THE IMPACT OF PRE-DEFINED TERMS ON THE VOCABULARY OF COLLABORATIVE INDEXING SYSTEMS

Kowatsch, Tobias, Hochschule Furtwangen University, Robert-Gerwig-Platz 1, 78120 Furtwangen, Germany, tobias.kowatsch@hs-furtwangen.de

Maass, Wolfgang, Hochschule Furtwangen University, Robert-Gerwig-Platz 1, 78120 Furtwangen, Germany, wolfgang.maass@hs-furtwangen.de

Abstract

Collaborative indexing systems have attracted an increasing amount of attention over the last three years. One fundamental limitation to such a system is the uncontrolled nature of its vocabulary, as this consists of terms users freely choose to index resources. As a result, the vocabulary can be poorly structured, making it difficult to harvest knowledge from the user community. Pre-defined terms are suggested to reduce this uncontrolled vocabulary. However, this suggestion has not yet been proven. This work therefore focuses on an empirical study of the adoption of pre-defined terms and its impact on the community's vocabulary by implying innovation diffusion theory. A research model is formulated to explain the relationship between the degree of adoption and its impact on the vocabulary in order to indicate consolidated term usage. For this purpose, constructs of social network analysis are applied. The model is then validated by one lab experiment (n=172), before being cross-validated by two open web experiments (n=254, n=160). Results indicate that to a remarkable extent, pre-defined terms are appropriate for reducing the uncontrolled nature of the community's vocabulary, such that the utility of collaborative indexing systems can be increased.

Keywords: Collaborative Indexing System, Online Communities, Social Network Analysis, Pre-defined Terms, Adoption, Uncontrolled Vocabulary

1 INTRODUCTION

A collaborative indexing system is either a stand-alone web portal such as Delicious or an add-on service of social networking systems such as Xing[1]. It offers basically two features. First, it is used to index resources such as individuals, websites or images for future retrieval. Second, it allows the exploration of other users' resources, e.g. to find users with the same interests. Indexing systems have attracted considerable attention over the last four years. A survey of collaborative indexing systems conducted by the Pew Internet and American Life Project in December 2006 found "that 28% of [the American] internet users have tagged or categorized content online such as photos, news stories or blog posts" (Rainie, 2007, p. 1). Current empirical data based on the indexing system Delicious indicate constant usage rates as Table 1 shows: over 300 000 active users indexed over two million websites within five months from late 2006 to early 2007. In particular, usage rates for Delicious have remarkably increased over the last two years as indicated by the live statistics of Keller[2].

Delicious	Sep '06	Oct '06	Nov '06	Dec '06	Jan '07	Total
Active user	4 282	4 413	4 430	4 141	4 873	329 984
Indexing rates	16 274	16 488	15 061	14 167	16 072	2 358 891

Table 1. Active users and indexing rates on a per-day basis; this data was calculated based on the work of Maass et al. (2007).

In light of these facts, there is a need for a deeper understanding of collaborative indexing systems and their potential to harvest network information of its underlying online community. Up until now, research has been focused on the principles and architecture of these systems. Furthermore, usage and vocabulary patterns have been studied, and several algorithms have been suggested to optimize the indexing process as well as the retrieval and visualization of resources. The deployment of collaborative indexing systems in both the public sector and in commercial enterprises has also been described.

However, prior work indicates that there is one fundamental limitation of such a system: the uncontrolled nature of its vocabulary. To recapitulate by citing Marlow et al. (2006), an indexing system can be modelled as a network of users, terms, and resources. In this case, the partial network of all terms represents the vocabulary. As terms are freely chosen by users to index their resources, they are not restricted to a controlled vocabulary. Although this kind of indexing freedom contributed a lot to the huge success of collaborative indexing systems, their vocabularies are uncontrolled with grave effects. For instance, the vocabulary problem in human-system communication as described by Furnas et al. (1987) causes the partial network of terms to be loosely coupled because terms are used inconsistently amongst users. For search or recommendation features, however, a network of strong links is required to identify significant clusters of semantically related users, resources or terms, such that users of a collaborative indexing system benefit from it.

Due to this uncontrolled nature of the vocabulary, the construct of pre-defined terms is suggested as a partial solution, since it helps the user index resources with terms that were already used by other users or were recommended for use by the indexing system. In such, if users would adopt pre-defined terms to index their resources, consolidated term usage would be expected to result in a network of terms consisting of strong links by which the uncontrolled nature of the vocabulary is reduced. Since this suggestion has not yet been validated, this article is devoted to an empirical study of the adoption of pre-defined terms in collaborative indexing systems and its impact on the community's vocabulary.

