrelevant to a specific information need” (Macgregor & McCulloch, 2006). Thus, in most enterprise KMS, to enable the discovery and retrieval processes of knowledge objects, individuals assign keywords or descriptors to knowledge resources according to their content. But it is important to note that in most enterprise KMS there is no consistency of resource formats. The vast majority of digital enterprise knowledge assets are stored in several forms i.e., structured, semi-structured and unstructured forms. For instance, CA labs recently reported that, “about 80% of all corporate data is unstructured” (Gupta, 2008). Structured data, data having a data model or a plan that is easily usable by a computer program, can easily be handled by KMS but it is the unstructured data that provides a challenge. Unstructured data is presented in both textual and non-textual formats and includes emails, reports, spreadsheets, presentations, images, video, and audio (Inmon & Nesavich, 2008). There is a world of promise and opportunity in the knowledge buried in the unstructured data environment. However, with unstructured data, keyword searches prove to be inefficient because they are not context sensitive. What is needed is “the ability to interpret the semantic meaning of words in documents—something that keyword searches cannot do” (Nelson Mattos as cited in Red Herring, 2005).

For years a number of scholars in Library and Information Science have maintained that to ensure effective indexing and system retrieval efficiency it is necessary to apply some degree of control to the indexing process by using what is commonly referred to as “controlled vocabulary” (Cattuto et al., 2007; Golder & Huberman, 2005; Golder & Huberman, 2006; Gudivada, Raghavan, Grosky, & Kasanagottu, 1997; Macgregor & McCulloch, 2006; Shirky, 2005; Spiteri, 2007). Controlled vocabulary refers to predefined, authorised terms that have been preselected by the designer of the vocabulary e.g., the Library of Congress Subject Headings, and “used to tag units of information (document or work) so that they may
be more easily retrieved by a search” (Wikipedia, 2008). In enterprise KMS the use of controlled vocabulary ensures that the same or similar terms are assigned to describe the same or similar resources, thereby facilitating easy user discovery of relevant resources. However, some scholars have argued that controlled vocabularies are not always adequate for online resource discovery (Macgregor & McCulloch, 2006; Mai, 2004; Nicholson et al., 2001). Nicholson et al. (2001), for example, comment on the lack of specificity in the subject areas as one of the key limitations of controlled vocabularies and point out the importance of the user’s own vernacular terms (what terms users are likely to use) and the need for user interaction in the subsequent creation of useful systems vocabulary. Moreover, enterprise-specific terminology is always evolving. As a result, the use of controlled terminology rather than vernacular phrases has proven impractical for some of the content being created and used in enterprises today; especially unstructured data content (Golder & Huberman, 2005). Thus the recent advent and popularity of collaborative tagging has been heralded by some researchers as a potentially useful way of employing user-created index terms, thereby superseding or complementing the subject indexing role of experts (Macgregor & McCulloch, 2006; Shirky, 2005).

Collaborative tagging, also referred to as social tagging mechanisms, refers to the process whereby users enrich diverse resources by choosing their own index terms or tags and then associating resources with these tags “in a totally uncoordinated fashion, for their own use” (Cattuto, 2006). In Web-based applications like Delicious (http://delicious.com) and CiteULike (http://www.citeulike.org/), for example, users assign their own index terms or tags to web-based resources (Cattuto, 2006). The process allows anyone to freely attach terms or tags to content. The users do not have to own the content being tagged. The advantage is that end users do subject indexing instead of only experts, and the assigned tags are shown immediately on the Web (Voss, 2007). The resulting set of tags, also referred to as “folksonomies”, is sometimes employed as a sort of semantic map or “tag clouds” that enables users to navigate the site’s contents. Tags also have the “additional effect of grouping related URLs together” (Shirky, 2005). It is important to highlight that collaborative tagging as a method of organizing knowledge resources “is contradictory to the ethos of controlled vocabularies” (Macgregor & McCulloch, 2006, p. 292). In contrast to the creation of a controlled vocabulary, collaborative tagging permits any user to assign keywords (or “tags”) to Web content (Golder & Huberman, 2006). In this way tagging can be viewed as document subject indexing without controlled vocabulary.