[1] http://www.xing.com
[2] http://deli.ckoma.net/stats

Correspondingly, a measurement is developed to calculate the degree of the adoption of pre-defined terms. Constructs of social network analysis are used to describe the vocabulary for identifying consolidated term usage. These aspects are then formulated within one integrative research model, which is validated by conducting online experiments. Finally, differences of vocabularies that emerged with and without pre-defined terms indicate whether or not the uncontrolled nature of the vocabulary is reduced when pre-defined terms are available to users.

Based on this research design, the current work starts with related work on collaborative indexing systems and social network analysis. Then, we motivate this work by referring to the uncontrolled vocabulary and raise three research questions. Afterwards, we introduce the research model and present corresponding hypotheses. The methodical approach is then described. Subsequently, we present the results and discuss them. Finally, we conclude and give an outlook on further work.

2 RELATED WORK

Basic principles of collaborative indexing systems are presented by Mathes (2004), who discusses the indexing process and describes the uncontrolled nature of the resulting vocabulary. Due to the fact that several kinds of indexing systems are available on the web, there has been a need for their classification. Following the preliminary work of Hammond et al. (2005), Marlow et al. (2006) introduced a more detailed taxonomy. As this taxonomy indicates, pre-defined terms belong to the dimension of indexing support. Marlow et al. also identify multiple incentives for using collaborative indexing systems. Two high-level categories based on organizational and social incentives are indicated. In combination with these incentives, users augment terms with several semantics. Golder and Huberman (2006) as well as Zhichen et al. (2006) describe such semantics used within collaborative indexing systems. Zhichen et al. point out that pre-defined terms should cover several semantics to be adopted by users. Correspondingly, they introduce an algorithm with the objective of optimizing the indexing process.

Hotho et al. (2006a), Dubinko et al. (2006) and Aurnhammer et al. (2006) introduce algorithms to optimize the retrieval and visualization of indexed resources. The extraction of ontologies from the vocabulary of collaborative indexing systems is described by Heymann and Garcia-Molina (2006), Begelman et al. (2006), Schmitz (2006) and Mika (2005). All of these techniques strongly depend on the quality of underlying vocabularies. Therefore, they could benefit from features such as pre-defined terms if the uncontrolled nature of the vocabulary was reduced. For further reading on collaborative indexing systems, see the additional literature reviewed by Voss (2007).

Constructs of social network analysis are applied to the partial network of terms. To identify consolidated term usage, the current work refers to the standard work of Wasserman and Faust (1994), which describes general structural patterns of social networks as well as methods to analyze them, respectively.

3 MOTIVATION

Terms are one of the core elements of collaborative indexing systems, which are freely chosen by users to index resources. Correspondingly, they link users and resources, and form a network as described by Marlow et al. (2006). The partial network of all distinct terms represents the vocabulary, which in turn represents the underlying framework for several applications as described in the following list:

- **Search and Exploration:** Terms establish links between users and resources and vice versa (Marlow et al. 2006). Hence, they represent the structural knowledge of each individual user according to Wu et al. (2006), as cited in Diekhoff and Diekhoff (1982). Additionally, terms are also linked to each other and their links may be weighted based on frequency rates (Maass et al. 2007), reputation scores (Zhichen et al. 2006) or other relevant factors like time-based coefficients.

These links are usually visible to the whole community of a collaborative indexing system, thus facilitating search and exploration features (Hotho et al. 2006a).

- **Describing the community:** By applying frequency analysis (Maass et al. 2007) or clustering and ontology extraction techniques (Heymann and Garcia-Molina 2006, Begelman et al. 2006, Schmitz 2006, Mika 2005) to the vocabulary, the community of a collaborative indexing system can be described in detail according to their usage of terms. Also, groups of interest can be identified if they use terms and resources in a consistent manner (Wu et al. 2006).
- **Identifying domain experts:** With the same methods used to describe the community, information leaders and domain experts in a collaborative indexing system can both be identified (Huberman 2004). In particular, this application could prove useful in finding appropriate staff members within larger organizations or on the web (Millen et al. 2006, Kanawati and Malek 2002).
- **Indicating historical developments:** Since timestamp information is stored with each indexing task, historical developments of user, term or vocabulary characteristics can be computed. Examples of such characteristics include the frequency rates of an individual user indexing resources and the density or the most widely used terms of a community's vocabulary, which would illustrate a change in the structural knowledge within a given period of time (Maass et al. 2007). Based on historical developments, topic-specific trends can also be identified (Hotho et al. 2006b), which would be useful for organizations developing new products and services.

These applications depend strongly on the structure of the vocabulary. For instance, a poorly structured vocabulary consisting of a couple of terms connected by only a few links would make it difficult to identify significant clusters of semantically related users, resources, or terms. In such, the formation of a distinctive and meaningful vocabulary is desired, but the uncontrolled nature of the vocabulary in collaborative indexing systems restricts the emergence of the desired structure as explained below. Since there are no rules defining the use of terms to index resources, they are freely chosen by users and therefore not restricted to a controlled vocabulary like they usually are in libraries (see Reitz 2004, p. 177). In detail, Golder and Huberman (2006) identify polysemy, synonymy and basic level variation problems by analyzing the vocabulary of Delicious. Basically, Furnas et al. (1987) refer to such findings as the vocabulary problem in human-system communication.

Pre-defined terms, which are provided by some collaborative indexing systems to assist the user in finding appropriate terms for indexing their resources, might be a partial solution towards the uncontrolled nature of the vocabulary as suggested by Marlow et al. (2006) and Zhichen et al. (2006). Since this suggestion has not yet been proven, our approach is to empirically study the use of pre-defined terms and their impact on the community's vocabulary. As pre-defined terms are shown to the user as innovation that can be adopted or rejected to index a resource, we take up the perspective of innovation diffusion theory (Rogers 2003). Therefore, we are interested in the adoption of pre-defined terms and its impact to reduce the uncontrolled vocabulary of collaborative indexing systems. This is done in two steps. First, the actual degree of adoption must be developed as an objective measurement, because the impact of pre-defined terms on the vocabulary is only significant if users will actually adopt them. Therefore, our first research question is formulated as follows:

1. To which degree do users of collaborative indexing systems adopt pre-defined terms?

Second, we have to identify properties of the vocabulary that appropriately measure consolidated term usage. Correspondingly, these properties must indicate the impact of the adoption of pre-defined terms on the community's vocabulary. Hence, the second research question addresses these properties:

2. Which properties of the vocabulary of collaborative indexing systems are relevant for measuring consolidated term usage?

4 RESEARCH MODEL

4.1 Degree of adoption of pre-defined terms

According to the definition provided by Rogers (2003, p. 473), adoption corresponds to a "decision to make full use of an innovation as the best course of action available". Here, pre-defined terms are meant to be innovations for the user who wants to index a specific resource. The adoption of pre-defined terms is determined by their first usage as definitively new ones. The second time of usage, the same terms are reused but not adopted.

To calculate the degree of adoption, one must distinguish between pre-defined, personal and newly created terms. Pre-defined terms are provided by the system and stem from other users' indexing tasks, or through automated content or context analysis techniques for which Zhichen et al. (2006) introduced the construct of virtual users. By contrast, personal terms represent those terms already utilized by the user for prior indexing tasks. Personal terms underlie the control of each user but may overlap with pre-defined terms since they are system-controlled. Additionally, terms may be newly created by the user, in which case they do not belong to the set of personal or pre-defined terms before usage, although they might belong to other users' personal or pre-defined terms. After usage, all terms chosen by the user – whether pre-defined, personal or newly created – belong to the set of personal terms. Hence, one particular pre-defined term can only be adopted once. In summary, adoption of pre-defined terms occurs only if a user utilizes a pre-defined term for indexing that is neither part of his or her personal terms, nor has been newly created.

By means of the set theory, all terms being adopted by a user x $T^x_{adopted}$ are a subset of pre-defined terms $T^x_{pre\text{-}defined}$, but do not belong to the set of personal terms $T^x_{personal}$ or new terms T^x_{new} before they are used to index a resource, c.f. Equation 1. By contrast, adopted terms $T^x_{adopted}$ belong to the new set of personal terms $T^x_{personal}$ after indexing, c.f. Equation 2.

(1) Before indexing a resource: $T^x_{adopted} \subseteq T^x_{pre\text{-}defined} \setminus (T^x_{personal} \cup T^x_{new})$

(2) After indexing a resource: $T^x_{adopted} \subseteq T^x_{personal}$

Given the set of adopted terms by a user x after all former indexing tasks $T^x_{adopted}$, the user's degree of adoption $D^x_{adoption}$ is defined by the fraction of the cardinality of $T^x_{adopted}$ and $T^x_{personal}$, reflecting the ratio of adopted terms to all of a user's personal terms:

$$(3) \qquad D^x_{adoption} = \frac{|T^x_{adopted}|}{|T^x_{personal}|}$$

The values of $D^x_{adoption}$ range from zero if no terms have been adopted at all to one if all personal terms have been adopted from the list of pre-defined terms. To calculate the community's degree of adoption $\overline{D}_{adoption}$, one must build the sum of $D^x_{adoption}$ for all N users divided by N:

$$(4) \qquad \overline{D}_{adoption} = \frac{1}{N} \cdot \sum_{x=1}^{N} D^x_{adoption}$$

With this mathematical framework, preconditions are set up to answer the first research question regarding the degree to which users adopt pre-defined terms.