Collaborative tagging has some shortcomings, however. One main problem is the tags’ lack of precision (Guy & Tonkin, 2006). As W. G. Stock observed, in contemporary popular systems that allow for tagging, different word forms, nouns in singular, nouns in plural, abbreviations, and misspelled words are all often present. There is also no control of polysemy, that is, words with multiple related meanings, and synonymy, multiple words with the same or similar meanings (Fichter, 2006; Spiteri, 2007; Stock, 2007). In addition, collaborative tagging suffers from “meta noise” limitations in the form of incorrectly spelled tags or irrelevant tags, which burdens users and decreases the system’s utility” (Wu, Zubair, & Maly, 2006, p. 112).

Collaborative tagging and the resulting folksonomies are in their infancies but exhibit a lot of potential benefits to the design of KMS. For enterprise KMS “collaborative tagging systems have the potential of becoming a technological infrastructure for harvesting enterprise knowledge” (Wu, Zubair, & Maly, 2006, p. 112). Tagging “directly reflects the vocabulary of users” and thus represents authentically the use of language inside enterprise communities (Mathes, 2004, p. 7). As Clay Shirky (2005) commented, with collaborative tagging there is an inherent kind of quality control: The more knowledge workers tag a document, the more relevance the tagged document seems to have for the enterprise and its workers. Through the use of tagging, knowledge workers can create value-added knowledge by annotating existing knowledge, providing metatags, and aggregating heterogeneous documents into named collections for future use (Macgregor & McCulloch, 2006; Mathes, 2004). Not unimportant to enterprise KMS is the fact that tagging provides a cheap method of indexing. For an enterprise KMS “tags
enable a huge amount of user-produced organizational value, at vanishingly small cost” (Shirky, 2005). A further strength of tagging is serendipity, which complements non-goal directed searching and browsing by introducing the knowledge worker to potentially invaluable resources that would otherwise have been undiscoverable (Mathes, 2004). Folksonomies also contain additional relevant information about the social context of the tags. Collaborative tagging offers promise for organizational knowledge management systems and thus warrants further investigation for the following reasons: (i) because users can annotate the resources they tag, knowledge workers can keep track of the sites they have visited as reminders of useful content for future projects; (ii) content that is shared through tagging may reveal patterns “that may be presumed to be representative of user interests and expertise - an ‘I tag, therefore I know’ indication by the user” (John & Seligmann, 2006, p. 1); (iii) users can see which other employees are using the same tags, and what sites they have visited, thereby making more visible the patterns and processes in knowledge work; (iv) the core strength of collaborative tagging to knowledge discovery depends on users’ ability to identify and connect with other users who share the same interests: when a user tags a resource he or she can see which other users have also tagged it. In this way tagging enables the discovery of shared interests and expertise within an organization and thus facilitates the identification and formation of communities of expertise or practice; and (v) by making visible user interests to all users, tagging provides a way of keeping pace with the knowledge worker’s changing interests without the need for the worker to update his or her skill profiles.

Current work on collaborative tagging indicates a huge gap between the knowledge worker’s vernacular terms and the controlled index terms found in intranet or KMS metadata (Fichter, 2006; Macgregor & McCulloch, 2006; Mathes, 2004; Millen, Feinberg, & Kerr, 2006). Our study posits that collaborative tagging aids in the representation, discovery, retrieval, and dissemination of knowledge objects in KMS. The focus of this study is not on discussing whether collaborative tagging should replace controlled vocabularies; rather the aim is to contribute to the understanding of how tag vocabulary evolves among users with a common interest. We hypothesize that by employing tagging in KMS, the gap between the knowledge worker’s vernacular terms and the controlled index terms could be reduced. We develop the following arguments:

1. As users belonging to a community tag more articles the unique tag vocabulary they each use also increases.
2. Unique tag growth across time is strongly influenced by the increase in the number of new users joining the community-of-interest.
3. Users belonging to a group or community-of-interest tend to reuse each others’ tags, thereby contributing to the building of domain-specific tags.
4. Tags assigned by users belonging to a community tend to have strong discriminating values, thereby establishing the intellectual value of tag vocabulary and usefulness to KMS designers.

The goal of this study, therefore, is to contribute to the current debate on collaborative tagging for knowledge resource indexing. As stated before, the focus of this paper is not on discussing whether collaborative tagging should replace controlled vocabularies; rather the argument is that the development and the maintenance of controlled vocabularies used in organizational KMS can profit from tagging, because tagging provides a rich source of authentic organization-specific term material.
Related Work

Very few empirical studies on the dynamics of collaborative tagging have been conducted. There are three distinct elements of interest in the world of tagging research: the tag, the item being tagged (i.e., the object) and the person doing the tagging (i.e., the user). Using four days of Delicious data, Golder and Huberman found that the frequency of tag use and what the tags themselves describe vary greatly among users. The study confirmed some measure of regularity in “user activity, tag frequencies, kinds of tags used, bursts of popularity in bookmarking and a remarkable stability in the relative proportions of tags within a given url” (Golder & Huberman, 2006, p. 1). Golder and Huberman proposed a “dynamical model of collaborative tagging” that predicts stable tagging patterns. They hypothesized that users who tag resources stabilize on a set of tags in large part because they are influenced by the tagging behaviours of others. While their proposed model can be a useful contribution to tagging systems, Golder and Huberman’s study as a whole is too limited in scope to be generalized, since their results on tagging behaviour dynamics rely on only four days of tracked activity. Furthermore, their hypothesis still remains untested.

Using one month’s worth of Delicious data Cattuto (2006) examined the amount of tagging data associated with a single popular resource as a function of time. By computing the standard frequency-rank distributions for tags and plotting the data to find a power-law behaviour corresponding to a generalized Zipf’s law, Cattuto observed that once a tag becomes popular it remains popular. Using Delicious, Schmitz et al. (2006) explored the association between elements drawn from the tripartite model of folksonomies, i.e., users, resources and tags. Their study revealed that: (i) users who assign certain tags to some resources often also assign another set of tags to those resources; and (ii) users who assign certain resources a set of tags often also assign those tags to another set of resources. While Cattuto and Schmitz identify the existence of relationships among different tags, users, or resources, they do not provide any information about the nature of these relationships across time. Using 52 million photographs with 3.7 million tags in Flickr, Sigurbjörnsson and van Zwol (2008) analyzed how users tag their photographs. They observed that the majority of the photographs were assigned only a few tags and presented tag recommendation strategies to support user tagging. Marlow et al. (2006) examined the tagging incentives, tag usage, and vocabulary size of 25,000 Flickr users and found that on top of the desire to tag for personal benefit (such as organization and future retrieval of resources), the social networking element plays an important role. This analysis is helpful in outlining and motivating possible future directions of research in tagging systems, but since it does not present comprehensive empirical results its results cannot be easily generalized.

Sen et al. (2006) used a movie recommendation system, MovieLens, to investigate factors that influence people’s choices of tags and tagging behaviour. Their analysis showed that habit, investment, and community influence play important roles in vocabulary evolution: users reuse individual tags from their vocabulary and tend to create tags resembling other tags they see in the community. Louise Spiteri (2007) explored 30 days’ worth of tags from Delicious, Furl and Technorati, and analyzed and compared the tag structure the norms used in the construction of controlled vocabularies. Her results showed that in general, the tags were well constructed and followed many of the norms used in controlled vocabularies, but found ambiguity and polysemy as notable weaknesses associated with the structure of tags. Robert Bruce (2008) compared the controlled vocabulary from the ERIC database and the tags created by users of CiteULike and found that “CiteULike users do not use the same terminology as subject specialists who maintain descriptors in the ERIC thesaurus.” Hammond et al.’s (2005) study of Connotea includes tag convergence, recommendations, and directory creation. They observed that the distribution of shared tags in Connotea follows a “classic power law”. A shortfall of this paper, however, is that its observations are not based on any empirical research; furthermore the data set collected was not substantial enough to permit a generalization of their observations.
Using a dataset including 550,000 ‘tag assignments’ on CiteULike, Capocci and Caldarelli (2008) analyzed the frequency of tags as a function of time and established that tags in collaborative systems can be approximated with a Zipfian distribution, i.e., the frequency of any tag is inversely proportional to its rank in the tag frequency table, e.g., the most frequent tag occurs approximately twice as often as the second most frequent tag. They also explored the semantic relationships of tags as a way of uncovering the hidden relationships between tags. They established that tags on CiteULike exhibit complex structures and “users typically use tags hierarchically, labeling a resource by tags related to the same topics” (Capocci & Caldarelli, 2008, p. 5). Farooq et. al. (2007) analyzed two years’ worth of CiteULike data and explored six tag metrics: tag growth, tag reuse, tag non-obviousness, tag discrimination, tag frequency, and tag patterns. Results indicate that CiteULike exhibits consistent growth in tag vocabulary and that users do not reuse others’ tags while many tags were not reused, a few tags were reused many times and users reused tags from their personal collections; this may be because in CiteULike, tags from outside a user’s personal collection are not visible during the time of tagging. A larger data set that analyzes tag growth, reuse and discrimination across time among users with common domain interest and belonging to a group would strengthen the conclusions of these two studies.