4.2 Formalisation of the community's vocabulary

Based on the previous work of Mika (2005) and Maass et al. (2007), an indexing task can be described as a quadruple consisting of four entities: < user, term*, resource, timestamp >. Correspondingly, a

specific resource is indexed by one user with none, or one or more terms at a certain time. For our approach, only the syntax of the entity term is considered with the exception of case sensitivity (i.e. *house* and *House* are referred to as the same term). The semantics of terms are not analyzed in this work. As described in Section 4.1, terms belong to the set of pre-defined terms, personal terms, or are newly created by the user before indexing. In this section, we focus only on indexing tasks that have already been completed. Hence, the terms always belong to the set of personal terms, thus reflecting each user's vocabulary, or when aggregated over all N users, the community's vocabulary. To identify adequate properties that indicate the adoption of pre-defined terms, the vocabulary V can be described by a frequency matrix $F(V)$, such that the value on the *i*th row and the *j*th column, denoted as $f(i,j)$, is equal to the frequency rate of two co-occurring terms ($i \neq j$) or the frequency rate of one term ($i = j$). It should be noted that the frequency rate of one term t_i might not be the sum of all frequency rates of terms, which co-occur with t_1, as one term or more terms can be used to index a resource at the same time.

Figure 1 shows the frequency matrix $F(V)$ of a vocabulary consisting of five terms. For instance, $term_1$ co-occurs twice with $term_2$, once with $term_4$ and five times with $term_5$. Overall, $term_1$ was used six times to index resources.

i/j	1	2	3	4	5
1	6	2	0	1	5
2	2	4	0	0	4
3	0	0	1	0	0
4	1	0	0	3	2
5	5	4	0	2	7

Figure 1. Frequency matrix of a vocabulary consisting of five terms.

4.3 Hypotheses and research model

The first property adequate for indicating consolidated term usage represents the size of the community's vocabulary. The size of the vocabulary $S(V)$ is given by the cardinality of the set of all terms $T(V)$ that belong that vocabulary. This is also the total number of rows or columns of the frequency matrix $F(V)$

(5) Size of the vocabulary: $S(V) = |T(V)| = T$

This size would not be increased if pre-defined terms were adopted because they already belong to the community's vocabulary. This consideration implicitly suggests that users created new terms that are not syntactically equivalent with terms of the community's vocabulary. As a result, the adoption of pre-defined terms yields a constant vocabulary size by preventing the creation of new terms. Therefore, the first effect of pre-defined term adoption refers to the size of the community's vocabulary depending on the community's degree of adoption:

H1 The community's degree of adoption $\overline{D}_{adoption}$ will have a negative relationship with the size of the community's vocabulary $S(V)$.

The density of the community's vocabulary represents the second property, which is adequate for measuring consolidated term usage. According to Wasserman and Faust (1994, p. 143), the density of the vocabulary $\delta(V)$ is given by the average frequency values of all co-occurring terms, thus reflecting a low or a high-weighted vocabulary. The weights on the diagonal values in the frequency matrix are therefore omitted:

(6) Density of the vocabulary: $\delta(V) = \frac{1}{T \cdot (T-1)} \cdot \sum_{i=1}^{T} \sum_{j=1}^{T} f(i,j) \quad \text{with} \quad i \neq j$

Based on consolidated term usage, we assume a higher strength of co-occurring terms if at least two pre-defined terms have been adopted within one indexing task. The likelihood that pre-defined terms are connected to each other is obviously greater than a connection between a newly created term and all pre-defined terms or – in the worst case – when both terms are newly created yet had not previously belonged to the community's vocabulary. Furthermore, the introduction of at least one newly created term could considerably reduce the density due to the denominator $T \cdot (T-1)$. By contrast, connecting pre-defined terms, which had not been previously linked, through adoption would increase the density by establishing a new connection between them and simultaneously preserving the size of the vocabulary. Therefore, we formulate the following effect based on the density of the vocabulary and the degree of adoption:

H2 The community's degree of adoption $\overline{D}_{adoption}$ will have a positive relationship with the density of the community's vocabulary $\delta(V)$.

Finally, we will introduce the average term frequency as a property that indicates consolidated term usage. Correspondingly, the average term frequency is given by the sum of all diagonal values $f(i,j)$ with $i = j$ divided by the size of the vocabulary:

(7) Average term frequency: $\bar{f}(V) = \frac{1}{T} \cdot \sum_{i=1}^{T} \sum_{j=1}^{T} f(i,j) \quad \text{with} \quad i = j$

The average term frequency predicts consolidated term usage since each adoption of pre-defined terms leads to an increased frequency rate of the terms that already belong to the community's vocabulary. In contrast to the adoption of pre-defined terms, newly created terms that do not belong to the community's vocabulary will reduce the average frequency rate because their rates always start with one. We therefore hypothesize the following relationship:

H3 The community's degree of adoption $\overline{D}_{adoption}$ will have a positive relationship with the average term frequency of the community's vocabulary $\bar{f}(V)$.