Empirical research on tag vocabulary evolvement in enterprise settings is still rare. Millen, Feinberg, and Kerr (2006) conducted a study on Dogear, an enterprise social bookmarking system developed by IBM, during the first eight-week period of its deployment. They analyzed tag usage on the system and how the system helps users navigate resources. Their study revealed that the majority of users surveyed found Dogear helpful in finding information on both the corporate intranet and the external Web, but suggested the need for longitudinal studies. Damianos et al. (2007) conducted a case study at MITRE Corporation that explored the potential benefits of tagging in an enterprise using a trial social bookmarking system called “onomi”. They established that this kind of system is valuable in a corporate environment for sharing and disseminating, promoting information discovery, supporting communities and social networks, and feeding expertise finding. John and Seligman (2006) discussed the potential of tagging in the enterprise and presented an approach to rank experts based on tagging activity; they suggested that tagging can be used as a location tool for experts in the company. Farrell and Lau hypothesized that (i) “tagging is an effective way to keep track of and manage contacts; and (ii) tagging can provide information about individuals and their relationships to others inside the organization” (Farrell & Lau, 2006 , p. 4-5); both these hypotheses, however, still remain untested. Van Damme et al. (2008) explored ways to enrich folksonomies or to turn folksonomies into ontologies. They presented a model for deriving a “lightweight” corporate ontology from a corporate folksonomy; their results and models, however, have not yet been empirically tested and require further validation to be conclusive.

The analysis of the related literature suggests that the debate on collaborative tagging for organizational knowledge discovery, retrieval and dissemination is still in its infancy. While a few notable initiatives have been recorded highlighting the potential usefulness of collaborative tagging to organizational KM, more empirical research on tag vocabulary evolvement in enterprise settings is still very much needed. As noted above, organizations tend to develop and grow their own unique vocabulary. The unanswered question is, does collaborative tagging aid in the development and growth of domain-specific or organizational specific-vocabulary? Also, does collaborative tagging aid in the discovery, retrieval, and dissemination of knowledge objects in organizational KMS? Relatively little emphasis has been given to indexing strategies that simultaneously exploit tagging terms and controlled vocabularies, in particular, harvesting organization-specific user-created vocabulary from tags.
Methodology

Our study was conducted on CiteULike, a free online social bookmarking service that allows users to store, organize, and share scholarly papers. The analysis was performed on a specific dataset of data created by users belonging to a specific group or community of interest, extrapolated from the comprehensive dataset retrieved from the CiteULike data dump available at http://static.citeulike.org/data/2007-05-30.bz2. CiteULike has a number of unique features that can demonstrate the potential value of collaborative tagging in creating and building community-centred user-driven vocabulary. Unique to CiteULike is a Group feature that enables a user to start his or her group or join existing groups. Although tagging on CiteULike is initially done for the individual user's personal benefit, because of the growth in group numbers and member participation there is also a growing community focus. It is evident that contributing to a group is an important motivator to many users on CiteULike. Also, CiteULike is specifically designed to work with academic papers. In this regard, CiteULike can be characterized as attracting a somewhat niche set of users, a feature that is of interest and value to our study of enterprise KMS. Our interest is centered on examining a service that is not just aimed at a generalized audience but built for specific users; the tagging within a niche service is more specific. Because of CiteULike’s group feature and niche user characteristics, relevant communities-of-interest are formed and the tags within these communities are likely to be meaningful to the community members, a feature that can be useful in the creation of domain-specific vocabulary. As Emamy and Cameron pointed out, for example, “a tag consisting of the term 'evolution' applied to the World Wide Web as a whole could correspond to many possible interpretations of the word. [But within the context of CiteULike] the scope of such a tag is likely to be much narrower, and users searching on that term will retrieve many more targeted results” (Emamy & Cameron, 2007).