To summarize, Figure 2 illustrates the impact of the community's degree of pre-defined term adoption on the properties size, density and average term frequency of the community's vocabulary.

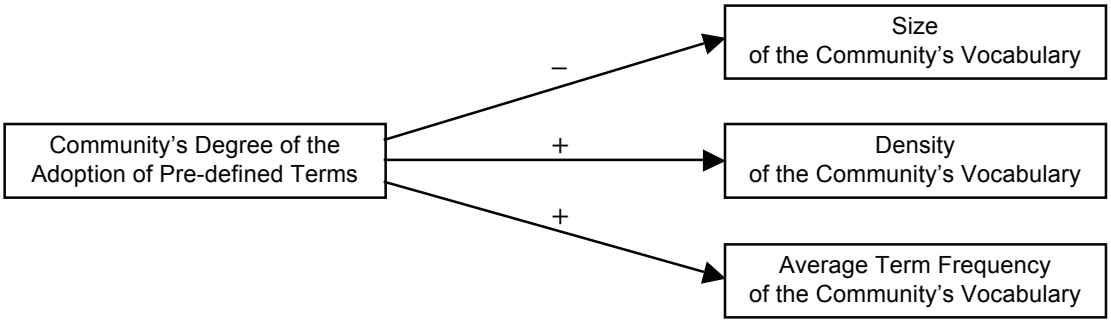

Figure 2. Impact of the adoption of pre-defined terms on the community's vocabulary

5 METHOD

An online experiment for German and English-speaking subjects was developed to calculate the community's degree of pre-defined term adoption and to verify the hypotheses. In the first part of this experiment subjects were instructed to index several websites. Before each indexing task, a screenshot of the corresponding website was presented to the participants. Pre-defined terms were only shown to the experimental group; they were deactivated for the control group to calculate and compare the impact on each vocabulary that resulted from all participants. Wording, layout and functionality of the

indexing page (e.g. listing and auto-complete mechanism of pre-defined terms) were adapted from Delicious, and the wording for German-speaking subjects was adapted from the German collaborative indexing system Mister Wong (http://www.mister-wong.de).

To increase external validity, we identified two semantic domains from the list of Delicious' *popular tags* (http://del.icio.us/tag/) that were both common for this kind of free-for-all collaborative indexing system: Photography & Images and Jobs & Career. Then, websites for each domain and the corresponding pre-defined terms were selected from Delicious and Mister Wong for the English and German versions of the online experiment, respectively. These websites were shown randomly to the subjects for indexing. Two well-known websites were also selected for the first two indexing tasks so that the subjects could get acquainted with the indexing system. In addition, two well-known websites were put between the shuffled websites of the two semantic domains to make indexing more interesting. The appendix lists all websites, their order, and the corresponding pre-defined terms for reference. After indexing the websites, subjects had to complete a questionnaire to provide personal data and feedback on the length and comprehensibility of the experiment.

One pretest and two studies were conducted. The pretest was accomplished with 15 students and had the objective of assessing and improving a preliminary version of the online experiment. After modifications were made, empirical data was collected from two studies. The first study was conducted with German students of Furtwangen University in a lab environment (n=172). By contrast, the second study was conducted as an open web experiment with German (n=254) and English-speaking (n=160) participants to validate the findings of Study 1, as well as to add external validity to the research model presented above. The number of websites has been reduced from 17 to 14 for Study 2 because of the participant's feedback on the length of the experiment in Study 1. Subjects for Study 2 have been acquired through mailing lists and weblogs that addressed collaborative indexing. In addition, the Web Experimental Psychology Lab and the Web Experiment List were utilized to sample participants for the online experiment (Reips 2001).

6 RESULTS AND DISCUSSION

As a general overview of the assessment of the online experiment's external validity Table 2 shows the average term usage rates in all indexing tasks for the experimental group and the control group. According to these figures, the subjects of both studies indexed each website with 2.81 to 3.60 terms on average. Since this range minimally exceeds the findings of previous field research (Maass et al., 2007), it adds external validity to the results of the next sections.

	Experimental Group	**Control Group**
Study 1 (n=86) 86 * 16 = 1462 Indexing tasks	3.27	3.60
Study 2a: German-speaking subjects (n=127) 127 * 14 = 1778 Indexing tasks	2.81	2.96
Study 2b: English-speaking subjects (n=80) 80 * 14 = 1120 Indexing tasks	3.39	3.20
Maass et al. (2007) Delicious: 452 806 Indexing tasks Connotea: 92 333 Indexing tasks CiteULike: 3 798 Indexing tasks	2.58	 2.71 2.57

Table 2. Average term usage rates.

6.1 Degree of adoption of pre-defined terms

The community's degree of adoption is presented in Figure 2. Findings indicate that more than half of all personal terms belong to the set of pre-defined terms. In other words, at least every other personal term was adopted from the set of pre-defined terms, which answers our first research question. In both

studies, the adoption rates of the topics Photography & Images (P&I) and Jobs & Career (J&C), which were covered by a number of similar websites, were lower than the adoption rates of the four general topics that were covered by four well-known websites only. Thus, it is suggested that multiple websites on one topic (as conceived for the topics P&I and J&C) were differentiated between more than the websites on general topics. This means that subjects created new terms rather than adopting pre-defined terms only to differentiate between websites within the same semantic domain. As a result, each newly generated term resulted in a reduction of the adoption rate.

A second aspect refers to the quantity of websites within the topics P&I and J&C. Analogue to the suggestion above, adoption rates of the first study were lower than those of the second because more websites had been indexed within one semantic domain in Study 1 (seven for P&I and six for J&C) than in Study 2 (five each for P&I and J&C).

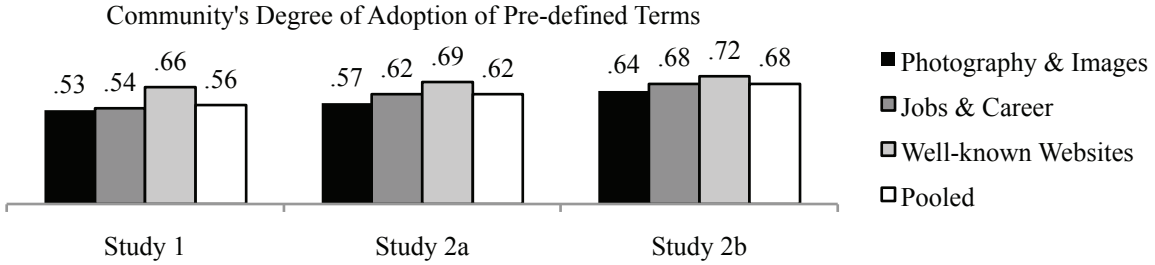

Figure 2. Adoption rates of pre-defined terms for the topics Photography & Images, Jobs & Career, well-known websites and all websites together. Note: Study 1 (n=86, German-speaking subjects), Study 2a (n=127, German-speaking subjects) and Study 2b (n=80, English-speaking subjects); n applies only for the experimental group.

6.2 Impact on the community's vocabulary

Taking the findings of the last section into account, the average degree of adoption resulted in remarkable effects on the community's vocabulary. The size of the vocabulary created by the experimental group was nearly half the size of the control group's vocabulary in Study 2 for the English-speaking subjects. Moreover, the density of the same vocabulary was three times higher and the average term frequency was twice as high as the corresponding measures of the control group's vocabulary. An overview of all indices is provided in Figure 3 for the size, density and average term frequencies of the community's vocabulary. Our findings indicate consolidated term usage with respect to the degree of adoption. As a result, all three hypotheses are supported by the empirical data. Hence, size, density and average term frequency of the community's vocabulary are relevant properties for measuring consolidated term usage, which answers our second research question.

6.3 Discussion

The findings of the online experiment show that subjects adopted pre-defined terms to a remarkable extent. At least 50 percent of their personal terms stem from the list of pre-defined terms. Regarding the impact of adoption on the uncontrolled vocabulary, consolidated term usage was identified by the vocabulary's size and density, as well as the average term frequency. The suggestions of Marlow et al. (2006) and Zhichen et al. (2006) are therefore supported by these results, as the uncontrolled nature of the vocabulary was significantly reduced when pre-defined terms were available to participants.

However, there are some restrictions in relation to the explanatory power of the results. In the first place, external validity was limited due to the experimental setting, especially for the first study that was conducted in a lab environment. Correspondingly, websites were pre-selected and could therefore not meet the personal interests of each individual participant in detail. Also, pre-defined terms were restricted to five items for each indexing task and subject. Therefore, they were not dynamically

changing over time and could not represent all of the semantic facets of the corresponding websites as expected at Delicious or Mister Wong. Moreover, resources were restricted to textual content and did not cover individuals, images, podcasts or video clips.

By contrast, the degree of adoption was introduced to indicate the adoption of pre-defined terms objectively, and social network analysis was effectually deployed to identify consolidated term usage. The internal validity of the online experiments was increased due to the standardized environment. On the other hand, consistency of results in both studies (lab and web environment) and language versions increased external validity. In summary, the research model was formulated and empirically validated to better understand the impact of pre-defined terms on the community's vocabulary of collaborative indexing systems.