We searched CiteULike for groups matching the group name or description “business or management,” using the search syntax “business | management”, to filter out members with a common interest in business or management related topics. Sixty-four matching groups were found. From these 64 groups, groups with more than five participating members were selected. We selected groups with more than 5 members to ensure the retrieval of enough data for manipulation. Eight groups were selected. From these 8 groups we selected the group with the highest number of tagging activity as evidenced by the amount of tagging within the group and the number of active group users. The selected group was the “Knowledge_Management” group. We then extrapolated from the CiteULike comprehensive dataset the “Knowledge_Management” group dataset using the unique IDs of the group members. Inspired by the work of Farooq et al. (2007) we analyzed data from CiteULike to uncover tag activity and the value of the resulting collaboratively generated vocabulary using three tag metrics, tag growth, tag reuse, and tag discrimination. We investigated whether the evolving vocabulary is introducing any unique and meaningful vocabulary to the Knowledge_Management community, and whether the community members are making use (reusing) of this unique vocabulary to discover and retrieve new knowledge. The computing of the tag metrics; tag growth, tag reuse, and tag discrimination, was valuable in exploring collaborative tagging dynamics on CiteULike as discussed below.

Data Analysis

The extrapolated “Knowledge_Management” community tag dataset on CiteULike used for this study covered the period from December 2004 to September 2008 and comprised 242 distinct articles with 2,697 tags assigned to them. There were 449 distinct users and 709 distinct tags. A tag is regarded as distinct if it is introduced by the user as a tag; tag-variants or semantically-related tags were not considered to be similar, e.g. “KM” and “knowledgeManagement” were counted as two unique or distinct tags. The most popular tag was “email” (n=119 tag applications). The highest tagged article received 103 tags, and on average 11.15 tags were assigned per article. 19.8 % of the users (n=89) assigned 5 or more
tags. Table 1 below contains general descriptive statistics of the dataset employed in the analysis. The most prolific user applied 37.6% of the tags (n=1014). The average number of tag assignments per user was 6.01. The median was 2. The tag application range is 1013, suggesting that users differed greatly in their tagging activities. With a skewness value of over 20 it is evident that the data is highly asymmetric. The distribution of tags is significantly skewed to the right and the relatively high kurtosis value (438.6) indicates a distinct peak in the distribution of tags per user with a rather rapid decline. To put it into perspective a skewness value of 0 and a kurtosis value of 3 indicate a normal distribution.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tags</strong></td>
<td>Mean</td>
<td>6.0067</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>2.0000</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>2299.533</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>47.95345</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1014.00</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>1013.00</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>20.828</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>438.644</td>
</tr>
</tbody>
</table>

We explored tag distribution by analyzing the community’s tag activity across time. To ensure data manageability and meaningful analysis, for the temporal analysis, month was chosen as the unit of analysis. Figure 1 shows the community’s tag activity assigned over 46 months. The results indicate that within the community, tag activity as measured by the number of tags assigned each month varies across time. The most active month recorded 282 tag assignments, whereas the least active had 2 tags assigned. Although there are a few variations, the overall trend shows a high peak of tag activity in the middle of the 46 month study period. This shows that tag patterns exist within this community’s users’ tagging behaviour.
We then computed the correlation between the number of articles tagged and the number of unique tags for each user. The results in Table 2 suggest that there is a positive statistically significant relationship between the number of articles tagged and the number of unique tags for each user. In other words as users tag more articles the unique tag vocabulary they each use also increases, reflecting that as users discover new interests they add new and distinct tags to categorize and describe them.

**Table 2**

**Correlations Articles Tagged and Unique Tags**

<table>
<thead>
<tr>
<th></th>
<th>Tag</th>
<th>Correlation Coefficient</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman's rho</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>.062(**)</td>
<td></td>
<td>2697</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>2697</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Article</td>
<td>Correlation Coefficient</td>
<td></td>
<td>2697</td>
</tr>
<tr>
<td></td>
<td>.062(**)</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.001</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>2697</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed).**