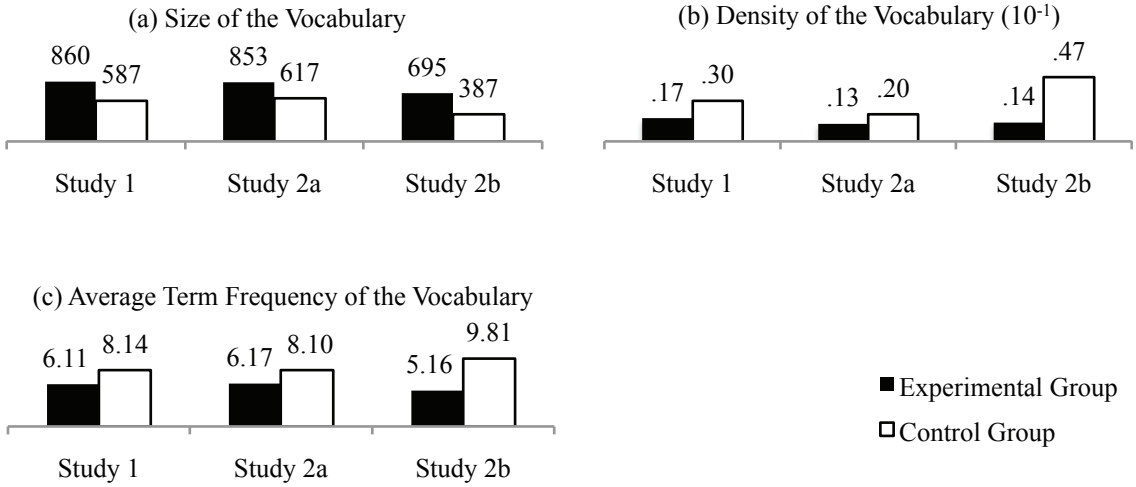

Figure 3. Size (a), density (b) and average term frequency (c) of the community's vocabulary for the experimental and the control group; Note: Study 1 (n=86, German-speaking subjects), Study 2a (n=127, German-speaking subjects) and Study 2b (n=80, English-speaking subjects); n applies for both experimental and control group.

7 CONCLUSION AND FUTURE WORK

One fundamental limitation of collaborative indexing systems is the uncontrolled nature of the community's vocabulary because terms are freely chosen by users to index their resources and are not restricted to a controlled vocabulary. Because pre-defined terms have been suggested to reduce this limitation through consolidated term usage, we studied the adoption of pre-defined terms and its impact on the community's vocabulary empirically. Three properties were identified that measure consolidated term usage depending on the community's degree of the adoption of pre-defined terms: the size, density and average term frequency of the community's vocabulary. As a managerial recommendation, collaborative indexing systems should support users with pre-defined terms to reduce the uncontrolled nature of the vocabulary as done by Delicious. Accordingly, pre-defined terms can support search tasks and therefore increase the utility of collaborative indexing systems and other knowledge-intensive information systems.

With regard to future work, some challenges still remain. As described by Golder and Huberman (2006) and Zhichen et al. (2006) terms are augmented with several semantics by users of collaborative indexing systems. Correspondingly, the adoption of pre-defined terms may depend on the coverage of these semantics. In addition, terms that are identified as part of one semantic domain by frequency analysis (Maass et al. 2007) or clustering techniques (Mika 2005) are also suggested as being candidates for recommendation. For this reason, further clarification is needed to optimize the selection process and the quantity of pre-defined terms.

References

Aurnhammer, M., Hanappe, P., and Steels, L. (2006). Augmenting navigation for collaborative tagging with emergent semantics. In Proc. of the 5th International Semantic Web Conference (ISWC '05), Athens, GA, USA.

Begelman, G., Keller, P., and Smadja, F. (2006). Automated tag clustering: Improving search and exploration in the tag space. In Collaborative Web Tagging Workshop at the 15th International World Wide Web Conference (WWW '06), Edinburgh, Scotland.

Cattuto, C., Loreto, V., and Pietronero, L. (2006). Collaborative tagging and semiotic dynamics. ArXiv Computer Science e-prints, http://arxiv.org/abs/cs.CY/0605015.

Diekhoff, G. M. and Diekhoff, K. B. (1982). Cognitive maps as a tool in communicating structural knowledge. Educational Technology, 22(4), 28–30.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., and Tomkins, A. (2006). Visualizing tags over time. In Proc. of the 15th International World Wide Web Conference (WWW '06), 193–202, New York, USA. ACM Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.

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

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

Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Stanford University.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006a). Information retrieval in folksonomies: Search and ranking. In Sure, Y. and Domingue, J., editors, The Semantic Web: Research and Applications, volume 4011 of LNAI, 411–426, Heidelberg, Germany. Springer.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006b). Trend Detection in Folksonomies. In Avrithis, Y. S., Kompatsiaris, Y. and Staab, S., and O'Connor, N. E., editors, Proc. of the 1st International Conference on Semantics And Digital Media Technology (SAMT), volume 4306 of LNCS, 56-70, Heidelberg, Germany. Springer.