**Analysis of Tag Growth**

We looked at the growth of the tag vocabulary as a measure of tag evolvement within the Knowledge_Management community on CiteULike. We explored tag growth over time as measured by the number of new tags applied per month. We categorized a tag as new when the tag was applied for the first time in the Knowledge_Management community, again without considering tag-variants. Figure 2 shows the number of new tags assigned over 46 months. Since December 2004 the tag vocabulary as indicated by the number of new tags assigned per month seems to be consistently growing over time, with a peak in the middle of the 46 months study period and a steady decrease in growth following the peak. One indicator of tag vocabulary growth is growth at a diminishing rate over time; this steady decrease of tag vocabulary growth might therefore suggest increasing stability in the community vocabulary over time and the establishment of domain-specific vocabulary. The stability of tag frequencies presumably
relies on both the interaction between community members (imitation) and the shared cultural knowledge of users. This is reminiscent of the behaviour observed in Cattuto’s study (2006) where the usage pattern of tags is subjected to some kind of frequency-bias. The tag vocabulary growth on CiteULike reveals an important aspect of collaborative tagging that might prove beneficial to organizational KMS. The fact that the distribution of tags stabilizes with time, denotes the emergence of a specific vocabulary that defines clearly the resource. Also the long-term stability of tag proportions makes the emergent vocabulary effective against noise limitations in the form of incorrectly spelled tags or irrelevant tags, which burdens users and decreases the system’s utility as mentioned above. Incorporating tags into KMS and taking these aspects into consideration might contribute greatly to the actual usability of KMS.

**Figure 2**

**Tag Growth: New Tags Assigned Over Time**

![Tag Growth Graph](image)

We also hypothesize that unique tag growth across time is strongly influenced by the increase in the number of new users. It is possible that tag growth is a result of the introduction of new tags from the new members who join the community. We therefore examined the relationship between tag growth and user growth within the community. We define a user as new when he/she applied a tag for the first time in this community; membership in this community has grown steadily over time. Figure 3 shows a positive relationship between the number of new users and the number of new tags applied across time.

**Figure 3**

**New Users and New Tag Assignment over time**

![New Users and Tags Graph](image)

Since the Knowledge Management community is still growing, one can conclude that the tag vocabulary growth rate within this community is gradually approaching a state of stability as noted above.
Evidence suggests that with time, this niche community will achieve critical mass, resulting in increasing stability in the community’s tag vocabulary.

### Table 3

**New Tag & New User Correlations**

<table>
<thead>
<tr>
<th></th>
<th>newuser</th>
<th>newtag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman's rho</td>
<td>1.000</td>
<td>.455(**)</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0</td>
<td>.001</td>
</tr>
<tr>
<td>N</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>newtag correlation</td>
<td>.455(**)</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.001</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>46</td>
<td>46</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).

### Analysis of Tag Reuse

One way of measuring whether the users in the community of study are increasingly collaborative is by looking at whether the tag vocabulary created is being reused among group members. By examining the use of previously used tags within the community one can explain how often the tag vocabulary created within a community is recycled among its users, which is a good indicator, therefore, of how domain-specific vocabulary is evolving. For a highly-collaborative community one would expect the tag vocabulary to converge and tag reuse to increase significantly over time. We posit that users belonging to the Knowledge Management community tend to reuse each others’ tags, thereby contributing to the building of domain-specific tags. Tag reuse among group members was determined by computing the average number of users who apply a tag using the metric first developed by Sen et al. (2006) and adapted by Farooq et al. (2007):

1. Tag reuse = Σ (# of distinct users for each tag) / # of tags

Since every tag is applied by at least one user the minimum value of this metric is 1.0. The tag reuse for this community was 23, which suggests that a fairly high number of tags was reused. Users were indeed reusing tags from the community’s collection of tags, and were thus increasingly collaborative. A high tag reuse value over time thus indicates an encouraging evolution of the tagging vocabulary: It shows that tagging is facilitating the development of core vocabulary for the users. This further supports the finding in the literature that users typically reuse tags. Collaborative tagging is also a vehicle for knowledge dissemination. Since on CiteULike tags used by others are not visible during the time of tagging, ruling out the effect of the user interface on tag reuse, one may conclude that users might be exploring the group’s library of tags and then tagging articles that they have seen others tagging, evidence to the strength of collaborative tagging in facilitating content sharing. The convergence of community members around a tag facilitates the sharing of tagged content among community members. Collaborative tagging therefore provides community members with new ways of discovering interesting resources. This supports the argument that collaborative tagging can be beneficial to the users’ search strategies by providing an increased number of entry points and a measure of serendipity. By browsing a community’s tags, the user can be introduced to potentially invaluable resources that would otherwise have been undiscoverable.
Analysis of Tag Discrimination

We also evaluated the value of tagging in creating distinct domain-specific vocabulary by examining the abilities of the tags used by the Knowledge_Management community users to discriminate between resources in the community’s library. Tag discrimination points to the informational or intellectual value of the tag. As Farooq et al (2007) stated “the information gained by tagging a collection of resources is proportional to how well the tag distinguishes between this collection of resources and other resources” (Farooq et al., 2007). In other words the importance of a tag as a keyword is inversely proportional to its frequency of occurrence. Mirroring the work of Farooq et al. (2007), we first computed the average number of distinct documents in my dataset that are associated per tag across time using the simple formula:

\[
\text{Tag discrimination} = \frac{\sum (\text{# of distinct articles for each tag})}{\text{# of tags}}
\]