Huberman, B. A. (2004). New ways of identifying and using organizational information. IST News. http://cordis.europa.eu/ictresults/index.cfm/section/news/tpl/article/BrowsingType/Features/ID/69226

Kanawati, R. and Malek, M. (2002). A multi-agent system for collaborative bookmarking. In Proc. of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '02), 1137–1138, New York, USA. ACM Press.

Maass, W., Kowatsch, T., and Muenster, T. (2007). Vocabulary patterns in Free-for-all Collaborative Indexing Systems. In Proc. of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE) at the 6th International Semantic Web Conference (ISWC '07), Busan, Korea.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read. In Proc. of the 17th Conference on Hypertext and Hypermedia (HYPERTEXT '06), 31–40, New York, USA. ACM Press.

Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata. Technical report, Graduate School of Library and Information Science, University of Illinois.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In Proc. of the 4th International Semantic Web Conference (ISWC '05), Galway, Ireland.

Millen, D. R., Feinberg, J., and Kerr, B. (2006). Dogear: Social bookmarking in the enterprise. In Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06), 111–120, New York, USA. ACM Press.

Rainie, L. (2007). Tagging play, forget dewey and his decimals, internet users are revolutionizing the way we classify information – and make sense of it. Pew Internet and American Life Project. http://pewresearch.org/pubs/402/tagging-play

Reips, U.-D. (2001). The web experimental psychology lab: Five years of data collection on the internet. Behavior Research Methods, Instruments, & Computers, 33(2), 201–211.

Reitz, J. M. (2004). Dictionary for Library and Information Science. Libraries Unlimited, Westport, Conn., USA.

Schmitz, P. (2006). Inducing ontology from flickr tags. In Collaborative Web Tagging Workshop at the 15th International World Wide Web Conference (WWW '06), Edinburgh, Scotland.

Voss, J. (2007). Tagging, folksonomy & co - renaissance of manual indexing? ArXiv Computer Science e-prints, http://arxiv.org/abs/cs/0701072.

Wasserman, S. and Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, UK.

Wu, H., Zubair, M., and Maly, K. (2006). Harvesting social knowledge from folksonomies. In Proc. of the 17th Conference on Hypertext and Hypermedia (HYPERTEXT '06), 111–114, New York, USA. ACM Press.

Zhichen, X., Yun, F., Jianchang, M., and Difu, S. (2006). Towards the semantic web: Collaborative tag suggestions. In: Collaborative Web Tagging Workshop at the 15th International World Wide Web Conference (WWW '06), Edinburgh, Scotland.

Appendix: Websites and Pre-defined Terms

English websites and pre-defined terms based on the collaborative indexing system Delicious, http://del.icio.us. German websites and pre-defined terms for German-speaking participants are available from the author. W: Well-known website

ID	Domain	URL (http://www.*)	Pre-defined Terms
W1	Shopping	amazon.com	amazon, shopping, books, music, dvd
W2	Web Search	google.com	searchengine, websearch, google, search, web
W3	Auctions	ebay.com	ebay, shopping, auctions, buy, shop
W4	Video	youtube.com	video, videos, youtube, media, web2.0
J1	Jobs & Career	simplyhired.com	jobs, job, career, jobsearch, employment
J2	Jobs & Career	jobster.com	jobs, job, career, jobsearch, socialnetworking
J3	Jobs & Career	theItJobBoard.com	jobs, job, IT, search, employment
J4	Jobs & Career	jobserve.com	jobs, job, career, jobsearch, recruitment
J5	Jobs & Career	jobsearch.com	jobs, job, career, jobsearch, agency
J6	Jobs & Career	monster.com	jobs, job, career, jobsearch, search
P1	Photography & Images	picsearch.com	search, images, photos, photo, pictures
P2	Photography & Images	flickr.com	photos, flickr, photo, photography, sharing
P3	Photography & Images	sxc.hu	photography, history, photos, photo, images
P4	Photography & Images	photobucket.com	photo, photos, hosting, images, photography
P5	Photography & Images	freefoto.com	photos, free, images, photography, photo
P6	Photography & Images	freedigitalphotos.net	photos, free, images, photography, stock
P7	Photography & Images	picfindr.com	photos, free, images, photography, stock

Order of the experiment's websites:

Pretest and Study 1: W1, W2, J1, J5, P4, J6, J3, P5, W3, J2, P1, W4, P2, P3, J4, P6, P7

Study 2: W1, W2, J4, P5, J6, J2, P4, J3, W3, P2, W4, P3, P1, J1