For the Knowledge_Management community the tag discrimination value traced across time was 4.11 distinct articles per tag, which suggests that each tag is associated with 4.11 distinct articles in the community’s library. This is a reasonably good value considering that there were 242 distinct articles with 2,697 tags assigned to them. We went a step further and looked at the discrimination capabilities of the five most popular tags for this community as evidenced by their number of tag assignments. The top five tags for this group were: email (n=119), project--email (n=108), hci (n=77), visualization (n=71), and management (n=67). The most discriminating tag was “project--email” which tagged a total of 100 distinct articles and thus discriminated 41.3% of the articles in the community library (100/242 x 100). This represents a good indicator of the informational value of the tag “project--email” to the members of the Knowledge_Management community. This is a relatively good tag discrimination value, considering that the tag could have discriminated as low as 0.41% of the articles if the tag was associated with one article only. Tracing tag discrimination across time can be beneficial to KMS designers and administrators as they evaluate the usefulness of tags over time in their ability to discriminate among knowledge objects in the system. If a tag discriminates all of the 242 articles, for instance, it means that the tag is being randomly assigned to the articles in the library without any intellectual or informational gain attached to it. The tag does not result in any informational gain or any enrichment to the existing system and therefore can be removed from the system. Thus, the information on the discriminating value of a tag is useful to KMS designers who can determine which user-created vocabulary has intellectual value and should be added to or removed from the system. It is important, however, to note that “a tag can change in its discriminating value as more papers [or knowledge objects] are added and tagged in the system (Farooq et al., 2007).

Conclusion

This study has revealed that collaborative tagging can play a potentially important role in understanding and designing future features of organizational KMS. Some of the relative merits of collaborative tagging for knowledge representation, discovery, retrieval and sharing have been highlighted in this study. We have observed that as members of the Knowledge_Management community on CiteUlike tag more articles the unique tag vocabulary they each use also increases and the growth of these unique tags across time is strongly influenced by the increase in the number of new users joining the community. Although the community users exhibit a great variety in their sets of tags and frequency of tagging, the study reveals stable patterns emerging over time in the community’s growth of tags and reuse of tags, evidence of the community’s increasingly collaborative behaviour. Furthermore, the results show that the community users reuse each others’ tags, thereby contributing to the building of domain-specific tags. Tags assigned by the Knowledge_Management users have strong discriminating values, thereby establishing the intellectual value of the resulting tag vocabulary and usefulness to KMS designers. This
study therefore, provides implications for the design of KMS for indexing and accessing knowledge spaces. For knowledge management systems developers, collaborative tagging can help to develop enhanced vocabulary useful to the system. First, user-created tags can be embedded alongside professional indexing. Second, since developing a controlled vocabulary can be very costly and time consuming for KMS designers, tag vocabulary may serve as a useful resource for vocabulary choice or maintenance by KMS managers; for example a frequently-occurring tag can denote a community-driven, preferred term to describe or categorize a given concept. Moreover, as noted above, the core strength of collaborative tagging to knowledge discovery depends on users’ ability to identify and connect with other users who share the same interests. Using user-created tags therefore, enables the discovery of shared interests and expertise within an organization and thus facilitates the identification and formation of communities of expertise or practice. Third, users can add tags to professionally-indexed resources in an organizational knowledge base, thereby personalizing their collections.

Since the dataset used for this study is specific to group or community tagging behavior on CiteULike, it is hard to generalize or extend the results to other general social bookmarking systems. Further investigation is necessary to explore indexing strategies that simultaneously exploit tagging terms and controlled vocabularies in community specific environments such as the CiteULike Knowledge_Management group. In particular to enterprise KM, harvesting enterprise vocabulary from tags, domain specific subject headings and assessing a number of different integrating strategies are of great interest. Collaborative tagging as an organizational method opens up possibilities for knowledge representation, knowledge discovery, knowledge retrieval and knowledge dissemination that were previously impossible. We believe that this study has provided fresh insight into the properties that affect the evolution and utility of tagging communities, insight that can be used by KMS designers to shape the behavior of their own systems.

References


